ELSEVIER
ACADEMIC
PRESS

Encyclopedia of

SOCIAL MEASUREMENT

Editor-in-Chief

Kimberly Kempf-Leonard

Volume 1 A–G

Editorial Board

Editorial Board

Editor Biography

Dr. Kempf-Leonard is Professor of Sociology, Crime and Justice Studies, and Political Economy at the University of Texas at Dallas. Prior to her appointment at UTD in 2000, she was Associate Professor and Graduate Director of Criminology and Criminal Justice at the University of Missouri at St. Louis. She also served for ten years as a gubernatorial appointee to the Missouri Juvenile Justice Advisory Group. She received her Ph.D. at the University of Pennsylvania in 1986; M.A. at the University of Pennsylvania in 1983; M.S. at the Pennsylvania State University in 1982; B.S. at the University of Nebraska in 1980. Her book *Minorities in Juvenile Justice* won the 1997 Gustavus Myers Award for Human Rights in North America. Her publications have appeared in: *Criminology, Justice Quarterly, Journal of Criminal Law & Criminology, Crime & Delinquency, Journal of Quantitative Criminology, Advances in Criminological Theory, Punishment & Society, Corrections Management Quarterly, the Journal of Criminal Justice, Criminal Justice Policy Review, The Justice Professional, Youth and Society, The Corporate Finance Reader*, and *The Modern Gang Reader.*

Editorial Board

Foreword

Not long ago, and perhaps still today, many would expect an encyclopedia of social measurement to be about quantitative social science. The *Encyclopedia of Social Measurement* excellently defies this expectation by covering and integrating both qualitative and quantitative approaches to social science and social measurement. The *Encyclopedia of Social Measurement* is the best and strongest sign I have seen in a long time that the barren opposition between quantitative and qualitative research, which has afflicted the social sciences for half a century, is on its way out for good. As if the Science Wars proper—between the social and natural sciences—were not enough, some social scientists found it fitting to invent another war within the social sciences, in effect a civil war, between quantitative and qualitative social science. Often younger faculty and doctoral students would be forced to take sides, and the war would reproduce within disciplines and departments, sometimes with devastating effects. This, no doubt, has set social science back. We cannot thank the editors and contributors to the *Encyclopedia of Social Measurement* enough for showing us there is an effective way out of the malaise.

This volume demonstrates that the sharp separation often seen in the literature between qualitative and quantitative methods of measurement is a spurious one. The separation is an unfortunate artifact of power relations and time constraints in graduate training; it is not a logical consequence of what graduates and scholars need to know to do their studies and do them well. The *Encyclopedia of Social Measurement* shows that good social science is opposed to an either/or and stands for a both/and on the question of qualitative versus quantitative methods. Good social science is problem-driven and not methodology-driven, in the sense that it employs those methods which for a given problematic best help answer the research questions at hand. To use a simple metaphor, asking whether social science is best served by qualitative or quantitative methods is about as intelligent as asking a carpenter whether a hammer or a saw is the better tool.

So far every effort has been unsuccessful in the social sciences at arriving at one canon for how to do science, most conspicuously the attempt at emulating the natural science model. Different explanations exist of this phenomenon, from Anthony Giddens' so-called double hermeneutic to Hubert Dreyfus' tacit skills argument. It is a great strength of the *Encyclopedia of Social Measurement* that it stays clear of the unity of science argument for social science, and of any other attempts at imposing one dominant paradigm on what social science is and how it should be conducted. The editors and most of the contributors have rightly seen that success in social science and social measurement lies with the type of methodological and epistemological pluralism, which is a distinguishing feature of the encyclopedia. Together with its impressive substantive breadth—covering the full range of social measurement from anthropology, sociology, political science, economics, and business administration over urban studies, environment, geography, demography, history, criminology, and law to neuroscience, biomedicine, nursing, psychology, linguistics, and communication—this healthy pluralism will prove the *Encyclopedia of Social Measurement* to be a robust and indispensable companion to all working social scientists for many years to come.

BENT FLYVBJERG
Professor of Planning,
Department of Development and Planning,
Aalborg University, Denmark

Preface

> *Methodology . . . [has] developed as a bent of mind rather than as a system of organized principles and procedures. The methodologist is a scholar who is above all analytical in his approach to his subject matter. He tells other scholars what they have done, or might do, rather than what they should do. He tells them what order of finding has emerged from their research, not what kind of result is or is not preferable. This kind of analytical approach requires self-awareness on the one hand, and tolerance, on the other. The methodologist knows that the same goal can be reached by alternative roads.*
>
> (Lazarsfeld and Rosenberg, 1955, p. 4)

In the social sciences we use methodology to try to answer questions about how and why people behave as they do. Some types of behavior are very common or routine, while others happen rarely or only in certain situations. When you realize that every conceivable type of behavior is within the realm of possible subjects for us to study, you can begin to appreciate the scope of social science. Beyond identifying human activities and the boundaries in which they occur, social scientists also want to explain why behaviors happen. In looking for causes, social scientists pursue all dimensions of the social world. We look at personal traits of individuals, characteristics of interactions between people, and contextual features of the communities and cultures in which they live. We study people who lived in the past, try to improve the quality of life today, and anticipate what the future will hold. It is difficult to think of a topic that involves people for which a social scientist could not investigate.

Given all we do, it is good that there are so many of us. You will find social scientists in university departments as professors of sociology, psychology, anthropology, political science, and economics. You will also find professors of geography, history, philosophy, math, management, planning, finance, journalism, architecture, humanities, and art who are social scientists. Even this multidisciplinary list is not exhaustive. There are important and prevalent social science investigations that influence decision-making in the world outside of universities too. Social scientists are world-wide and work in all branches of government, large and small organizations, and many types of businesses. Daily life for most people is influenced by social science research in marketing, insurance, and government. However, not everyone in these positions is a social scientist; the distinction involves scientific inquiry, or the approach used to try to answer questions about behavior. As the definition cited above conveys, good science includes tolerance and appreciation for many methodological paths. This encyclopedia of social science methodology provides 356 entries written by social scientists about what they do.

The entries in this encyclopedia cover many forms of measurement used by social scientists to study behavior. Eleven substantive sections delineate social sciences and the research processes they follow to measure and provide knowledge on a wide range of topics. The encyclopedia has an extensive index too, because many topics include issues that are relevant in more than one section. From many perspectives and strategies, these volumes describe the research questions social scientists ask, the sources and methods they use to collect information, and the techniques they use to analyze these data and provide answers to the important questions.

Each section includes entries that address important components of quantitative and qualitative research methods, which are dissected and illustrated with examples from diverse fields of study. The articles convey research basics in sufficient detail to explain even the most complicated statistical technique, and references for additional information are noted for each topic. Most entries describe actual research experiences to illustrate both the realm of possibilities and the potential challenges that might be encountered. Some entries describe major contributions and the social scientists who made them. The authors are accomplished methodologists in their fields of study. They explain the steps necessary to accomplish the measurement goals, as well as provide their practical advice for ways in which to overcome the likely obstacles.

Collectively, the entries in this encyclopedia also convey that no single approach, type of data, or technique of analysis reigns supreme. Indeed, plenty of disagreements exist among social scientists about what constitutes the "best" measurement strategy. Often distinctions are made between quantitative and qualitative methodologies, or are

discipline-specific. Some preferences can be linked to a specific field of study or research topic; others, related to time and location, coincide with how new ideas and advances in technology are shared. Sometimes we don't even agree on what is the appropriate question we should try to answer!

Although our views differ on what is ideal, and even on what are the appropriate standards for assessing measurement quality, social scientists generally *do* agree that the following five issues should be considered:

1. We agree on the need to be clear about the scope and purpose of our pursuits. The benchmarks for evaluating success differ depending on whether our intent is to describe, explain, or predict and whether we focus extensively on a single subject or case (e.g., person, family, organization, or culture) or more generally on patterns among many cases.
2. We agree on the need to make assurances for the ethical treatment of the people we study.
3. We agree on the need to be aware of potential sources of measurement error associated with our study design, data collection, and techniques of analysis.
4. We agree it is important to understand the extent to which our research is a reliable and valid measure of what we contend. Our measures are reliable if they are consistent with what others would have found in the same circumstances. If our measures also are consistent with those from different research circumstances, for example in studies of other behaviors or with alternate measurement strategies, then such replication helps us to be confident about the quality of our efforts. Sometimes we'd like the results of our study to extend beyond the people and behavior we observed. This focus on a wider applicability for our measures involves the issue of generalizability. When we're concerned about an accurate portrayal of reality, we use tools to assess validity. When we don't agree about the adequacy of the tools we use to assess validity, sometimes the source of our disagreements is different views on scientific objectivity.
5. We also agree that objectivity merits consideration, although we don't agree on the role of objectivity or our capabilities to be objective in our research. Some social scientists contend that our inquiries must be objective to have credibility. In a contrasting view of social science, or epistemology, objectivity is not possible and, according to some, not preferable. Given that we study people and are human ourselves, it is important that we recognize that life experiences necessarily shape the lens through which people see reality.

Besides a lack of consensus within the social sciences, other skeptics challenge our measures and methods. In what some recently have labeled "the science wars," external critics contend that social scientists suffer "physics envy" and that human behavior is not amenable to scientific investigation. Social scientists have responded to "anti-science" sentiments from the very beginning, such as Emile Durkhiem's efforts in the 19th century to identify "social facts." As entertaining as some of the debates and mudslinging can be, they are unlikely to be resolved anytime soon, if ever. One reason that Lazarsfeld and Rosenberg contend that tolerance and appreciation for different methodological pathways make for better science is that no individual scientist can have expertise in all the available options. We recognize this now more than ever, as multidisciplinary teams and collaborations between scientists with diverse methodological expertise are commonplace, and even required by some sources of research funding.

Meanwhile, people who can be our research subjects continue to behave in ways that intrigue, new strategies are proffered to reduce social problems and make life better, and the tool kits or arsenals available to social scientists continue to grow. The entries in these volumes provide useful information about how to accomplish social measurement and standards or "rules of thumb." As you learn these standards, keep in mind the following advice from one of my favorite methodologists: "Avoid the fallacy fallacy. When a theorist or methodologist tells you you cannot do something, do it anyway. Breaking rules can be fun!" Hirschi (1973, pp. 171–2). In my view nothing could be more fun than contemporary social science, and I hope this encyclopedia will inspire even more social science inquiry!

In preparing this encyclopedia the goal has been to compile entries that cover the entire spectrum of measurement approaches, methods of data collection, and techniques of analysis used by social scientists in their efforts to understand all sorts of behaviors. The goal of this project was ambitious, and to the extent that the encyclopedia is successful there are many to people to thank. My first thank you goes to the members of the Executive Advisory Board and the Editorial Advisory Board who helped me to identify my own biased views about social science and hopefully to achieve greater tolerance and appreciation. These scientists helped identify the ideal measurement topics, locate the experts and convince them to be authors, review drafts of the articles, and make the difficult recommendations required by time and space considerations as the project came to a close. My second thank you goes to the many authors of these 356 entries. Collectively, these scholars represent well the methodological status of social science today. Third, I thank the many reviewers whose generous recommendations improved the final product. In particular I extend my personal thanks to colleagues at the University of Texas at Dallas, many of whom participated in large and small roles in this project, and all of whom have helped me to broaden my appreciation of social

measurement. Finally, I thank Scott Bentley, Kirsten Funk, Kristi Anderson, and their colleagues at Elsevier for the opportunity and their encouragement when the tasks seemed overwhelming. Scott's insights to the possibilities of a project such as this and the administrative prowess of both Kirsten and Kristi helped make this a reality.

Good science is a cumulative process, and we hope this project will be ongoing and always improving. Despite our best efforts to identify topics and authors, sometimes we failed. If you have suggestions, criticisms, or information worth considering, I hope you will let me know.

Hirschi, Travis (1973). Procedural rules and the study of deviant behavior. *Social Problems* **21**(2), 159–173.

Lazarsfeld, Paul and Morris Rosenberg (1955). *The Language of Social Research.* The Free Press, New York.

KIMBERLY KEMPF-LEONARD

Paleodemography

James W. Wood
Pennsylvania State University, University Park, Pennsylvania, USA

Glossary

age-at-death distribution The distribution of individuals in a sample (e.g., a skeletal sample) by known or estimated ages at death. The empirical age-at-death distribution provides the fundamental data for paleodemographic analysis.

age mimicry A statistical bias whereby the estimated age-at-death distribution of a target sample (q.v.) tends to reproduce the known age-at-death distribution of whatever reference sample (q.v.) was used to generate it. This bias can be eliminated with the proper statistical methods.

ancient DNA (aDNA) Genetic material (mitochondrial or nuclear DNA) recovered from archaeological or paleontological samples using the polymerase chain reaction. In paleodemography and paleopathology, the aDNA of interest may be that of the actual humans being investigated or may be that of the pathogens infecting them or food remains associated with them.

demographic nonstationarity The failure of a population to conform to the theoretical ideal of a stationary population, i.e., one which is closed to in- and out-migration and has an intrinsic rate of increase equal to zero, age-specific schedules of fertility and mortality that are unchanging over time, and the equilibrium age distribution uniquely determined by those age-specific birth and death rates. In the presence of demographic nonstationarity, age-specific mortality rates cannot be estimated unless the effects of population increase are removed statistically.

paleopathology The branch of osteology concerned with the identification and classification of disease-related features in ancient tissues, especially in bone but also in dry soft tissue. The ultimate goal of paleopathology is to make inferences about disease frequency and health in ancient populations, concerns that also fall under the heading of paleoepidemiology.

reference sample A sample of skeletons with documented ages at death whose characteristics are used as a standard for estimating unknown ages in a target sample (q.v.). The documented ages are usually treated as known without error, which is probably never strictly the case and is sometimes very far from the truth.

taphonomy The study of the formation of paleontological or archaeological samples, including the processes of deposition, preservation, and recovery. Several biases in paleodemography are taphonomic in nature, including the tendency for skeletons of the very young to be underrepresented because of poor preservation.

target sample A sample of skeletons of unknown ages at death, usually from an archeological site, whose ages are to be estimated as a first step in paleodemographic analysis.

Paleodemography is the branch of biological anthropology devoted to the reconstruction of past populations using skeletal samples from archaeological excavations. Skeletons may be studied at the time of (or soon after) excavation, or may sit in museum collections for many years before being examined. In either case, paleodemographers use data on the distribution of skeletons by age at death and sex to reconstruct general population characteristics, especially mortality patterns. Paleopathological data on bony lesions indicative of growth faltering, infections, or chronic illness may provide supplementary information on individual-level health. Although serious methodological problems have dogged the field from its inception, recent advances hold promise that paleodemography will become an important and reliable source of information on the many populations in the past that did not leave more conventional demographic records.

Introduction

Paleodemography attempts to reconstruct past population structure using samples of human skeletons from archaeological excavations. Its chief claim to legitimacy

is that it provides demographic information—albeit of a limited, indirect, and uncertain sort—about the many human populations in the past that left no written records. In principle, it also allows reconstruction of demographic trends over time spans unattainable by any other branch of population science. Because of persistent methodological problems, however, paleodemographic analysis has achieved limited credibility among mainstream demographers. Yet while it is fair to say that past paleodemographic analyses were often too crude to be believable, it is also true that methodological advances over the past decade or so—advances with which most demographers are unfamiliar—have done much to place paleodemography on a firm scientific footing. The most important advances have been in the areas of age estimation, mortality analysis, adjustments for the effects of demographic nonstationarity on skeletal age-at-death distributions, and corrections for biases in the distribution of pathological lesions caused by selective mortality.

Age Estimation

Osteologists have made great progress in identifying reliable skeletal markers of age. Information on age at death is provided by such things as dental development, annual increments in dental cementum, closure of long-bone epiphyses and cranial sutures, and changes in the articular surfaces of the pelvis. Ages based on such features are subject to differing degrees of error arising from the inherent variability of the underlying processes of maturation and senescence; juveniles can be aged much more reliably than adults, and younger adults more reliably than older adults. But all paleodemographic age estimates are inherently error-prone and always will be. No matter how much osteological work is done to reduce the error and identify new age indicators, a large degree of aging error will always be a part of paleodemography. The deepest problems of paleodemographic age estimation are thus statistical rather than purely osteological.

In addition to a target sample (the archaeological skeletons whose ages are to be estimated), the paleodemographer needs access to a reference sample of skeletons whose ages at death are known (or at least documented, which is not always the same thing). Several well-known reference samples—for example, the Hamman-Todd and Terry Collections—provide reasonably accurate data on the joint distribution of c and a, where c is a vector of skeletal traits that provide information on age and a is age itself. For the target sample, however, we know only the marginal distribution of c, from which we hope to estimate the marginal distribution $\Pr(a)$ of ages at death. One of several parametric or nonparametric methods can be applied to data from the reference sample to estimate the conditional probability density or mass function $\Pr(c \mid a)$. If these estimates are to be used in aging archaeological skeletons, we need to make an invariance assumption that the joint distribution of c and a is identical in the two populations from which the reference and target samples were drawn. It is by no means clear that this assumption is warranted for many skeletal traits, and an ongoing goal of paleodemography is to identify indicators that are both informative about age and reasonably invariant across human populations.

Insofar as the invariance assumption is correct, it would seem to make sense to combine data on $\Pr(c)$ in the target sample and the joint distribution of a and c in the reference sample to estimate $\Pr(a \mid c)$ for each individual skeleton. But according to Bayes' theorem,

$$\Pr(a \mid c) = \frac{\Pr(c \mid a)\Pr(a)}{\int_0^\infty \Pr(c \mid x)\Pr(x)\, dx},$$

where $\Pr(a)$ is the age-at-death distribution in the target sample, which is unknown. Müller *et al.* have recently suggested using the marginal distribution

$$\Pr(c) \int_0^\infty \Pr(c \mid a)\Pr(a)\, da$$

as a basis for maximum likelihood estimation of $\Pr(a)$. If we provide a parametric specification for $\Pr(a)$—for example, as a Gompertz-Makeham, bi-Weibull, or Siler function—we can estimate the parameters of the age-at-death distribution by maximizing the likelihood function

$$L = \prod_{i=1}^{n} \Pr(c_i) = \prod_{i=1}^{n} \int_0^\infty \Pr^*(c_i \mid a)\Pr(a \mid \theta)\, da,$$

where n is the size of the target sample, c_i is the vector of skeletal characteristics observed in the ith skeleton in the target sample, θ is a set of parameters for the parametric $\Pr(a)$ model (to be estimated), and the asterisk denotes an empirical estimate from the reference sample. Once the parameters of $\Pr(a \mid \theta)$ have been estimated, the expected ages of individual skeletons can be found by a straightforward application of Bayes' theorem. This approach to age estimation is now called the *Rostock protocol* because it grew out of a series of workshops held at the Max Planck Institute for Demographic Research in Rostock, Germany, although significant parts of the protocol were anticipated in the earlier work of Konigsberg and Frankenburg.

It will seem strange to orthodox paleodemographers that they need to estimate the entire age-at-death distribution before they can estimate individual ages—the reverse of their usual procedure. But the Rostock protocol actually solves a number of problems that have long plagued paleodemography, including the problem of so-called age mimicry (the tendency of the estimated age-at-death distribution to be biased in the direction

of the age distribution—often highly unusual—of the reference sample) first noted by Bocquet-Appel and Masset. In addition, the method can be used to obtain, not just point or interval estimates of age, but the entire error structure of the age estimates. There are important statistical problems that remain to be solved, such as whether to use discrete categories or "staged" traits versus more continuous age indicators and how best to use multivariate skeletal data when traits are correlated in their age trajectories. But these problems can all be attacked within the framework of the Rostock protocol.

Mortality Analysis

For years paleodemographers have used skeletal age-at-death data to compute life tables based on some simple modifications of conventional life-table techniques originally developed by Acsádi and Nemeskéri. Though this approach is still a common one, the paleodemographic use of life tables can be criticized on several grounds. First, paleodemographic studies do not produce the kinds of data needed to compute life-table mortality rates using standard methods—specifically, the numbers of deaths among people at each (known) age and the number of person-years of exposure to the risk of death at that age during some well-defined reference period. Second, the use of fixed age intervals in the life table implies that the ages of all skeletons are known within the same margin of error, including those of fragmentary skeletons that exhibit only a few, unreliable indicators of age. Third, the life table is a wasteful way to use the small samples typical of paleodemographic studies—samples that are often on the order of a few dozen or a few hundred skeletons. In computing a life table we need to estimate one parameter (an age-specific mortality rate) for every age × sex category in the table. Few paleodemographic samples will support such an extravagant approach to estimation.

The Rostock protocol supports an alternative approach to paleodemographic mortality analysis. If unbiased estimates of the parameters of $\Pr(a \mid \theta)$ can be obtained for the target population of interest, and if the effects of demographic non-stationarity can be removed (see below), the parameter estimates can be used to derive the survival function, age-specific hazard of death, life-expectancies, and anything else we might hope to learn from life-table analysis.

Demographic Nonstationarity

Another shortcoming of traditional paleodemographic life-table analysis is that it assumes that the population under investigation was stationary—that it was closed to migration and had an intrinsic rate of increase equal to zero, age-specific schedules of fertility and mortality that were unchanging over time, and an equilibrium age distribution induced by those age-specific birth and death rates. Only in this special case is the empirical age distribution of skeletons expected to have a simple, straightforward relationship to the cohort age-at-death column in the life table.

As demographers have long realized, the age structure of a nonstationary population (and thus the number of its members at risk of death at each age) is more sensitive to changes in fertility than to similar changes in mortality. Thus, age-at-death distributions from different populations are at least as likely to reflect fertility differences as genuine differences in mortality. This incontrovertible fact of mathematical demography has given rise to the odd notion that paleodemographic age-at-death estimates are more informative about fertility than mortality. In fact, all we can ever hope to estimate about fertility from such data is the crude birth rate, which is scarcely a measure of fertility at all. But if we could correct for demographic nonstationarity, we could extract quite a bit of information about age-specific mortality from skeletons, and perhaps even estimate the population's growth rate.

Let $f_0(a)$ be the expected age-at-death distribution for a single birth cohort in the target population. If the target population was stationary, the same distribution holds for all deaths occurring in the population. But even if we cannot take it for granted that the population was stationary, it may be reasonable to assume that it was stable. That is, we may be able to make all the assumptions listed above for the stationary population, except allowing for the possibility of a nonzero growth rate. (The assumption of stability is much less restrictive than that of stationarity: even when fertility and mortality rates are changing and migration is occurring, most human populations still closely approximate a stable age distribution at any given time.) In a stable but nonstationary population, the age-at-death distribution is only partly a function of age-specific mortality; it is also influenced by the number of living individuals at risk of death at each age, which is influenced in turn by population growth. More precisely, the probability density function for ages at death in a stable population with growth rate r is

$$f_r(a) = \frac{\mu(a)S(a)e^{-ra}}{\int_0^\infty \mu(x)S(x)e^{-rx}\,dx} = \frac{f_0(a)e^{-ra}}{\int_0^\infty f_0(x)e^{-rx}\,dx},$$

where $\mu(a)$ is the force of mortality and $S(a)$ is the survival function at age a. This expression also applies to all the skeletons accumulated by a stable population over some more or less extended span of time—for example, the period over which skeletons were deposited in a particular cemetery. In principle, then, we can treat $f_r(a)$ as the $\Pr(a \mid \theta)$ function in the Rostock protocol and

estimate r as an additional parameter of the model, if we can assume that the population was stable. If it was not stable, at least approximately, we have probably reached the limits of what we can ever hope to learn about age-specific mortality from skeletal samples.

Selective Mortality and Paleopathological Analysis

When trying to interpret the mortality patterns of past populations, it would seem to make sense to use the information on individual-level health provided by the kinds of bony lesions studied by paleopathologists. Such lesions include marks in bones and teeth indicative of stress-related periods of growth faltering during preadult life, other dental and bony lesions caused by undernutrition (both general protein-calorie malnutrition and deficiencies of specific micronutrients), changes in the structure of bones caused by infection, and signs of chronic conditions such as arthritis and certain malignancies. Since most paleodemographers are also paleopathologists, the two kinds of analysis almost inevitably go together. There are, however, important problems in interpreting paleopathological findings at the population level that have received too little attention in the past.

The most obvious of these problems are the poor sensitivity and specificity of osteological indicators of health and disease. Most disease processes affect bone only in rare and unusually advanced cases; hence, bony lesions do not reveal most cases of the associated disease. In addition, specificity is poor because several unrelated disease processes can induce indistinguishable changes in bone. These problems are aggravated by the fact that most common diseases of the preindustrial past cause no bony changes whatsoever and thus go unmarked in the skeletal record. Much work needs to be done to develop more general osteological indicators of health and disease. One recent development that holds some promise is the extraction of microbial DNA from ancient bone, making it possible to detect the presence of infections that do not cause gross histological changes.

A deeper problem has to do with a form of selectivity bias that is inherent in the formation of skeletal samples. (This bias is quite distinct from the common taphonomic biases caused by differential burial and preservation.) All the individuals represented in a skeletal sample share one important characteristic: they are dead. As such, they are unlikely to be representative of the living community from which they were drawn. People who die at any given age presumably have characteristics that differentiate them from the larger group of living people originally at risk of death at that age. In particular, they are more likely to have pathological conditions, some of which may leave bony markers. In short, mortality is selective with respect to the very features of interest to the paleopathologist. While it is true that everyone dies eventually and (in theory) becomes available for paleopathological examination, it is also true that those who die are likely to have conditions, perhaps acquired soon before the time of death, that distinguish them from surviving members of their cohort.

Because of selective mortality, the frequency of pathological lesions in a skeletal sample is generally a poor reflection of the prevalence of the associated disease process in the living population. To understand the relationship between the two, it is necessary to model the process of selective mortality itself. This requires us, at a minimum, to model the individual-level distribution of risks of developing the pathological condition and the influence of the condition on the relative risk of death at each age. Recent work in this area has drawn upon models of heterogeneous frailty developed by mathematical demographers over the past two decades. While this work is promising, the problem of selective mortality is still a profound one for paleopathological analysis.

Prospects

During the past decade, the most important developments in paleodemography have been methodological, not substantive. But now that paleodemographic methods have become more sophisticated, there is every reason to expect that important empirical results will be forthcoming in the near future. It is likely, too, that the findings of paleodemography will be strengthened by the study of DNA extracted from ancient bones—a field that is already starting to provide insights into the ancestry and kinship structure of past populations, as well as the pathogens that infected them. There is also a new and encouraging movement to bring archaeological settlement studies—long an established approach to past population dynamics—into the purview of paleodemography. While mainstream demographers were once justified in dismissing the field of paleodemography, it may be time for them to rethink their skepticism.

See Also the Following Articles

Attrition, Mortality, and Exposure Time • Demography • Mathematical Demography

Further Reading

Acsádi, G., and Nemeskéri, J. (1970). *History of Human Life Span and Mortality.* Akadémiai Kiadó, Budapest.

Bocquet-Appel, J. P., and Masset, C. (1982). Farewell to paleodemography. *J. Human Evol.* **11,** 321–333.

Hoppa, R. D., and Vaupel, J. W. (eds.) (2002). *Paleodemography: Age Distributions from Skeletal Samples*. Cambridge University Press, Cambridge, UK.

Horowitz, S., Armelagos, G., and Wachter, K. (1988). On generating birth rates from skeletal samples. *Am. J. Phys. Anthropol.* **76,** 189–196.

Kolman, C. J., and Tuross, N. (2000). Ancient DNA analysis of human populations. *Am. J. Phys. Anthropol.* **111,** 5–23.

Konigsberg, L. W., and Frankenberg, S. R. (1992). Estimation of age structure in anthropological demography. *Am. J. Phys. Anthropol.* **89,** 235–256.

Konigsberg, L. W., and Frankenberg, S. R. (1994). Paleodemography: "Not quite dead." *Evolution. Anthropol.* **3,** 92–105.

Meindl, R. S., and Russell, K. F. (1998). Recent advances in method and theory in paleodemography. *Ann. Rev. Anthropol.* **27,** 375–399.

Milner, G. R., Wood, J. W., and Boldsen, J. L. (2000). Paleodemography. In *Biological Anthropology of the Human Skeleton* (M. A. Katzenberg and S. R. Saunders, eds.), pp. 467–497. Wiley–Liss, New York.

Müller, H. G., Love, B., and Hoppa, R. D. (2002). Semiparametric method for estimating paleodemographic profiles from age indicator data. *Am. J. Phys. Anthropol.* **117,** 1–14.

Paine, R. R. (ed.) (1997). *Integrating Archaeological Demography: Multidisciplinary Approaches to Prehistoric Population*. Southern Illinois University Press, Carbondale, IL.

Sattenspiel, L., and Harpending, H. C. (1983). Stable populations and skeletal age. *Am. Antiquity* **48,** 489–498.

Stone, A. C. (2000). Ancient DNA from skeletal remains. In *Biological Anthropology of the Human Skeleton* (M. A. Katzenberg and S. R. Saunders, eds.), pp. 351–371. Wiley–Liss, New York.

Wood, J. W., Milner, G. R., Harpending, H. C., and Weiss, K. M. (1992). The osteological paradox. *Curr. Anthropol.* **33,** 343–370.

Partial Credit Model

Geoff N. Masters
Australian Council for Educational Research, Camberwell, Australia

Glossary

dichotomous scoring The use of two ordered categories to record the outcome of an individual's interaction with an item (usually "Right"/"Wrong").
objective comparison The possibility of comparing the locations of two parameters on a measurement variable independently of all other parameters and through two alternative observable events.
ordered response categories A set of ordered alternatives for recording the outcome of an individual's interaction with an item.
partial credit scoring The use of more than two ordered categories to record the outcome of an individual's interaction with an item (usually to recognize degrees of correctness or completion).
Rasch model A model for measuring based on the principle of objective comparison.

The partial credit model is a statistical model for the analysis of tests and questionnaires that provide several ordered response categories for each item. Examples of such items are test and examination questions that award part marks for partially complete or partially correct answers, rated student performances, and graded essay questions. The partial credit model is used to calibrate items of this general kind and to measure respondents along a latent measurement variable. The model is best understood as Rasch's model for dichotomies applied to adjacent pairs of response categories in an ordered sequence of categories.

Introduction

It is common in the social and behavioral sciences to use ordered response categories to record individuals' responses to measurement items. For example, respondents may be asked to indicate the frequency with which they engage in a particular activity by choosing one of several ordered time periods. They may be asked to indicate the strength of their agreement or disagreement with a statement by choosing from a set of ordered alternatives on a Likert scale. Or they may be asked to record their response by choosing one of a number of points on a provided scale (e.g., semantic differential).

In addition to measurement instruments that provide respondents with ordered response alternatives, many other instruments require judges to evaluate individuals' responses and to assign them to one of several ordered categories. For example, a psychologist may rate an aspect of a child's language development or the child's performance on a psychomotor task using a set of ordered rating points. In educational contexts, many areas of learning require judgments of the quality of students' responses to tasks, including instrumental music, dance, drama, written expression, art portfolios, research projects, oral language, computer programming, and technology (food, metals, ceramics, etc). In these areas of student learning it usually would be inadequate to record a response as either right or wrong. Judgments must be made of the quality of students' responses and work, and these judgments invariably are made and recorded against a scale of ordered possibilities. Points on this scale sometimes are labeled with letter grades (e.g., A to E) or numbers (e.g., 1 to 10).

In most systems of measurement, and certainly in the measurement of physical properties, individual instruments are calibrated against general measurement scales. The result is that measures such as 35 cm and 28°C have a meaning that does not depend on the particulars of the instrument used to obtain them. This characteristic is sometimes referred to as a distinguishing feature of measures.

Ordinary test, questionnaire, and examination scores do not have this property. A score of 35 items correct on a test does not have a general meaning; the ability required to score 35 will depend on the number and difficulties of the items on that particular test. The partial credit model is designed to convert performances on instruments measuring the same variable into numerical measures on a common scale. It is one of a number of statistical models developed to do this. The distinguishing properties of the model are that it can be used to calibrate items that provide multiple ordered response alternatives, and it belongs to the Rasch family of models.

Specific Objectivity

The partial credit model is a particular application of the model for dichotomies developed by Danish mathematician Georg Rasch. An understanding of the partial credit model thus depends on an understanding of Rasch's model for dichotomies, the properties of this model, and, in particular, Rasch's concept of specific objectivity.

Rasch used the term specific objectivity in relation to a property of the model for tests he developed during the 1950s. He considered this property to be especially useful in the attempt to construct numerical measures that do not depend on the particulars of the instrument used to obtain them.

This property of Rasch's model can be understood by considering two people A and B with imagined abilities θ_A and θ_B. If these two people attempt a set of test items and a tally is kept of the number of items N_{10} that person A answers correctly but B answers incorrectly and of the number of items N_{01} that person B answers correctly but A answers incorrectly, then under Rasch's model, the difference $\theta_A - \theta_B$ in the abilities of these two people can be estimated as:

$$\ln\left(\frac{N_{10}}{N_{01}}\right)$$

What is significant about this fact is that this relationship between the parameterized difference $\theta_A - \theta_{BAB}$ and the tallies N_{10} and N_{01} of observed successes and failures applies to *Any* selection of items when test data conform to Rasch's model. In other words, provided that the responses of A and B to a set of items are consistent with the model, the difference $\theta_A - \theta_{BAB}$ can be estimated by simply counting successes and failures without having to know or estimate the difficulties of the items involved. Any subset of items (e.g., a selection of easy items, hard items, even-numbered items, or odd-numbered items) can be used to obtain an estimate of the relative abilities of A and B from a simple tally table (Table I). The possibility of obtaining an estimate of the relative abilities of A and B that is not dependent on the details of the items used was referred to by Rasch as the possibility of specifically objective comparison.

Table I Tally Table

Person A \ *Person B*	*Wrong*	*Right*
Right	N_{10}	N_{11}
Wrong	N_{00}	N_{01}

Figure 1 Rasch's model begins with the idea of a measurement variable on which two objects A and B have imagined locations ξ_A and ξ_B.

Rasch's Model

In its most general form, Rasch's model begins with the idea of a measurement variable on which two objects A and B have imagined locations ξ_A and ξ_B (see Fig. 1). The possibility of estimating the relative locations of objects A and B on this variable depends on the availability of two observable events:

- An event X indicating that ξ_B exceeds ξ_A.
- An event Y indicating that ξ_A exceeds ξ_B.

Rasch's model relates the difference between objects AA and B to the events X and Y that they govern:

$$\xi_B - \xi_A = \ln\left(\frac{P_X}{P_Y}\right) \quad (1)$$

where P_X is the probability of observing X and P_Y is the probability of observing Y. Notice that, under the model, the odds P_X/P_Y of observing X rather than Y are dependent only on the direction and distance of ξ_B from ξ_A, and are uninfluenced by any other parameter. In 1977, Rasch described the comparison of two objects as objective if the result of the comparison was "independent of everything else within the frame of reference other than the two objects which are to be compared and their observed reactions."

An estimate of the difference between objects A and B on the measurement variable can be obtained if there are multiple independent opportunities to observe either event X or event Y. Under these circumstances, $\xi_B - \xi_A$ can be estimated as:

$$\ln\left(\frac{p_X}{p_Y}\right) = \ln\left(\frac{N_X}{N_Y}\right)$$

where p_X and p_Y are the proportions of occurrences of X and Y, and N_X and N_Y are the numbers of times X and Y occur in $N_X + N_Y$ observation opportunities.

Dichotomous Test Items

The most common application of Rasch's model is to tests in which responses to items are recorded as either wrong (0) or right (1). Each person n is imagined to have an ability θ_n and each item i is imagined to have a difficulty δ_i, both of which can be represented as locations on the variable being measured (see Fig. 2). In this case, observable event X is person n's success on item i, and observable event Y is person n's failure on item i, as shown in Table II Rasch's model applied to this situation is:

$$\tilde{\theta}_n - \delta_i = \ln\left(\frac{P_1}{P_0}\right) \tag{2}$$

If person n could have multiple independent attempts at item i, then the difference $\tilde{\theta}_n - \delta_i$ between person n's ability and item i's difficulty could be estimated as:

$$\ln\left(\frac{p_1}{p_0}\right) = \ln\left(\frac{N_1}{N_0}\right)$$

Although this is true in theory, and this method could be useful in some situations, it is not a practical method for estimating $\tilde{\theta}_n - \delta_i$ from test data because test takers are not given multiple attempts at the same item (and if they were, they would not be independent attempts). To estimate the difference $\tilde{\theta}_n - \delta_i$ from test data, it is necessary to estimate θ_n from person n's attempts at a number of items, and to estimate δ_i from a number of n's attempts at that item. In other words, the difficulties of a number of test items and the abilities of a number of test takers must be estimated simultaneously.

Comparing and Measuring Individuals

In the application of Rasch's model to tests, every person has an imagined location on the variable being measured. Two people m and n have imagined locations θ_m and θ_n (see Fig. 3). It follows from Eq. (2) that, if m and n attempt the same item and their attempts at that item are independent of one another, then the modeled difference between n and m is:

$$\tilde{\theta}_n - \theta_m = \ln\left(\frac{P_{10}}{P_{01}}\right) \tag{3}$$

where P_{10} is the model probability of person n succeeding but m failing the item, and P_{01} is the probability of person m succeeding but n failing that item.

It can be seen that Eq. (3) is Rasch's model, Eq. (1), applied to the comparison of two people on a measurement variable. The two observable events involve the success of one person but failure of the other in their attempts at the same item, as shown in Table III.

In this comparison of m and n, nothing was said about the difficulty of the item being attempted by these two people. This is because Eq. (3) applies to every item. The odds of it being person n who succeeds, given that one of these two people succeeds and the other fails, is the same for every item and depends only on the relative abilities of m and n.

Because the modeled odds P_{10}/P_{01} are the same for every item, the difference $\theta_n - \theta_m$ can be estimated as:

$$\ln\left(\frac{N_{10}}{N_{01}}\right)$$

where N_{10} is the number of items that person n has right but m has wrong, and N_{01} is the number of items that person m has right but n has wrong.

When test data conform to the Rasch model, the relative abilities of two individuals can be estimated in this way using any selection of items without regard to their difficulties (or any other characteristics). By making multiple pairwise comparisons of this kind, it is possible to estimate the relative locations of a number of people on the same measurement variable.

Comparing and Calibrating Items

In the application of Rasch's model to tests, every item has an imagined location on the variable being measured.

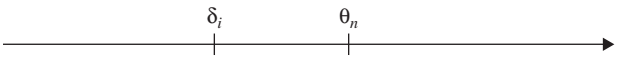


Figure 2 Each person n is imagined to have an ability θ_n and each item i is imagined to have a difficulty δ_i, both of which can be represented as locations on the variable being measured.

Figure 3 In an application of Rasch's model to tests, two people, m and n, have imagined locations θ_m and θ_n.

Table II Dichotomous Testing

$\xi_B - \xi_A$	*Observation opportunity*	*Observable event* X	Y
$\theta_n - \theta_i$	Person n attempts item i	1	0

Table III Dichotomous Testing, Comparing Individuals

$\xi_B - \xi_A$	*Observation opportunity*	*Observable event* X	Y
$\theta_n - \theta_m$	Individuals n and m independently attempt the same item	1,0	0,1

Figure 4 In an application of Rasch's model to tests, two items i and j have imagined locations δ_i and δ_j.

Table IV Dichotomous Testing, Comparing Items

$\xi_B - \xi_A$	*Observation opportunity*	*Observable event* X	Y
$\delta_i - \delta_j$	Items i and j independently attempted by the same person	0,1	1,0

Two items i and j have imagined locations δ_i and δ_j (see Fig. 4). It follows from Eq. (2) that, if items $i$and j are attempted by the same person and this person's attempts at items i and j are independent of one another, then the modeled difference between items i and j is:

$$\delta_i - \delta_j = \ln\left(\frac{P_{01}}{P_{10}}\right) \tag{4}$$

where P_{10} is the model probability of the person succeeding on item i but failing item j, and P_{01} is the probability of the person succeeding on item j but failing item i.

It can be seen that Eq. (4) is Rasch's model, Eq. (1), applied to the comparison of two items on a measurement variable. The two observable events involve the person's success on one item but failure on the other, as shown in Table IV.

In this comparison of items i and j, nothing was said about the ability of the person attempting them. This is because Eq. (4) applies to every person. The odds of success on item i given success on one item but failure on the other is the same for every person and depends only on the relative difficulties of items i and j.

Because the modeled odds P_{01}/P_{10} are the same for every person, the difference $\delta_i - \delta_j$ can be estimated as:

$$\ln\left(\frac{n_{01}}{n_{10}}\right)$$

where n_{10} is the number of people with item i right but j wrong, and n_{01} is the number of people with j right but i wrong.

When test data conform to the Rasch model, the relative difficulties of two items can be estimated in this way using any group of people without regard to their abilities (or any other characteristics). By making multiple pairwise comparisons of this kind, it is possible to estimate the relative locations of a number of items on the measurement variable.

Figure 5 When the model is applied to tests in which responses to items are recorded in several ordered categories labeled 0, 1, 2, . . . , K_i, each person n is imagined to have an ability θ_n and each item i is imagined to have a set of K_i parameters δ_{i1}, $\delta_{i2}, \ldots, \delta_{iK_i}$, each of which can be represented as a location on the variable being measured.

Table V Dichotomous Testing, Ordered Categories

$\xi_B - \xi_A$	*Observation opportunity*	*Observable event* X	Y
$\theta_n - \delta_{ik}$	Person n attempts item i	k	$k-1$

Application to Ordered Categories

The partial credit model applies Rasch's model for dichotomies to tests in which responses to items are recorded in several ordered categories labeled 0, 1, 2, . . . , K_i. Each person n is imagined to have an ability θ_n and each item i is imagined to have a set of K_i parameters δ_{i1}, $\delta_{i2}, \ldots, \delta_{iK_i}$, each of which can be represented as a location on the variable being measured (see Fig. 5), where δ_{ik} governs the probability of scoring k rather than $k-1$ on item i, as shown in Table V.

The Rasch model applied to this situation is:

$$\tilde{\theta}_n - \delta_{ik} = \ln\left(\frac{P_k}{P_{k-1}}\right) \tag{5}$$

In polytomous test items, objective comparison (and thus objective measurement) continues to depend on the modeling of the relationship between two imagined locations on the variable and two observable events. This comparison is "independent of everything else within the frame of reference," including other possible outcomes of the interaction of person n with item i. The conditioning out of other possible outcomes to focus attention only on the two observable events that provide information about the relative locations of the two parameters of interest is a fundamental feature of Rasch's model.

The conditioning on a pair of adjacent response alternatives has parallels with McFadden's 1974 assumption that a person's probability of choosing to travel by car rather than by bus should be independent of the availability of other options (e.g., train). McFadden referred to this as the assumption of "independence from irrelevant alternatives." In a similar way, it is assumed in this application of Rasch's model that a person's probability of

choosing or scoring k rather than $k-1$ is independent of all other possible outcomes.

When a person responds to an item with several ordered response categories, he or she must make a choice, taking into account all available alternatives. The partial credit model makes no assumption about the response mechanism underlying a person's choice. It simply proposes that, if category k is intended to represent a higher level of response than category $k-1$, then the probability of choosing or scoring k rather than $k-1$, should increase monotonically with the ability being measured.

As for dichotomously scored items, if person n could have multiple independent attempts at item i, then the difference $\tilde{\theta}_n - \delta_{ik}$ could be estimated from proportions or counts of occurrences of k and $k-1$:

$$\ln\left(\frac{p_k}{p_{k-1}}\right) = \ln\left(\frac{N_k}{N_{k-1}}\right)$$

However, because multiple independent attempts at test items usually are not possible, this method is not feasible in practice.

Comparing and Measuring Individuals

In the application of Rasch's model to tests in which responses to items are recorded in several ordered categories, every person has an imagined location on the variable being measured (see Fig. 6). It follows from Eq. (5) that, if individuals m and n attempt the same item and their attempts at that item are independent of one another, then the modeled difference between n and m is:

$$\tilde{\theta}_n - \theta_m = \ln\left(\frac{P_{k,k-1}}{P_{k-1,k}}\right) \tag{6}$$

where $P_{k,k-1}$ is the model probability of person n scoring k but m scoring $k-1$, and $P_{k-1,k}$ is the probability of person m scoring k but n scoring $k-1$ on that item.

It can be seen that Eq. (6), which applies for all values of k ($k = 1, 2, \ldots, K_i$), is Rasch's model, Eq. (1) (see Table VI). If one of either m and n scores k on an item, and the other scores $k-1$, then the probability of it being person n who scores k is the same for every item and depends only on the relative abilities of m and n.

Figure 6 In the application of Rasch's model to tests in which responses to items are recorded in several ordered categories, every person has an imagined location on the variable being measured.

Because the modeled odds $P_{k,k-1}/P_{k-1,}$k are the same for every item, the difference $\theta_n - \theta_m$ can be estimated as:

$$\ln\left(\frac{N_{k,k-1}}{N_{k-1,k}}\right)$$

where $N_{k,k-1}$ is the number of items on which person n scores k and m scores $k-1$, and $N_{k-1,k}$ is the number of items on which person m scores k and n scores $k-1$.

Once again, when test data conform to Rasch's model—the relative abilities of two people can be estimated in this way using any selection of items. And by making multiple pairwise comparisons of this kind, it is possible to estimate the relative locations of a number of people on the measurement variable.

Comparing and Calibrating Items

In polytomous items, each item parameter δ_{ik} ($k = 1, 2, \ldots, K_i$) is a location on the variable being measured. The parameters δ_{ik} and δ_{jk} from two different items i and j can be compared on this variable (see Fig. 7). It follows from Eq. (5) that, if items i and $j$are attempted by the same person and this person's attempts at items i and $j$are independent of one another, then the modeled difference between parameters δ_{ik} and δ_{jk} is:

$$\tilde{\delta}_{ik} - \delta_{jk} = \ln\left(\frac{P_{k-1,k}}{P_{k,k-1}}\right) \tag{7}$$

where $P_{k,k-1}$ is the probability of the person scoring k on item i but $k-1$ on item j, and $P_{k-1,k}$ is the probability of the person scoring k on item j but $k-1$ on item i.

It can be seen that Eq. (7), which applies for all values of k ($k = 1, 2, \ldots, K_i$), is Rasch's model, Eq. (1) (see Table VII).

Table VI Dichotomous Testing, Comparing Individuals with Ordered Categories

		Observable event	
$\xi_B - \xi_A$	*Observation opportunity*	*X*	*Y*
$\beta_n - \beta_m$	Individuals n and m independently attempt the same item	$k, k-1$	$k-1, k$

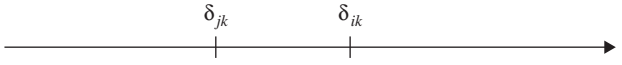


Figure 7 In polytomous items, each item parameter δ_{ik} ($k = 1, 2, \ldots, K_i$) is a location on the variable being measured. The parameters δ_{ik} and δ_{jk} from two different items i and j can be compared on this variable.

Table VII Polytomous Testing, Comparing Items

$\xi_B - \xi_A$	*Observation opportunity*	*Observable event* X	Y
$\delta_{ik} - \delta_{jk}$	Items i and j independently attempted by the same person	$k-1, k$	$k, k-1$

In this comparison of items i and j, nothing was said about the ability of the person attempting them. This is because Eq. (7) applies to every person. When a person attempts items i and j, the probability of the person scoring k on item i given that he or she scores k on one item and $k-1$ on the other is the same for every person.

Because the modeled odds $P_{k-1,k}/P_{k,k-1}$ are the same for every person, the difference $\delta_{ik}-\delta_{jk}$ can be estimated as:

$$\ln\left(\frac{n_{k-1,k}}{n_{k,k-1}}\right)$$

where $n_{k,k-1}$ is the number of people scoring k on item i but $k-1$ on item j, and $n_{k-1,k}$ is the number of people scoring k on item j but $k-1$ on item i.

When test data conform to Rasch's model, the difference $\delta_{ik}-\delta_{jk}$ can be estimated in this way using any group of people without regard to their abilities (or any other characteristics).

Comparisons with Other Models

The partial credit model is one of a number of models that have been introduced for the analysis of ordered response category data. To understand similarities and differences between these models, it is useful to identify two broad classes of models.

Models with Discrimination Parameters

In some models proposed for the analysis of test data, in addition to a location θ_n for each person n and a location δ_i for each item i, a discrimination parameter α_i is proposed for each item i. Among models for ordered response categories that include a discrimination parameter are Samejima's graded response model and Muraki's generalized partial credit model.

These models differ from the partial credit model in that they enable specifically objective comparisons, as described by Rasch. The reason for this can be seen most easily in two-parameter dichotomous item response theory (IRT) model:

$$\alpha_i(\tilde{\theta}_n - \delta_i) = \ln\left(\frac{P_1}{P_0}\right) \tag{8}$$

If we follow the steps outlined earlier and consider independent attempts of two people m and n at item i, then for the two-parameter IRT model we obtain:

$$\alpha_i(\tilde{\theta}_n - \theta_m) = \ln\left(\frac{P_{10}}{P_{01}}\right) \tag{9}$$

where P_{10} is the probability of person n succeeding but m failing item i, and P_{01} is the probability of person m succeeding but n failing.

It can be seen from Eq. (9) that the odds of person n succeeding but m failing, given that one of these two people succeeds and the other fails, is not the same for all items. Rather, the odds depend on the discrimination of the item in question.

To compare the locations of m and n on the measurement variable, it is not possible to ignore the particulars of the items involved and simply tally occurrences of (1,0) and (0,1). The comparison of θ_n and θ_m on the measurement variable is dependent not only on the two observable events (1,0) and (0,1) that they govern, but also on the details (the discriminations) of the items these two people take. For this reason, the two-parameter IRT model does not permit an objective comparison in the sense described by Rasch.

Models with Cumulative Thresholds

A second class of models for ordered response categories includes as parameters cumulatively defined thresholds. Each threshold parameter is intended to divide all ordered response alternatives to an item up to and including alternative $k-1$ from response alternatives k and above. L. L. Thurstone, who used the normal rather than logistic function to model thresholds referred to them as "category boundaries."

The threshold notion is used as the basis for Samejima's graded response model. Her model also includes an item discrimination parameter, but that is ignored here for the sake of simplicity. Samejima's model takes the form:

$$\theta_n - \gamma_{ik} = \ln\left[\frac{(P_k + P_{k+1} + \cdots + P_{K_i})}{(P_0 + P_1 + \cdots + P_{k-1})}\right]. \tag{10}$$

In this model, the item threshold γ_{ik} governs the probability of scoring k or better on item i.

Table VIII compares Samejima's graded response model with the partial credit model for an item with four ordered response alternatives labeled 0, 1, 2, and 3.

Table VIII Comparison of Samejima and Rasch Models for Polytomous Items

	Samejima	*Rasch*
Elementary equations (person n, item i, $K_i = 3$)	$\theta_n - \gamma_{i1} = \ln[(P_1 + P_2 + P_3)/P_0]$ $\theta_n - \gamma_{i2} = \ln[(P_2 + P_3)/(P_0 + P_1)]$ $\theta_n - \gamma_{i3} = \ln[P_3/(P_0 + P_1 + P_2)]$	$\theta_n - \delta_{i1} = \ln[P_1/P_0]$ $\theta_n - \delta_{i2} = \ln[P_2/P_1]$ $\theta_n - \delta_{i3} = \ln[P_3/P_2]$
Events being compared	Compound (e.g., response in category 1 or 2 or 3 rather than 0)	Simple (comparison of adjacent response categories)
Item parameters	Global/unconditional (each γ relates to all available response categories)	Local/conditional (each δ relates to adjacent response categories only)
Relationship of elementary equations	Dependent $(P_1 + P_2 + P_3)/P_0 > (P_2 + P_3)/(P_0 + P_1) > P_3/(P_0 + P_1 + P_2)$	Independent (e.g., odds of response in category 1 rather than 0 are independent of odds of response in category 2 rather than 1)
Implications for item parameters	$\gamma_{i1} < \gamma_{i2} < \gamma_{i3}$	δs are unfettered and free to take any value
Model for ordered categories	When brought together, the elementary equations provide a model for ordered response categories in which the person parameters cannot be conditioned out of the estimation procedure for the items	The elementary equations provide a model for ordered response categories in which the person parameters can be conditioned out of the estimation procedure for the items and vice versa
Specific objectivity	No	Yes

From Table VIII it can be seen that the observable events in this model are compound events, that is,

Event X: response in category 1 or 2 or 3
Event Y: response in category 0

The consequence is that the elementary equations in this model are not independent because $(P_1 + P_2 + P_3)/P_0 > (P_2 + P_3)/(P_0 + P_1) > P_3/(P_0 + P_1 + P_2)$. As a result, thresholds are not independent, but are always ordered $\gamma_{i1} < \gamma_{i2} < \gamma_{i3}$.

The elementary equations in Samejima's model lead to the following expressions for the probabilities of person n scoring 0, 1, 2, and 3 on item i:

$$P_{ni0} = 1 - \exp(\theta_n - \gamma_{i1})/[1 + \exp(\theta_n - \gamma_{i1})]$$

$$P_{ni1} = \exp(\theta_n - \gamma_{i1})/[1 + \exp(\theta_n - \gamma_{i1})] - \exp(\theta_n - \gamma_{i2})/[1 + \exp(\theta_n - \gamma_{i2})]$$

$$P_{ni2} = \exp(\theta_n - \gamma_{i2})/[1 + \exp(\theta_n - \gamma_{i2})] - \exp(\theta_n - \gamma_{i3})/[1 + \exp(\theta_n - \gamma_{i3})]$$

$$P_{ni3} = \exp(\theta_n - \gamma_{i3})/[1 + \exp(\theta_n - \gamma_{i3})]$$

It is not possible to condition one set of parameters (either the person parameters or the item thresholds) out of the estimation procedures for the other in this model.

In contrast, the elementary equations for the Rasch model (see Table VIII) lead to the following expressions for the probabilities of person n scoring 0, 1, 2, and 3 on item i:

$$P_{ni0} = 1/\Psi$$

$$P_{ni1} = \exp(\theta_n - \delta_{i1})/\Psi$$

$$P_{ni2} = \exp(2\theta_n - \delta_{i1} - \delta_{i2})/\Psi$$

$$P_{ni3} = \exp(3\theta_n - \delta_{i1} - \delta_{i2} - \delta_{i3})/\Psi$$

where Ψ is the sum of the numerators. In general, the partial credit model takes the form:

$$P_{nik} = \exp(k\theta_n - \delta_{i1} - \delta_{i2} - \cdots - \delta_{ik})/\Psi. \qquad (11)$$

It is possible to condition the person parameters out of the estimation procedures for the item parameters, and vice versa, in this model.

Other Rasch Models

As a member of the Rasch family of item response models, the partial credit model is closely related to other members of that family. In 1984, Masters and Wright described several members of this family and showed how each has as its essential element Rasch's model for dichotomies. Andrich's 1978 model for rating scales, for example, can be thought of as a version of the partial credit model with the added expectation that the response categories are defined and function in the same way for each item in an instrument. With this added expectation, instead of modeling a set of

m_i parameters for each item, a single parameter δ_i is modeled for item i, and a set of m parameters (τ_1, $\tau_2, \ldots, \tau_m$) are proposed for the common response categories. To obtain the rating scale version of the partial credit model, each item parameter in the model is redefined as $\delta_{ix} = \delta_i + \tau_x$. Wilson also has proposed a generalized version of the partial credit model in 1993.

Substantive Calibration of Measurement Variables

When the partial credit model is applied, it provides estimates of the item parameters (δ_{i1}, $\delta_{i2}, \ldots, \delta_{im_i}$) for each item i. When these estimates are substituted into Eq. (11), the estimated probabilities of scoring $0, 1, \ldots, m_i$ on item i are obtained for any specified ability $\hat{\theta}$

Figure 8A shows the model probabilities of scoring 0, 1, 2, and 3 on a four-category item calibrated with the partial credit model. Notice that the maxima of the response curves are in the order $0 < 1 < 2 < 3$ from left to right. This is a basic feature of the model; the response curve maxima are always ordered $0 < 1 < 2 < \cdots < m_i$ on the measurement variable.

The item parameters (δ_{i1}, δ_{i2}, $\ldots$, δ_{im_i}) have a simple interpretation in this picture. Each parameter δ_{ix} corresponds to the position on the measurement variable at which a person has the same probability of responding in category x as in category $x-1$ (i.e., $P_{ijx} = P_{ijx-1}$). The parameter estimates for the item in Figure 8A are thus at the intersections of the response curves for categories 0 and 1, 1 and 2, and 2 and 3.

The person ability parameter θ_j in the partial credit model is the modeled location of person j on the variable being measured. For each item i scored $0, 1, \ldots, m_i$, the model defines a set of m_i item parameters. These parameters, all of which are locations on the measurement variable, can be used to map the qualitative meaning of the variable and to interpret person parameters.

In the case of items scored right or wrong, only one parameter is estimated for each item and these item difficulties mark out and give qualitative meaning to a variable. Levels of ability are differentiated in terms of the kinds of items likely to be answered correctly. By convention, "likely" means with a probability ≥ 0.5, although in some testing programs higher probabilities of success (e.g., ≥ 0.7 or ≥ 0.8) are specified.

Open-ended and performance tasks of the kind the partial credit model is designed to analyze are usually intended to be accessible to a wide range of abilities and to differentiate among test takers on the basis of their levels of response. Response categories for each item capture this response diversity and thus provide the basis for the qualitative mapping of measurement variables and the consequent interpretation of ability estimates. For items with more than two response categories, however, the mapping of response categories on to measurement variables is a little less straightforward than for right/wrong scoring.

As previously noted, because each item parameter in the partial credit model is defined locally with respect to just two adjacent categories (rather than globally taking into account all categories simultaneously), the item parameters in the model can take any order. For the item in Fig. 8A they are ordered $\delta_{i1} < \delta_{i3} < \delta_{i2}$. Because they have this property, these parameters may not, in themselves, be helpful in marking out regions of an underlying variable. Three useful alternatives are illustrated in Fig. 8A–C.

The first method, illustrated in Fig. 8A, identifies regions of single most probable response to an item. The three shaded regions at the bottom of Fig. 8A indicate the single most probable response (0, 1, or 3) for the range of abilities shown.

A disadvantage of this method is that, in this case, it defines no region for a score of 2. In fact, it gives the impression that category 2 has disappeared entirely.

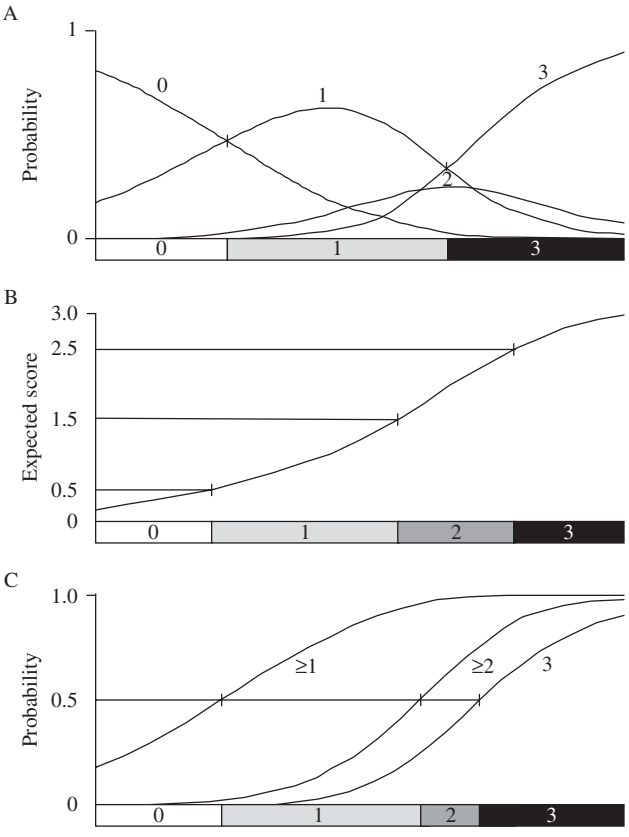


Figure 8 Probability and cumulative probability curves for the partial credit model.

This is not the case—a score of 2 on this item is a quite likely result ($p = 0.28$) for an ability at the junction of curves 1 and 3. When partial credit model item parameters are in the order $\delta_{i1} < \delta_{i2} < \delta_{i3}$, the parameters mark out regions of single most probable response for all categories. However, as can be seen from Fig. 8A, other orders of partial credit model item parameters are possible.

Some caution is required in the interpretation of "most probably" statements. Toward the far right of the shaded region labeled 1 in Fig. 8A, it is *not* the case that people will most probably score 1. Because the sum of the probabilities for categories 2 and 3 exceeds the probability of a response in category 1, people with abilities in this part of the continuum are more likely to score at least 2 than to score only 1. This observation makes it clear that the most useful method of mapping response categories onto underlying measurement variables is not obvious when items are scored in more than two categories.

The second method, illustrated in Fig. 8B, is based on the calculation of the expected score:

$$E_x = \sum_{h=0}^{m_i} hP_{ijh}. \tag{12}$$

In Fig. 8B, the abilities at which $E_x = 0.5$, 1.5, and 2.5 are identified. These are used to mark out an alternative set of response regions along the horizontal variable. An attraction of method B is that it is consistent with procedures for testing model-data fit that contrast a person's observed score on an item with his or her expected score. For expected scores between, say 1.5 and 2.5, the observed score that minimizes the observed-expected residual is a score of 2. In this sense, it might be argued that the 2 region of the measurement variable is best defined as that range of abilities for which the expected score is between 1.5 and 2.5. A disadvantage of this method is that it provides category regions that may seem implausibly wide.

The third method, illustrated in Fig. 8C, sums the curves in Fig. 8A. The curve labeled ≥ 1 is the result of summing curves 1, 2, and 3 in Fig. 8A. The curve labeled ≥ 2 is the result of summing curves 2 and 3. These cumulative ogives give the probability of scoring 1 or better, 2 or better, and 3 on this item. The abilities at which the cumulative probabilities equal 0.5 have been used to mark out a third set of regions. People with abilities in the 2 region in Fig. 8C will most probably (i.e., $p > 0.5$) score at least 2 on this item, but will most probably ($p > 0.5$) *not* score 3.

An attractive feature of this method is that it parallels the interpretation of variables constructed from dichotomously scored items. On a variable defined by dichotomous items, people's estimated location places them above items they will most probably pass ($p > 0.5$) and below items they will most probably fail ($p > 0.5$). Method C similarly interprets an individual's location by reference to thresholds they will most probably pass ($p > 0.5$) and thresholds they will most probably fail ($p > 0.5$).

Figure 8C also is helpful in distinguishing the partial credit model from models such as Samejima's graded response model. In the partial credit model, cumulative probabilities of the kind shown in Fig. 8C are not modeled directly but are the result of summing the category response functions in Fig. 8A, which in turn are the result of applying Rasch's model for dichotomies separately to each pair of adjacent categories. In the graded response model, cumulative probabilities *are* modeled directly. Thus, whereas the partial credit model gives the probability of person j scoring x on item i, the graded response model models the probability of person j scoring x or better on item i. As a result, the partial credit model is *not* a case of Samejima's graded response model (e.g., with equal item discriminations). This difference between the two models is sometimes misunderstood.

Parameter Estimation

Conditional and joint maximum likelihood procedures for estimating partial credit model parameters were described by Masters in 1982 and by Wright and Masters in 1982, who also describe a procedure based on the pairwise comparison of responses to items (PAIR) and a simplified procedure (PROX) based on the assumption that the effects of the person sample on item calibration and of the test on person measurement can be summarized by means and standard deviations on the variable. The essentials of the conditional and joint maximum likelihood procedures are summarized here.

Conditional Maximum Likelihood

The conditional maximum likelihood procedure begins with the conditional probability of the response vector (x_i) given test score r:

$$P\{(x_i); ((\delta_{ik}))|r\} = \frac{\exp\left(-\sum_i^n \sum_{k=0}^{xi} \delta_{ik}\right)}{\sum_{(x_q)}^r \exp\left(-\sum_i^n \sum_{k=0}^{xi} \delta_{ik}\right)}, \tag{13}$$

where $\delta_{i0} + 0$ and $\sum_{(x_q)}^r$ is the sum over all response vectors (x_q) that produce the score r.

The conditional probability of responding in category h of item i given score r is:

$$P^*_{irh} = \frac{\exp\left(-\sum_{k=0}^{h}\delta_{ik}\right)\sum_{x_q\neq i}^{r-h}\exp\left(-\sum_{q\neq i}^{n}\sum_{k=0}^{x_q}\delta_{qk}\right)}{\sum_{g=0}^{m_i}\left\{\exp\left(-\sum_{k=0}^{g}\delta_{ik}\right)\sum_{x_q\neq i}^{r-g}\exp\left(-\sum_{q\neq i}^{n}\sum_{k=0}^{x_q}\delta_{qk}\right)\right\}},$$
$$= \frac{\exp\left(-\sum_{k=0}^{h}\delta_{ik}\right)\gamma_{r-h,i}}{\gamma_r} \qquad (14)$$

where $\sum_{(x_{q\neq i})}^{r-h}$ is the sum over all response vectors $(x_{q\neq i})$ that exclude item i and produce the score $r-h$.

The conditional likelihood over N people with various scores is:

$$\Lambda = \prod_{j}^{N}\left[\frac{\exp\left(-\sum_{i}^{n}\sum_{k=0}^{x_{ij}}\delta_{ik}\right)}{\gamma_r}\right]$$
$$= \frac{\exp\left[-\sum_{j}^{N}\sum_{i}^{n}\sum_{k=0}^{x_{ij}}\delta_{ik}\right]}{\prod_{r}^{M-1}(\gamma_r)^{N_r}}, \qquad (15)$$

where $M = \sum_i^n m_i$ is the maximum possible score on the instrument:

$$\prod_{j}^{N}\gamma_r = \prod_{r}^{M-1}(\gamma_r)^{Nr}$$

and N_r is the number of people with a particular score r. The log likelihood can then be written:

$$\lambda = \log\Lambda = -\sum_{i}^{n}\sum_{k=1}^{m_i}S_{ik}\delta_{ik} - \sum_{r}^{M-1}N_r\log\gamma_r, \qquad (16)$$

where S_{ik} is the number of people scoring k or better on item i.

The estimation equations for the conditional maximum likelihood procedure require the first and second derivatives of the log-likelihood with respect to δ_{ih}. These are:

$$\frac{\partial\lambda}{\partial\delta_{ih}} = -S_{ih} - \sum_{r}^{M-1}\frac{N_r}{\gamma_r}\left(\frac{\partial\gamma_\lambda}{\partial\delta_{ih}}\right)$$
$$= -S_{ih} + \sum_{r}^{M-1}N_r\sum_{k=h}^{m_i}P_{irk} \qquad (17)$$

and

$$\frac{\partial^2\lambda}{\partial\delta_{ih}} = -\sum_{r}^{M-1}N_r\left(\sum_{k=h}^{m_i}P_{irk}\right)\left(1-\sum_{k=h}^{m_i}P_{irk}\right), \qquad (18)$$

where $\sum_{k=h}^{mi}P_{irk}$ is the probability of a person with score r scoring h or better on item i.

Joint Maximum Likelihood

The joint maximum likelihood procedure begins by modeling the likelihood of an entire data matrix (x_{ij}) as the continued product of the probabilities P_{ijx} over all people $j = 1, N$ and all items $i = 1, n$:

$$\Lambda = \prod_{j}^{N}\prod_{i}^{n}P_{ijx}$$
$$= \frac{\exp\sum_{j}^{N}\sum_{i}^{n}\sum_{h=0}^{x_{ij}}(\theta_j-\delta_{ih})}{\prod_{j}^{N}\prod_{i}^{k}\left[\sum_{h=0}^{m_i}\exp\sum_{h=0}^{h}(\theta_j-\delta_{ik})\right]}. \qquad (19)$$

The log-likelihood is:

$$\lambda = \log\Lambda = \sum_{j}^{N}\sum_{i}^{n}x_{ij}\theta_j - \sum_{j}^{N}\sum_{i}^{n}\sum_{h=1}^{x_{ij}}\delta_{ih}$$
$$-\sum_{j}^{N}\sum_{i}^{n}\log\left[\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h}(\theta_j-\delta_{ik})\right],$$

which can be rewritten:

$$\lambda = \sum_{j}^{N}r_j\theta_j - \sum_{i}^{n}\sum_{h=1}^{m_i}S_{ih}\delta_{ih}$$
$$-\sum_{j}^{N}\sum_{i}^{n}\log\left[\sum_{h=0}^{m_i}\exp\sum_{k=0}^{h}(\theta_j-\delta_{ik})\right], \qquad (20)$$

where r_j is the score of person j on the instrument, and S_{ih} is the number of people scoring h or better on item i.

The first derivatives of λ with respect to θ_j and δ_{ih} are:

$$\frac{\partial\lambda}{\partial\theta_j} = r_j - \sum_{i}^{n}\sum_{k=1}^{m_i}kP_{ijk} \qquad j = 1,\ N$$

and

$$\frac{\partial\lambda}{\partial\delta_{ih}} = -S_{ih} + \sum_{j}^{N}\sum_{k=h}^{m_i}P_{ijk} \qquad i = 1,n,\ \ h = 1,m_i.$$

The second derivatives of λ with respect to θ_j and δ_{ih} are:

$$\frac{\partial^2\lambda}{\partial\theta_j^2} = -\sum_{i}^{n}\left[\sum_{k=1}^{m_i}k^2P_{ijk} - \left(\sum_{k=1}^{m_i}kP_{ijk}\right)^2\right]$$

and

$$\frac{\partial^2\lambda}{\partial\delta_{ih}^2} = -\sum_{j}^{N}\left[\sum_{k=h}^{m_i}P_{ijk} - \left(\sum_{k=h}^{m_i}P_{ijk}\right)^2\right].$$

With these results, the joint maximum likelihood estimation equations for the partial credit model are:

$$\hat{\theta}_r^{t+1} = \hat{\theta}_r^{(t)} - \frac{r - \sum_i^n \sum_{k=1}^{m_i} kP_{irk}^{(t)}}{-\sum_i^n \left[\sum_{k=1}^{m_i} k^2 P_{irk}^{(t)} - \left(\sum_{k=1}^{m_i} kP_{irk}^{(t)} \right)^2 \right]} \tag{21}$$

and

$$\hat{\delta}_{ih}^{t+1} = \hat{\delta}_{ih}^{(t)} - \frac{-S_{ih} + \sum_r^{M-1} N_r \sum_{k=h}^{m_i} P_{irk}^{(t)}}{-\sum_r^{M-1} N_r \left[\sum_{k=h}^{m_i} P_{irk}^{(t)} - \left(\sum_{k=h}^{m_i} P_{irk}^{(t)} \right)^2 \right]}, \tag{22}$$

where $\hat{\theta}_r^{(t)}$ is the estimated ability of a person with score r on the n-item instrument after t iterations; $\hat{\delta}_{ih}^{(t)}$ is the estimate of δ_{ih} after t iterations; N_r is the number of people with score r; and M is the maximum possible score on the instrument (i.e., $M = \sum_i^n m_i$).

Of these two procedures, the conditional maximum likelihood is preferable on theoretical grounds. The joint procedure has the advantage of being relatively easy to program, and for this reason it is widely used in practice. The estimates it produces, however, contain a bias. Through simulation studies it has been shown that this bias—which is greatest for instruments with small numbers of items—can be significantly reduced by multiplying the final item estimates by $(n-1)/n$. For typical data sets, these corrected estimates are equivalent to those obtained from the conditional procedure.

See Also the Following Articles

Maximum Likelihood Estimation • Rasch, Georg

Further Reading

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* **43,** 561–573.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* **47,** 149–174.

Masters, G. N. (1988). The analysis of partial credit scoring. *Appl. Meas. Educ.* **1,** 279–297.

Masters, G. N., and Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika* **49,** 529–544.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (P. Zarempka, ed.) Academic Press, New York.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danmarks Paedogogiske Insitut, Copenhagen.

Rasch, G. (1977). On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. *Dan. Year Book Philos.* **14,** 58–94.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores.* Psychometrika Monograph Supplement no. 17.

Wilson, M., and Adams, R. J. (1993). Marginal maximum likelihood estimation for the ordered partition model. *J. Educ. Statist.* **18,** 69–90.

Wilson, M., and Masters, G. N. (1993). The partial credit model and null categories. *Psychometrika* **58,** 87–99.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis.* MESA Press, Chicago, IL.

Wu, M., Adams, R. J., and Wilson, M. (1998). CONQUEST computer program. Australian Council for Educational Research, Melbourne.

Participant Observation

Tamar Diana Wilson
University of Missouri, St. Louis, Missouri, USA

Glossary

advocacy The attempt to advance the interests and goals of the community or some of its members, often through networking them into services and goods provided by governmental or nongovernmental programs, agencies, and organizations.
cultural relativism The belief, ethical position, and approach that holds that cultures are diverse and unique in themselves and that behaviors and values must be understood in the cultural context in which they occur.
culture The shared patterns of thought, behavior, practice, lifestyles, and beliefs into which members of a society or community have been socialized; notably, cultures may be heterogeneous.
emic The insider or subjective viewpoint; emphasizes ideas and members' perceptions and interpretations of events, behaviors, relationships, and other phenomena of interest to the group or community.
ethnocentrism The assumption that the beliefs, practices, and lifestyle of the group to which one belongs are superior to those of other groups, who are consequently considered inferior.
ethnography The description and analysis of the culture or subculture (including behavioral patterns and lifestyles) of a group or community.
etic The outsider or objective viewpoint; emphasizes perceptions and interpretations of patterns of behavior and events using categories brought to the field by the researcher.
reciprocity The mutual exchange of goods and/or services and/or emotional support.
subculture The culture of a definable group or subsociety within a society or community.
work The generation of an income and/or the production or gleaning or processing of valued items for immediate use or sale.

Participation involves interacting, conversing, and sharing ongoing life experiences with members of a community. Observation of events, behavior, interactions, and conversations may be recorded in field notes or taped. Interviews may also be conducted and videos made. There are a number of continuums in the method, including that between observation and participation, that between passive participation and active participation, and that between active participation and advocacy/activism. Problems in participation—observation arise due to segregation along a number of axes in any given society and factionalisms that may divide it. Moving to insider status can involve essentialist characteristics of the researcher but also involves establishing commonalities and friendships, accepting fictive kinship designations, and participating in work and leisure activities with members of the community. Ethical issues include the principle of doing no harm, the necessity of being reflexive about one's impact and understandings gained, the importance of reciprocity relationships with community members, and the possibility of acting and advocating for the welfare of the community and its members.

The Method of Participant Observation

Development

Among the first to use participant observation as a data-gathering technique in anthropology were Bronislaw Malinowski in 1922 and Margaret Mead in 1928. Frank Cushing is also well credited with using the methodology, and even "going native" in his studies of the Zuni Pueblo in the 1880s. In the 1920s, the Chicago school, headed by Robert Ezra Park, sent sociologists into the field; they generated ethnographies on various aspects of the ethnically, occupationally, and socially diverse urban conglomeration of Chicago and elsewhere using methods of participant observation.

For most of the period following Malinowski and until the 1960s, participant observers stressed objective description and replicability, although subjective experience was sometimes taken into account and adopting a moral stance was not alien to early work. By the 1960s, emic and etic methods of approaching and gathering data had been elucidated and controversies arose regarding which was the best approach. In the 1980s, partially in response to earlier critical works, such as Vine Deloria's "Custer Died for Your Sins" and Edward W. Said's "Orientalism," as well as developing feminist critiques, ethnographers became more reflexive about their work and questioned the foundations and possibilities of complete "objectivity," often identifying its pursuit with colonialism, imperialism, gender discrimination, and ethnocentrism.

Definition

A working definition of participant observation as a methodology to gain information about and understanding of a group or community and create texts about their lives, behaviors, and beliefs (their "culture" or subculture) is as follows: Participant observation is (i) a qualitative methodology (which can put quantitative data into context and/or give it greater meaning); (ii) usually involving long-term research consisting of observing, interacting, conversing, sharing in work and leisure routines, and interviewing members of the community; and (iii) that takes an insider's (emic) point of view—often to build etic as well as emic models—to understand routine, everyday life or unique phenomena occurring within the setting(s) occupied by members of the group or community. In addition to utilizing a case study approach, participant observation consists of "a logic and process of inquiry that is open-ended, flexible, opportunistic, and requires constant redefinition of what is problematic, based on facts gathered in concrete settings of human existence" (Jorgenson, 1989, p. 14). An underlying assumption is that the researcher will be able to communicate with community members in their native language.

Participant observation is usually complemented by interviewing. Interviews can be unstructured, essentially guiding ordinarily occurring conversations toward topics of interest; semistructured, in which the interviewer openly suggests what he or she would like discussed; open-ended, with set questions designed to elicit a detailed response; or structured, in the form of a questionnaire administered to a sample of the population under study in a fixed format. This last form of interviewing may often lend itself to quantitative analysis. It also involves dialoguing with key informants—members of the community who lend themselves to supplying information about and interpreting local customs and perhaps attempting to socialize the researcher into meeting community expectations.

The results of participant observation are embodied in field notes, which document (i) the researcher's observations, conversations, interviews, interactions, and experiences as an insider or partial insider; (ii) emerging analysis and generalization; and, often, (iii) reflections on the researcher's role in and impact on the community and its members (reflexive ethnography). Field notes, once analyzed, form part of the basis of ethnographic understanding and are the foundation from which articles (or books) will be generated. Audio and video taping may also be used in capturing data.

Emics and Etics

Participation can lead to a greater emic (insider's, subjective, viewpoint) understanding; observation tends toward an etic (outsider's, objective, viewpoint) understanding. The emic point of view is concerned with the meaning the people ascribe to events, relationships, behaviors, and experiences. The etic point of view attempts to map patterns of behavior using categories identified not by insiders but through the perceptions and interpretations of the "outsider" often based on preexisting theory. Theoretical insights are thus brought to the field rather than being grounded in or "discovered" through the field experience. A completely emic view is often only an ideal because the researcher brings his or her own political and moral values into the field, which at a minimum will result in (unconscious) selective perception and recording of phenomena. Furthermore, hypotheses and categories for observation are often developed prior to entering the field, although they are subject to revision. On the other hand, a reliable and valid etic understanding, to not be ethnocentric, must take into account the emic viewpoint(s), which may be heterogeneous. (Viewpoints differ, for example, by gender, class position, or any number of achieved or ascribed statuses.) As Peter Berger pointed out in 1976, people can be assumed to be experts in their own lives and experts in defining the immediate problems (or joys and satisfactions) that confront them. Sometimes they even describe the larger context of their day-to-day problems: Etic approaches are useful in analyzing this context but may have already been appropriated by at least some members of the community; for example, the knowledge that the profit-making propensities of the factories they work for are responsible for their low wages or that Green Revolution packages are not adapted to the local agricultural regime.

In combating ethnocentrism, the belief in and ethical principle of "cultural relativism" is endorsed by most participant observers. There have been modifications in the theme of cultural relativism—which tends to reify "culture"—with the growing recognition (i) that behaviors, practices, norms, and values within any one culture may be heterogeneous due to differential access

to the means of production, power differences based on gender, ethnicity, and/or age and also the existence of subcultural groups; and (ii) that all cultures are and/or have been affected by world systemic processes (or processes of globalization), including colonialism, imperialism, capitalism, and consumerism.

Participant Observation Continuums

Some scholars equate passive participation with observation and distinguish three more interactive modes of participation: moderate (weighed toward observation and with only marginal membership), active (participation in as many events as possible with group or community members), and complete (becoming a full member of the group—"going native"). There are instances of all these modes of participation in ethnographies written by anthropologists and sociologists. However, degrees of participation/membership might also be conceived as occurring along three contiguous continuums. The first has as its poles "observation" and "participation" and runs from noninteractive "hanging out" and recording behavior one sees to participation as interacting, conversing, and interviewing with members of the community. The second continuum has as its poles "passive participation" and "active participation." Passive participation involves social interactions, dialoguing, conversing, and informal interviewing; active participation involves taking part in reciprocity networks, work routines, leisure activities, family and community events, and rites of passage and often being adopted as a core friend or fictive kinsman/kinswoman by one or more key informants. Active participation implies a greater socialization or enculturation into the society or subsociety under study. The third continuum runs from active participation to advocacy and activism. Advocacy and activism may involve independent input of effort and information, including networking community members into relevant government or nongovernmental organizations to help them or the community as a whole to obtain services or otherwise reach their goals (or even to envision new goals).

Problems Hindering Participant Observation

Segregation by Ascribed Status

Ascribed status includes gender, age, and ethnicity; they are considered essential and unchangeable characteristics of a person at any given point in time. They are unlike the achieved or acquired status of "college student," "warrior," or "engineer." Although assigning a role or a belief system to the holder of an ascribed status is considered essentialism, members of some groups and communities will do so. Some characteristics, such as marital status or parenthood, are on the ascriptive pole of an ascribed—acquired status continuum and are considered achieved statuses in others. Such characteristics can hinder (or help) access to people within a community who may conceive as oppositional statuses such as male vs. female, old vs. young, married vs. single, and so forth. It is well-known that within any given community male researchers may be denied access to women's spaces and activities, whereas female researchers may be blocked from entering men's spaces or watching men's activities. Even if the denial is not outright, the participant observer may be made to feel uncomfortable in settings reserved for the opposite sex. This is true not only in highly gender segregated societies, such as some in the Middle East, but also to some extent throughout the world (e.g., Australian women being expected not to enter men's segregated sections in public bars).

Many societies, especially tribal ones, are age graded as well. Informal segregation along the lines of same-aged peer groups is common in most societies. The participant observer may have easier access to people who are similar in age to his or her age. Martial status and parenthood may also affect the scope of interactions; some researchers have gotten more information and been more accepted by community members because they had children with them; others felt more restrained in the time they could devote to observation and participation because of child care responsibilities. Single researchers are often felt to be threatening and urged to find a spouse within the community. There are many instances, however, of ethnographers crossing into spaces normally segregated by age, gender, or any other ascribed (or acquired) characteristic.

Factionalism within the Group or Community

It has become well established that cultures are heterogeneous, dynamic, and changing, and that not all members of the community bearing the culture will have the same beliefs or habits. Within any community there may be opposing segments, parties, factions, or rivalries. Within the Mexican and Mexican American community of East Los Angeles, for example, neighborhood gangs may occupy streets that other gangs cross into only at risk. Political factionalism is common as well and can be found in such places as Mexican squatter settlements or rural population centers or in refugee camps throughout the world. Class factionalism is also common, whether in rural or urban communities. Often, the participant observer, because of the people he or she associates with or frequently interacts with, will be identified with the opposing faction and thus be denied access to or interactions with some people in the community.

Becoming an "Insider"

The Partial Insider

Partial insider is a rather essentialist term that has been used to describe a similarity of ethnicity or cultural background shared by the researcher and the people under study. On the one hand, the insider is only partially so due to differences in class (whether of origin or of actual class), work experience, education, overseas residence, urban or rural origin, gender, marital status, or sexuality. These differences may also cause the autoethnographer—the ethnographer studying his or her community of origin—to be a partial outsider as well, even being surprised by things that go on within his or her culture. This is specially true since (i) cultures are heterogeneous and (ii) and interpretations of events and behaviors in the same group can be highly variable, subject to change, and sometimes contradictory. On the other hand, commonalities may be established on the basis of any number of the many dimensions of ascribed or acquired status.

Establishing Commonalities

The ethnographer arrives in the field with a unique autobiography; over the life course the researcher has developed talents, acquired skills, and amassed bodies of knowledge and experiences that aid in adaptation to the field site and can be the basis for establishing communication though shared common interests and experience. The knowledge of wildlife, horticulture, livestock, boats, sports, mechanics, cooking, first aid, and so forth can aid in establishing rapport, facilitating discourse, and entering into reciprocity relationships with the community members.

There are at least five kinds of commonalities that ethnographers should seek in their fieldwork. First are the commonalities based on the hybridity, creolization, or interpenetration of cultures and the interconnectivity of the world. International migration has led to much primary interconnectedness as the mass media has led to much secondary interconnectedness. Thus, it would not be surprising, when doing fieldwork in North Africa, for example, if the researcher mentioned Paris, Marseilles, or Cadiz and a member of the community offered "I have a cousin in Paris" (or Marseilles or Cadiz) or even had been to one or more of these cities himself or herself. The same is true when mentioning Chicago, Los Angeles, or New York (among other places) in Mexico and Central or South America.

Second are commonalities based on ascriptive location (age, sex, race, and ethnicity) or personal situation (work, family, migration histories, marital status, and stage in the life cycle). Third are commonalities based on emotional experiences (joys and tragedies, celebrations and mournings, and births and bereavements). Fourth are commonalities that develop as the researcher is being socialized or enculturated into the community, as part of shared experiences. Fifth are commonalities based on sharing a common moral vision, what one can say "no" to; for example, torture, terrorism, and political repression; babies dying from lack of food or medical care; and children eating from garbage dumps. The establishment of such commonalities facilitates dialogue, conversation, interaction, empathy, and participation in events of importance to the researcher.

Friendship and Fictive Kinship

The researcher's presence is often considered intrusive and sometimes even "dangerous" to the normative order, especially since he or she is "out of place" in the sense of both not having a recognized status in the community and not knowing common courtesies, well-known social boundaries, or acceptable behaviors. Regarding the attempted normalization of "out-of-placeness," Freedman (1986, p. 54) recounts that upon her return as a widow to a village she had studied in Rumania, she was urged to remarry within the community because she was perceived as "a threat to the women and a temptation to the men."

People in a community may often try to regularize, tame, exorcise, or make sense of the anomalous role of the researcher by extending him or her the status of kinsman or kinswoman. The fictive kinship in which the researcher becomes involved may be *ad hoc* or a cultural institution, such as the practice of *compadrazgo* in Ibero-America as well as among other Catholic and some non-Catholic groups. In Latin America, becoming a godparent on the occasion of a baptism, a school graduation, a first communion, a coming out party, or a wedding (among other occasions) means that you have a responsibility for your godchild should his or her parents die, as well as reciprocal social and economic obligations with his or her parents, your compadres (coparents).

Male sociologists and anthropologists have been knitted into a community by members' informally adopting them as sons, brothers, fathers, or cousins and have been offered wives. Female researchers have been informally adopted as daughters and sisters and urged to find a husband with the group or community. Other researchers have gained entree into a community by being defined as a "friend" by a core member. Often, becoming friend or kin (however fictive) means becoming socialized into what are considered the site-specific or culturally unique requisites of such a role. In some cases, the researcher has found this to be a burden, and playing the kinship role correctly may limit access to other members of the community, especially of the opposite gender, or of antagonistic families. Such roles entail not only certain rights

but also obligations (including emotional upkeep), as do friendships and kinship ties worldwide.

Participation in Work and Leisure Activities

Commonalties can be established in the field by joining the work routines or play activities, including such sports as baseball or volleyball, fishing, dancing, horseback riding, and other pastimes that might fill part of the community members' leisure time. Working beside people can especially give knowledge into their daily lives and often opens up opportunities to know fellow workers and their families better. Fernández-Kelly worked in a *maquiladora* (an export-processing factory) in Cuidad Juarez, Chihuahua, Mexico, to gain insight into women employees' lives and to get to know them better. Kornblum worked in a steel mill, where many members of the South Chicago community he was studying (and living in) were employed. Phil DeVita established his credentials with the Acadian lobster fishermen through his mechanical and fishing abilities, part of the skills he brought into the field. Initially rejecting, community members eventually embraced him because of these abilities. In his words, "I became accepted as a mechanic and a fisherman, not as an anthropologist" (DeVita, 1992, p. 163). My work beside families of garbage pickers in Mexicali, corn farmers in Jalisco, and lettuce cutters in Salinas, California, gave me greater insight into the activities that occupied a good percentage of their waking lives; made me accepted as an "insider" who was willing to work alongside them, no matter how much drudgery it entailed; led me to understand the problems they faced at work both by experiencing it firsthand and because it more easily became a topic of conversation; and led to the families inviting me to events in their personal and social lives that extended beyond the workplace. Since I worked for free, and often added to the families' income by my efforts, I was also able to uphold my end in reciprocity obligations.

Ethical Issues

The Golden Rule

Ethical considerations in participant observation research are embodied in codes promulgated by the American Anthropological Association and the American Sociological Association. The golden rule is "to do no harm to the community or its members." This does not mean avoiding intervention to prevent one community member from doing harm to another, although some researchers believe they should not intervene in such incidents. It often means using pseudonyms for people and identifiable places. Sometimes, it means not publishing some of the information garnered. Covert research, although it appears to be the only means of securing information with some groups, is considered unethical by most scholars. Overt research means informing the group or community that one is doing research among them, informing them of the aims of the research, and explaining that the results of research may do them little good (which is true in most cases). Consent, sometimes written and sometimes not, is to be obtained from those who will be interviewed. The participant observer should also explain the aims of his or her study to people with whom he or she interacts in order to obtain information about the group or community. Assuring confidentiality and privacy is part of the "do no harm" rule.

Reflexivity

Reflexivity involves awareness of the impact of the researcher and the research on the community and self and comprises at least six aspects: (i) awareness of the possibility of ethnocentrism (i.e., the belief that the researcher's standards of behavior and values are superior to those of the group being studied) and the need to avoid it; (ii) awareness of one's positioning vis-à-vis the group under study, including differences in power and resources and differences (or similarities) in moral-political—ideological values; (iii) awareness of the impact the researcher has on the community in terms of changing community members' routines, interactions, behaviors, etc. (often the object of the ethnography); (iv) awareness that some of the information gleaned may reflect normative expectations rather than actual behavior—and this too gives insight into the community; (v) awareness of how any text produced might impact on the community or on its members; and (vi) awareness of how the field experience is also changing the researcher and his or her worldview, becoming a new chapter in autobiographical experience.

Reciprocity

Involvement in reciprocity relationships and networks is a means of establishing rapport, gaining friends, and consequently opening doors to more observation, participation, and information. Empathy and emotional ties (feelings of friendship) most usually develop in the course of reciprocity relationships. Reciprocity is also a way of paying back members of the group or community for their efforts on the researcher's behalf (and tolerance of his or her presence). At a minimum, reciprocity involves answering truthfully questions community members pose to the researcher in return for answering his or her questions. Many, if not most, participant observers have provided transportation; written letters; accompanied

individuals to offices, hospitals, or other institutionalized settings; extended first aid; filled in forms; and provided a variety of other services and goods. Reciprocity relations most often involve a flow of goods and services in both directions. Advocacy for the community or individuals within it is also a form of reciprocity. There is a moral imperative that at least some forms of reciprocity should be engaged in to "pay back" informants for their time and efforts on the researcher's behalf. Reciprocity is a basis for friendship in any case, within or outside of the field site(s).

Advocacy and Activism

Advocacy and activism concern the researcher's aiding the community to reach its goals or even envision new ones, such as education for its children. It may involve networking them into existing social services or governmental or nongovernmental organizations, programs, or agencies, such as orienting AIDS victims or drug addicts or abused women to existing treatment facilities. It may involve setting up new organizations or services, for example, engaging community members in building a school or a clinic. Those engaged in participant observation need not become advocates or activists for their group/community of study, although many do so and some have such plans before they enter the field. Others eschew advocacy or activism for fear of intruding on the daily routines and practices of the group whose life ways they wish to portray. Some participant observers take part in political rallies and organizational activities undertaken by the people; although possibly done merely to observe, the researcher may be asked for feedback, to hand out flyers, etc. (i.e., to become more of a participant). Opposing viewpoints on the responsibility for activism, and for taking a moral stance, have been documented in journals such as *Cultural Anthropology*. Researchers involved in reciprocity relationships may be expected to take some action on behalf of their friends or kinsmen.

See Also the Following Articles

Anthropology, Psychological • Ethical Issues, Overview • Ethnocentrism • Ethnography • Qualitative Analysis, Anthropology

Further Reading

Berger, P. L. (1976). *Pyramids of Sacrifice: Political Ethics and Social Change.* Anchor Books, Garden City, NY.

Davies, C. A. (1999). *Reflexive Ethnography: A Guide to Researching Selves and Others.* Routledge, New York.

Deloria, V. (1970). *Custer Died for Your Sins: An Indian Manifesto.* Macmillan, New York.

DeVita, P. R. (1992). Greasy hands and smelly clothes: Fieldworker or fisherman? In *The Naked Anthropologist: Tales from around the World* (P. R. Devita, ed.), pp. 156–164. Wadsworth, Belmont, CA.

DeWalt, K. M., and DeWalt, B. R. (2002). *Participant Observation: A Guide for Fieldworkers.* Altamira, Walnut Creek, CA.

Dumont, J. P. (1992). *The Headman and I: Ambiguity and Ambivalence in the Fieldworking Experience.* Waveland, Prospect Heights, IL.

Fernández-Kelly, M. P. (1983). *For We Are Sold, I and My People: Women and Industry on Mexico's Frontier.* State University of New York Press, Albany.

Freedman, D. C. (1986). Wife, widow, woman: Roles of an anthropologist in a Transylvanian village. In *Women in the Field: Anthropological Experiences* (P. Golde, ed.), pp. 333–358. University of California Press, Berkeley.

Golde, P. (ed.) (1986). *Women in the Field: Anthropological Experiences*. University of California Press, Berkeley.

Headland, T. N., Pike, K. L., and Harris, M. (eds.) (1990). *Emics and Etics: The Insider/Outsider Debate*. Sage, Newbury Park, CA.

Jorgenson, D. L. (1989). *Participant Observation: A Methodology for Human Studies.* Sage, Newbury Park, CA.

Kornblum, W. (1974). *Blue Collar Community.* University of Chicago Press, Chicago.

Malinowski, B. (1922). *Agronauts of the Western Pacific.* Dutton, New York.

Mead, M. (1928). *Coming of Age in Samoa.* Morrow, New York.

Nayaran, K. (1997). How native is a "native" anthropologist? In *Situated Lives: Gender and Culture in Everyday Life* (L. Lamphere, H. Ragoné, and P. Zavella, eds.), pp. 23–41. Routledge, New York.

Said, E. W. (1978). *Orientalism.* Pantheon, New York.

Sanjek, R. (ed.) (1990). *Fieldnotes: The Making of Anthropology*. Cornell University Press, Ithaca, NY.

Scheper-Hughes, N. (1995). The primacy of the ethical: Propositions for a militant anthropology. *Cultural Anthropol.* **36,** 409–420.

Path Analysis

Christy Lleras
Pennsylvania State University, University Park, Pennsylvania, USA

Glossary

endogenous variable A variable whose variation is explained by one or more variables within the model.
exogenous variable A variable whose variation is explained by factors outside the model and which also explains other variables within the model.
path analysis A statistical method used to examine hypothesized (causal) relationships between two or more variables.
recursive model A path model where all the causal relationships flow in a single direction with no reciprocal effects or feedback loops.
specification error Error that occurs when significant causal variables are excluded from the path model.
spurious effects When part of the association between two variables is due to shared causal influences.
standardized path coefficient Coefficients that have been converted into standardized *z*-scores which allow researchers to compare the relative magnitude of the effects of different explanatory variables in the path model.

Path analysis is a statistical technique used primarily to examine the comparative strength of direct and indirect relationships among variables. A series of parameters are estimated by solving one or more structural equations in order to test the fit of the correlation matrix between two or more causal models, which are hypothesized by the researcher to fit the data.

Introduction

Path Analytic Methods

One of the primary goals of social scientific research is to understand social systems through the explication of causal relationships. However, given the complexity of social life, disentangling the interrelationships among variables is often a difficult task. Path analysis is a methodological tool that helps researchers using quantitative (correlational) data to disentangle the various (causal) processes underlying a particular outcome. The path analytic method is an extension of multiple regression analysis and estimates the magnitude and strength of effects within a hypothesized causal system. In addition, path analysis can be used to test the fit between two or more causal models, which are hypothesized by the researcher to fit the data.

Since path analysis assesses the comparative strength of different effects on an outcome, the relationships between variables in the path model are expressed in terms of correlations and represent hypotheses proposed by the researcher. Therefore, the relationships or pathways cannot be statistically tested for directionality and the models themselves cannot prove causation. However, path models do reflect theories about causation and can inform the researcher as to which hypothesized causal model best fits the pattern of correlations found within the data set. One of the advantages of using path analysis is that it forces researchers to explicitly specify how the variables relate to one another and thus encourages the development of clear and logical theories about the processes influencing a particular outcome. Path analysis is also advantageous in that it allows researchers to break apart or decompose the various factors affecting an outcome into direct effects and indirect components.

History of Path Analysis

Path analysis was originally developed by geneticist Sewall Wright in the 1920s to examine the effects of hypothesized models in phylogenetic studies. Wright's

analysis involved writing a system of equations based on the correlations among variables influencing the outcome and then solving for the unknown parameters in the model. According to Wright, the path analytic method was intended to measure "... the direct effect along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause." Wright also acknowledged the fact that often causal relations were uncertain and cautioned that this method was not intended to deduce causal relations simply from correlation coefficients. Rather, the method utilized information provided by the statistical correlations in conjunction with qualitative information regarding the causal relationships to find the consequences of hypothesized structures.

Several decades later, path analysis was introduced into social scientific research by Blalock, Duncan, and others (Boudon and Turner). Sociologists Peter Blau and Otis Dudley Duncan were among the first to utilize path analysis extensively in their research on the processes involved in status attainment. In their book, *The American Occupational Structure*, Blau and Duncan utilized data collected from a sample of adult males and their parents to develop path models of the causal processes underlying educational and occupational outcomes.

During the 1970s, path analysis became even more popular and numerous papers were published featuring path analytic methods in sociology, as well as psychology, economics, political science, ecology, and other fields. Since the early 1980s, path analysis has evolved into a variety of causal or structural equation modeling (SEM) programs and computer packages. Unlike earlier path models, which were based on least squares regression, these new methods of causal modeling utilize the general linear model approach. The advantages of these new approaches are discussed below in "Extensions and Computer Software."

Elements of Path Models

The Path Diagram

Social scientific theories of causal relationships often specify a system of relationships in which some variables affect other variables and these in turn influence still other variables in the model. A single multiple regression model can only specify one response variable at a time. However, path analysis estimates as many regression equations as are needed to relate all the proposed theoretical relationships among the variables in the explanation at the same time.

To illustrate, consider the following hypothesis that a child's educational attainment is directly affected by family background, as well as individual academic achievement and engagement in school. In addition, academic achievement and school engagement depend on both the mother's educational level and parental income. Further, the level of student engagement in school is hypothesized to affect student achievement and, in turn, influence educational attainment.

A path diagram represents the hypothesized causal model in path analysis. To illustrate this hypothesis regarding educational attainment, the basic elements of a path diagram are depicted in the model shown in Fig. 1. The single straight arrows indicate a causal relationship, leading from the explanatory (causal) variable to the outcome variable (effect). For example, educational attainment is dependent upon mother's education, parental income, school engagement, and achievement. A child's level of school engagement is also dependent upon maternal education and parental income. Finally, student achievement is influenced by three factors: mother's education, parental income, and school engagement. The double-headed curved arrows linking maternal education and parental income indicate that the two

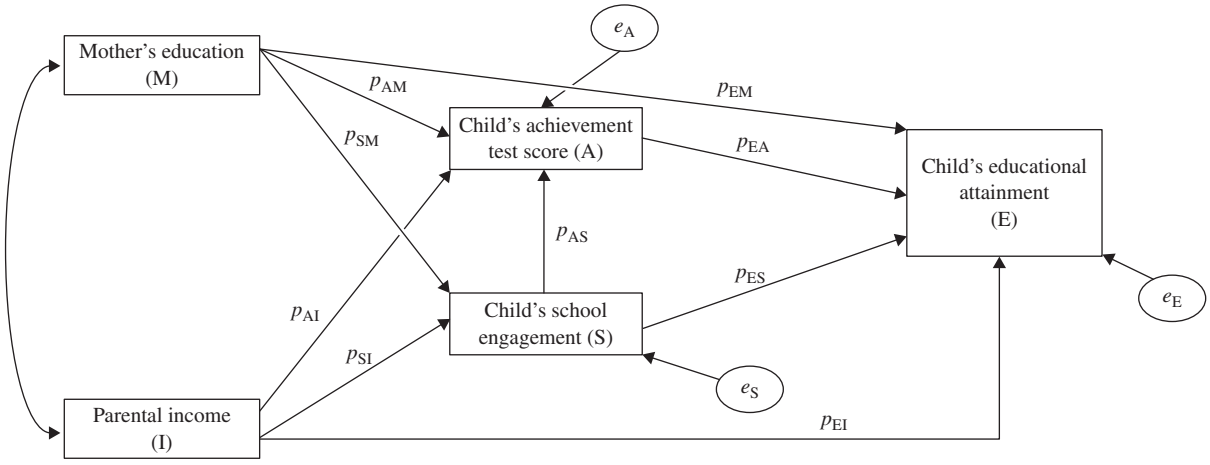


Figure 1 Path diagram for hypothesized model predicting child's educational attainment.

variables may be related, but no prediction is being made in the model as to the direction of the association.

Exogenous and Endogenous Variables

Variables often play more than one role in path models and this is reflected in the analytic language used in path analysis. Exogenous variables are variables whose cause is external to the model and whose role is to explain other variables or outcomes in the model. In Fig. 1, for example, the model does not explain the variability in maternal educational level and parental income. However, these exogenous variables are hypothesized to account for differences in child's achievement, school engagement, and educational attainment.

Endogenous variables are variables that are caused by one or more variables within the model. Endogenous variables have incoming arrows and can include outcome variables (only incoming arrows) and intervening causal variables. Endogenous variables, such as educational attainment, have only incoming arrows. School engagement and achievement are called intervening endogenous variables since they have both incoming and outgoing arrows. The hypothetical model in Fig. 1 indicates that engagement and achievement are influenced by other variables in the model (e.g., maternal education, parental income), and in turn have an effect on educational attainment.

Residual Error

Residuals or error terms (represented by e) are exogenous independent variables that are not directly measured and reflect unspecified causes of variability in the outcome or unexplained variance plus any error due to measurement. They are depicted in the diagram by arrows connecting the error terms with their respective outcome or endogenous variables. Residual error is assumed to have a normal distribution with a mean of zero and to be uncorrelated with other variables in the model. Note error terms are not always uncorrelated.

Path Coefficients

Although not required, path models often report the standardized regression coefficients (beta) or estimated path coefficients that have been converted into standardized z-scores, for each causal path depicted in the model. Standardized coefficients allow researchers to compare the relative magnitude of the effects of different explanatory variables in the path model by adjusting the standard deviations such that all the variables, despite different units of measurement, have equal standard deviations. These standardized path coefficients measure the relative strength and sign of the effect from a causal variable to an endogenous or outcome variable in the model. When more than one causal variable is present in the model, the standardized path coefficients represent partial regression coefficients that measure the effect of one variable on another, controlling for prior variables.

The subscripts for each pathway (p_{ij}) describe the causal relationship being estimated in the model. The first subscript (i) is the outcome variable and the second subscript (j) is the causal variable or variable whose influence on the outcome is under consideration. In Fig. 1, M, I, A, S, and E denote mother's education, parental income, achievement test score, school engagement, and educational attainment, respectively. The partial standardized regression coefficient depicted by the pathway p_{EA}, for example, is the estimated effect of child's achievement test score on educational attainment, controlling for family background and school engagement. The model hypothesizes that an increase in achievement test scores, holding parental income, maternal education, and engagement in school is associated with an increase in educational attainment. Similarly, the pathways p_{AM} and p_{SM} denote the standardized regression coefficients for the effect of mother's education on child's achievement and school engagement respectively, holding the antecedent variable, parental income, constant.

Structural Equations

Since the path analytic method follows the usual assumptions of ordinary least squares regression, all the relationships depicted in Fig. 1 are assumed to be linear, additive, and causal. Therefore, the model can be specified by a series of path or structural equations that describe the direct causal relationships between the variables. There are three endogenous or outcome variables in the model, so there are three sets of standardized coefficients to be estimated using ordinary least squares regression. In path models, to solve for the direct effects, each endogenous variable is regressed on all the variables with direct paths leading to it. Thus, the set of hypothesized direct causal relationships depicted in Fig. 1 correspond to the following path equations:

$$\begin{aligned}\text{Educational Attainment} &= p_{\mathrm{EM}}\mathrm{M} + p_{\mathrm{EI}}\mathrm{I} + p_{\mathrm{ES}}\mathrm{S} \\ &\quad + p_{\mathrm{EA}}\mathrm{A} + e_{\mathrm{E}} \qquad (1)\\ \text{Child's Achievement} &= p_{\mathrm{AM}}\mathrm{M} + p_{\mathrm{AI}}\mathrm{I} \\ &\quad + p_{\mathrm{AS}}\mathrm{S} + e_{\mathrm{A}} \qquad (2)\\ \text{Child's School Engagement} &= p_{\mathrm{SM}}\mathrm{M} + p_{\mathrm{SI}}\mathrm{I} + e_{\mathrm{S}} \qquad (3)\end{aligned}$$

In this model, educational attainment (Eq. 1) depends on the following partial regression coefficients: mother's education, parental income, school engagement, and achievement. Achievement (Eq. 2) is influenced by mother's education, parental income, and the level of

engagement in school. School engagement (Eq. 3) depends only upon mother's education and parental income. Each outcome or endogenous variable also has a residual path or error term (e) associated with it (i.e., e_E, e_A, e_S), which represent the variation left unexplained by the explanatory variables in the path model.

Decomposition of Path Effects

One of the unique contributions of path analysis to social scientific research is its ability to decompose the associations between several variables into causal (direct and indirect) and noncausal (e.g., spurious) components.

Direct and Indirect Causal Relationships

Causal relationships between variables may consist of direct and indirect effects. Direct causal effects are effects that go directly from one variable to another. Indirect effects occur when the relationship between two variables is mediated by one or more variables. For example, in Fig. 1, school engagement affects educational attainment directly and indirectly via its direct effect on achievement test score. Maternal education and parental income also have indirect effects on both achievement and educational attainment. Their indirect effects on achievement occur through their direct effects on school engagement. Their indirect effects on educational attainment occur through their influence on school engagement, through their influence on achievement, and through their effects on achievement and engagement, combined.

The magnitude of the indirect effects is determined by taking the product of the path coefficients along the pathway between the two causally related variables. Thus, the total indirect effect between two variables in a path model equals the sum of the products of each indirect effect. For example, child's school engagement affects educational attainment indirectly through its effect on achievement. Thus, the magnitude of the indirect effect between engagement and attainment can be estimated by multiplying the paths from school engagement to achievement and from achievement to educational attainment, $(p_{EA} \times p_{AS})$.

Calculating the total indirect effect between mother's education and child's educational attainment is a bit more complicated but follows the same logic. Maternal education affects educational attainment indirectly through child's achievement and the magnitude of the indirect effect is $(p_{EA} \times p_{AM})$. Maternal education also indirectly influences educational attainment via child's school engagement and the magnitude of the effect is $(p_{ES} \times p_{SM})$. In addition, mother's education influences child's educational attainment both through its effect on school engagement and on achievement. The magnitude of this indirect effect is $(p_{EA} \times p_{AS} \times p_{SM})$. Thus, the total indirect effect of mother's educational attainment on child's educational attainment is the sum of all of these indirect effects, $(p_{EA} \times p_{AM}) + (p_{ES} \times p_{SM}) + (p_{EA} \times p_{AS} \times p_{SM})$. Since mother's education is also correlated with parental income, all of these indirect effects also occur via this correlation.

Noncausal Relationships

In addition to decomposing relationships between variables into direct and indirect effects, path models also break down effects into noncausal components. Noncausal relationships may be spurious or due to unanalyzed prior associations. Spurious noncausal effects occur when the relationship between two endogenous variables is being influenced by a third variable. In other words, part of the association between two variables is due to shared causal influences. In the hypothesized model of educational attainment in Fig. 1, part of the relationship between child's school engagement and achievement reflects such a spurious effect. School engagement and achievement are influenced by both mother's education and by parental income. Thus, part of the association between school engagement and achievement is due to the fact that both variables are influenced by mother's education and both are also influenced by parental income.

Estimation and Testing

Estimation of Path Models

The previous example used to illustrate the elements of the path diagram and path decomposition depicts a recursive or unidirectional causal flow model. The model is recursive since all the causal linkages flow in one direction and none of the variables represent both cause and effect at the same time. This causal model is the only kind of model which can properly be called path analysis. In models where the hypothesized causality flows in a single direction, the estimation can be done relatively simply by using ordinary least squares (OLS) regression or maximum likelihood estimation (MLE) to solve the equations for each endogenous or outcome variable in the model. For a discussion of nonrecursive models, refer to "Extensions and Computer Software" below.

Model Specification

Path analysis is particularly sensitive to model specification. The inclusion of irrelevant variables or the exclusion

of important causal variables changes the value of path coefficients. In addition to recursivity, it is also assumed that all causally relevant variables have been included in the model. Specification error occurs when significant causal variables are excluded from the model. When this type of error occurs the value of the path coefficients reflect the common variance shared with these omitted variables. Since the strength of direct and indirect effects on the outcome variables in the model are evaluated using the path coefficients, the decision of whether or not to include different variables in the path model is critical to the interpretation of the underlying (causal) processes. Therefore, this method is most helpful in the testing of well-specified theories about the relationships between variables and not for exploratory purposes. Additional details on the assumptions in path analysis can be found in the literature by Kline and by Maruyama.

Data Demands

Since all the pathways or relationships must be capable of being estimated using multiple regression, the path analytic method requires the use of interval level data for all the variables included in the model. However, dichotomous and ordinal variables are commonly used in path analysis. For a review of literature on this topic see Jaccard and Wan (1996). In addition, path analysis only deals with measured variables or variables that can be observed and measured directly.

Goodness of Fit

Often researchers have more than one theory regarding which variables or what paths to include in the path model. Competing theories can be evaluated by estimating separate path models and then assessing the goodness-of-fit statistics to determine which hypothesized model best fits the correlation matrix in the observed data. Alternative theories can also be combined into a single path model and the researcher can assess which pathways are more significant by comparing the relative strength of different pathways within the same path model.

There is a variety of goodness of fit statistics that can be used to assess model fit and evaluate competing path models. The different computer programs commonly used to estimate path models generate a number of fit statistics as part of their output. While there is some disagreement over which specific tests are the "best" to use, it is commonly recommended that researchers examine more than one fit statistic when evaluating model fit. For example, Kline recommends using at least the following four tests: χ^2; GFI, NFI, or CFI; NNFI; and SRMR. For a discussion of these indexes and others refer to the work by Bollen and Long and by Jaccard and Wan.

Strengths and Limitations

Strengths of Path Analysis

The questions that are the subject of social inquiry often involve multiple causal influences. In order to explain a particular outcome it is therefore necessary to examine both the direct and indirect relationships among variables within a hypothesized model. As noted previously, one of the strengths of the path analytic method is that it estimates a system of equations that specify all the possible causal linkages among a set of variables. In addition, path analysis enables researchers to break down or decompose correlations among variables into causal (i.e., direct or indirect) and noncausal (e.g., spurious) components. Thus, path analysis helps researchers disentangle the complex interrelationships among variables and identify the most significant pathways involved in predicting an outcome.

Path analysis can also play a vital role in the theoretical or hypothesis testing stage of social research. While certainly experimental designs involving the random assignment of individuals to either a treatment or control group is the best way to test for causal effects, these experiments are often impossible to conduct given the sorts of questions that are the subject of social scientific inquiry. Since path analysis requires researchers to explicitly specify how they think the variables relate to one another within the path diagram, this method forces researchers to develop detailed and logical theoretical models to explain the outcome of interest. Thus, researchers using nonexperimental, quantitative, or correlational data can test whether their hypotheses about the relationships between variables are plausible and supported by the data and represent underlying (causal) processes.

Limitations of Path Analysis

While many researchers espouse the benefits of using path analysis in quantitative social research, the technique also has its critics. Since path analysis is an extension of multiple regression, it follows all the usual assumptions of regression. However, it is often difficult to meet these assumptions in social scientific research, particularly those of reliability and recursivity or unidirectional causal flow. The path model has to assume that each variable is an exact manifestation of the theoretical concepts underlying them and reasonably free of measurement error. In addition, the causality in the hypothesized model has to flow in one direction (no feedback loops or bidirectional causality); otherwise the model cannot be solved with ordinary least squares regression techniques. As will be discussed in the last section, however, methods have developed out of path analysis that can deal with these limitations of the least squares approach.

Finally, path analysis is a statistical tool used to evaluate whether the correlations between variables in a given data set reflect the causal hypotheses specified in the model. Since the models are based on correlations, path analysis cannot demonstrate causality or the direction of causal effects. However, as stated previously the path analytic method can indicate which of the path models best fits the pattern of correlations found in the data.

Extensions and Computer Software

Structural Equation Modeling

The limitations of the least squares method in path analysis discussed above led to the development of the general linear modeling (GLM) approach. This approach has evolved over the past 20 years into a variety of structural equation modeling (SEM) programs and computer packages. The main advantages of this approach is that it provides better measures of the theoretical constructs underlying variables and can estimate not only traditional recursive path models, but also nonrecursive models, models with measurement error, and models with unobserved variables (for further discussion see the literature from Maruyama.

The path analytic method underlies the structural equation modeling (SEM) approach, however SEM provides a more powerful alternative to path analysis and other regression techniques (e.g., multiple regression, time series). SEM allows more flexible assumptions including explicitly modeling correlated error terms, interactions, nonlinearities, and data level. In addition, while path analysis deals only with measured variables, SEM can model latent variables that cannot be directly observed in the data but rather are inferred from measured variables (i.e., factor analysis). The use of multiple indicators of a construct helps to reduce measurement error and increase data reliability.

Computer Software

Currently, path analysis can easily be conducted with any one of the following SEM computer programs: LISREL (e.g., Jöreskog and Sörbom), AMOS (e.g., Arbuckle), or EQS (e.g., Bentler). Once the model and data are entered into the program, the regression equations in the path model(s) are estimated simultaneously. The output includes the unstandardized and standardized regression or path coefficients, as well as a variety of goodness-of-fit statistics, which can be used to compare the fit of different hypothesized models against the correlation matrices, observed in the data.

LISREL is commonly used in sociology and throughout the social sciences and utilizes a programming language called SIMPLIS. AMOS is a more recent program that allows researchers to utilize either a user-friendly graphical interface to draw path diagrams or a programming language called BASIC, to estimate path models. An advantage of AMOS is that it produces high quality path diagrams and reads SPSS system files. Both LISREL and AMOS are put out by SPSS. Another program called EQS uses a graphical interface and runs the model based on the diagram drawn by the researcher using a drawing tool called *Diagrammer*. For a discussion of these and other SEM programs see the literature from Kano or Kline.

See Also the Following Article

Structural Equation Models

Further Reading

Alwin, D., and Hauser, R. (1975). The decomposition of effects in path analysis. *Am. Soc. Rev.* **40,** 37–47.

Bielby, W., and Hauser, R. (1977). Structural equation models. *Ann. Rev. Soc.* **3,** 137–161.

Blalock, H., Jr. (1964). *Causal Inferences in Non-Experimental Research.* University of North Carolina Press, Chapel Hill, NC.

Blau, P., and Duncan, O. (1967). *The American Occupational Structure.* Wiley, New York.

Bollen, K., and Long, S. J. (1993). *Testing Structural Equation Models.* Sage, Newbury Park, CA.

Boudon, R. (1965). A method of linear causal analysis: Dependence analysis. *Am. Soc. Rev.* **30,** 365–374.

Duncan, O. (1966). Path analysis: Sociological examples. *Am. J. Soc.* **72,** 1–16.

Everitt, B. S., and Dunn, G. (1991). *Applied Multivariate Data Analysis.* Arnold, London.

Jaccard, J., and Wan, C. K. (1996). *LISREL Approaches to Interaction Effects in Multiple Regression.* Sage, Thousand Oaks, CA.

Kline, R. (1998). *Principles and Practice of Structural Equation Modeling.* Guilford Press, New York.

Land, K. (1969). Principles of path analysis. In *Sociological Methodology* (E. F. Borgatta, ed.), pp. 3–37. Jossey–Bass, San Fransisco.

Maruyama, G. (1998). *Basics of Structural Equation Modeling.* Sage, Thousand Oaks, CA.

Wright, S. (1934). The method of path coefficients. *Ann. Math. Statist.* **5,** 161–215.

Pearson, Karl

M. Eileen Magnello

University College London/Wellcome Trust Center for the History of Medicine, London, United Kingdom

Glossary

biserial correlation Pearson devised this technique in 1909 to measure a linear relationship between one continuous variable and one dichotomous variable on the assumption that the underlying dichotomous variable is continuous and normally distributed.

chi-square goodness-of-fit test This test seeks to determine whether the observed distribution (constructed from observational data) conforms to the theoretical distribution with a "correction" of $n-1$ as termed by Pearson (or "degrees of freedom" later used by R. A. Fisher). Pearson considered first cases in which the theoretical probability is known a *priori*, where

$$\chi^2 = S\frac{(m' - m_s - \mu)^2}{m_s},$$

where m' is the observed (or empirical) frequencies in a distribution, m_s is the theoretical (or expected) distribution known *a priori*, μ is the population mean, and S is summation. A more contemporary formula for the chi-square goodness-of-fit test is $\chi^2 = \Sigma(O-E)^2/E$, where O is the observed values, E is the expected values, and Σ is summation.

chi-square test of independence Established by Pearson in 1904 to measure the differences between observed cell frequencies and expected cell frequencies. The chi-square statistic, as renamed by R. A. Fisher in 1923, is one of the most commonly used statistical tests to measure the association between two discrete (usually nominal) variables for manifold contingency tables. Pearson found the value of χ^2 as

$$\chi^2 = S\frac{(n_{uv} - v_{uv})^2}{v_{uv}},$$

where n_{uv} is the theoretical (expected) frequencies in cells, v_{uv} is the observed (empirical) frequencies in cells, and S is summation.

chi-square tests The statistical framework of Pearson's chi-square is a tripartite system because it not only incorporates a probability distribution and a goodness-of-fit test that he devised in 1900 but also includes a statistical technique for the analysis of contingency tables, which he introduced in 1904.

coefficient of variation Created by Pearson in 1896 as a standardized method for measuring the percentage of variation in the mean, where the standard deviation is treated as the total variation of the mean. Thus, C of $V = \bar{X}/\mathrm{SD} \times 100$.

method of moments Derived from Clapeyron's theorem of the three moments in mechanics, Pearson found that the covariance (Σxy) corresponded to the product moment of dynamics and the standard deviation (σ) corresponded to the moment of inertia. Pearson adopted the method of moments in 1892 for curve fitting of symmetrical and asymmetrical distributions by defining the following statistical parameters: central tendency (the mean), variability (the square of the standard deviation or the "variance"), skewness, and kurtosis.

multiple correlation Underpinned by matrix algebra, which was created by Cambridge mathematicians in the mid-19th century and which Pearson introduced into statistical theory to calculate the multiple correlation coefficient r in 1896, it measures the linear relationship between one dependent continuous variable and a set of independent continuous variables.

Pearson product–moment correlation coefficient A technique used to measure a linear relationship between two continuous variables whose values range from -1 to $+1$. Pearson's formula of the product–moment correlation coefficient is

$$r = \frac{\sum xy}{(S_X)(S_Y)} = \frac{\text{covariance}}{(\text{standard deviation of } x)(\text{standard deviation of } y)}.$$

phi coefficient This method was designed for so-called point distributions, which implied that two variables have two

points that are represented by some qualitative attribute for a 2 × 2 contingency table. This measures the relationship between two dichotomous variables that do not assume an underlying normal distribution. After finding the standard deviation, Pearson calculated the correlation between errors of the mean for two sets of variables. This led to his phi coefficient, the values of which range from −1 to +1:

$$r_{hk} = \frac{ad - bc}{\sqrt{(b+d)(a+c)(c+d)(a+b)}}.$$

regression A statistical system used for the linear prediction of two continuous variables. Pearson introduced the terms *dependent variable* and *independent variable* in 1896; this was an essential distinction to make for regression because the independent variable is the predictor and the dependent variable is the criterion. Pearson then showed that $Y' = a + bX$ was the equation for the regression (or predicted) line, and that the regression coefficients could be determined by finding the covariance of the variables x and y and dividing that by the variance of x. The regression coefficient is thus

$$b = \frac{\sum xy}{S^2 x} = \frac{\text{covariance}}{\text{variance of } x}$$

where X is the independent variable. The constant $a = \bar{Y} - b\bar{X}$.

standard deviation A technique for measuring variation at all points on a distribution and involves taking the square root of the deviational values squared and dividing by the total number in the sample. Pearson defined the standard deviation in 1893 as the error of the mean square, where $S = \sqrt{(\sum x^2/N)}$.

standard error of estimate Following Pearson's work on regression and correlation, he provided the standard deviation of correlation and also introduced the standard error of estimate in 1896 as the standard deviation of the regression coefficient:

$$\sigma\sqrt{(1-r)^2}.$$

tetrachoric correlation This measures the association between two continuous but dichotomous variables that are normally distributed and linearly related. Pearson devised this technique in 1900 when he began to analyze variables that could not be simply classified as continuous. Pearson derived the value of the tetrachoric correlation r from the value of the correlation parameter ρ of the bivariate normal distribution with frequencies from each of the four quadrants of the x, y plane by lines parallel to the coordinate axes. This division agreed exactly with four cell frequencies of the fourfold table in terms of arbitrary and precise dichotomous division of two discrete variables. The values of the tetrachoric correlation ranged from −1 to +1 and could be found by

$$r = \sin 2\pi \frac{(ad - bc)}{N^2} = \cos \pi \frac{b}{a+b}.$$

Pearson did not explain why he used the trigonometric functions of sine and cosine. In more contemporary terms, the tetrachoric correlation may be expressed as

$$r_{\text{tet}} = \frac{\cos 180^\circ}{1 + \sqrt{(BC/AD)}}.$$

Karl Pearson, Cambridge-trained English mathematician, philosopher, and statistician, was born in London in March 1857 and died at Coldharbour, Under Dorking, Surrey, in April 1936. Pearson's prodigious and innovative publications combined with a vast reservoir of energy and determination enabled him to create the discipline of mathematical statistics. He also established many institutional changes at University College London, including the creation of a department of structural (now civil) engineering, a department of astronomy with two observatories, and his department of applied statistics.

Introduction

Karl Pearson was one of the principal architects of the modern theory of mathematical statistics, or what he also termed biometrics. He was a polymath whose interests ranged from astronomy, mechanics, meteorology, and physics to the biological sciences in particular (including anthropology, eugenics, evolutionary biology, heredity, and medicine). He was also interested in the study of German folklore and literature, the history of the Reformation, and German humanists (especially Martin Luther). In addition to these activities, he also contributed hymns to the "Socialist Song Book." Pearson's writings were voluminous: He published more than 650 papers in his lifetime, of which 400 are statistical. Over a period of 28 years, he founded and edited six journals and was a cofounder (along with W. F. R. Weldon and Francis Galton) of the journal *Biometrika*. The main set of Pearson's collected papers, which consist of 235 boxes containing family papers, scientific manuscripts, and 16,000 letters, have been reposited at University College London (UCL).

Due mainly to his interests in evolutionary biology, Pearson created, almost single-handedly, the modern theory of statistics in his Biometric School at University College London from 1892 to 1903, which was practiced in the Drapers' Biometric Laboratory from 1903 to 1933. These developments were underpinned by Charles Darwin's ideas of biological variation and "statistical" populations of species, and they arose from the impetus of the statistical and experimental work of his colleague and closest friend, the Darwinian zoologist W. F. R. Weldon (1860–1906). This work led to his development of goodness-of-fit tests for asymmetrical

curves in 1892 and culminated in his chi-square goodness-of-fit test in 1900. Additional developments emerged from Francis Galton's (1822–1911) law of ancestral heredity. Pearson later devised a separate methodology for problems of eugenics, based on family pedigrees and actuarial death rates, in the Galton Eugenics Laboratory from 1907 to 1933.

In his creation of the emerging discipline of mathematical statistics, Pearson introduced a new vernacular for statistics, including such terms as standard deviation, mode, homoscedasticity, heteroscedasticity, kurtosis, and the product–moment correlation coefficient. Like a number of scientists at the end of the 19th century, Pearson was interested in the developing etymology in various disciplines, especially biology. Although he attempted to coin a number of biological words, the only word that survived him is "siblings," which he used "to cover a group of brothers and sisters regardless of sex."

Family and Education

The second son of William Pearson and Fanny Smith, Carl Pearson was born in London on March 27, 1857. The University of Heidelberg changed the spelling of his name in 1879 when he was enrolled as "Karl Pearson," and 5 years later he adopted this variant of his name and eventually became known as "KP." His mother came from a family of seamen and mariners and his father was a barrister and QC. The Pearsons were of Yorkshire descent because most of their ancestors came from the North Riding. They were a family of dissenters and of Quaker stock. By the time he was in his 20s, Pearson had rejected Christianity and had become a Freethinker, which involved the "rejection of all myths as explanation and the frank acceptance of all ascertained truths to the relation of the finite to the infinite." Although he did not regard himself as an atheist, "he vigorously denied the possibility of a god ... because the idea of one and all of them by contradicting some law of thought involves an absurdity." To Pearson, "religion was the relation of the finite to the infinite." Politically, he was a socialist whose outlook was similar to the Fabians, but he never joined the Fabian Society (despite requests from Sidney and Beatrice Webb). Socialism was a form of morality for Pearson; the moral was social and the immoral was antisocial in conduct.

There were a number of solicitors in the Pearson family, including William's brother Robert and Robert's son Hugh, as well as William's eldest child, Arthur, all of whom read law at the Inner Temple. William was a very hardworking and taciturn man who was never home before 7 PM; he continued to work until about midnight and was usually up at 4 AM reviewing his briefs. To both of his sons, William emphasized the importance of hard work quite regularly and especially once they were at Cambridge. Only during the holidays did the children spend any time with their father. In a letter to Karl, his elder brother Arthur described the experience of being home with their father as "simply purgatory ... the governor never spoke a word." In this desolate atmosphere, with her husband working incessantly and never talking to anyone when he was home, Fanny was deeply unhappy in her marriage. Thus, she transferred her love to her two sons, and she was deeply affectionate to Karl who was, without doubt, her favorite child.

For a short time in 1866, both boys received tuition at home from a Mr William Penn, who had started a school near Harrow. Both children were very unhappy being away from home, and their mother was disconsolate in their absence. As a child, Karl was rather frail, delicate, often ill, and prone to depression. There were a number of occasions when he received tuition at home because he was too unwell to go to school. After the Pearsons moved to 40 Mecklenburgh Square, in Holborn, in June 1866 (where they stayed until 1875), Karl and Arthur began attending University College London School.

When they went to Cambridge, at least one of the Pearson boys was expected to read mathematics. The Cambridge Mathematics Tripos was, at that time, the most prestigious degree in any British university. Although his father urged him to read mathematics, Arthur settled on classics. Thus, when Karl was 15 years old, his father was looking for a good Cambridge Wrangler to prepare him for the Mathematics Tripos. Less than a year later, Karl went to Hitchin, near Cambridge, where he stayed from January 28 to July 1, 1874, receiving tuition from the Reverend Louis Hensley. He was very unhappy at Hitchin and was ready to leave by the summer of 1874 so that he could be coached in mathematics in preparation for Cambridge. A couple weeks later, he decided to go to Merton Hall in Cambridge for tuition under John Edward Rendall Harris, John P. Taylor, and Edward John Routh. He stayed at Merton Hall from mid-July 1874 to April 15, 1875.

By the spring of 1875, Pearson was ready to take the entrance examinations at various colleges at Cambridge. His first choice was Trinity College, where he failed the entrance exam; his second choice was King's College, from which he received an Open Fellowship on April 15, 1875. Pearson found that the highly competitive and demanding system leading up to the Mathematical Tripos was the tonic he needed. Although he had been a rather frail, delicate, and sickly child with a nervous disposition, he came to life in this environment and his health improved. In addition to the highly competitive and intellectually demanding system, students of the Mathematics Tripos were expected to take regular exercise as a means of preserving a robust

constitution and regulating the working day. Pearson carefully balanced hard mathematical study against such physical activities as walking, skating, ice hockey, and lawn tennis.

As a diversion from studying mathematics, Pearson read works from such Romantics as Goethe and Shelley in his second year. He also read Rousseau and Dante, and he wrote a couple of articles on Spinoza for the *Cambridge Review*. Pearson's time at King's College left its legacy through his revolt over the compulsory divinity examination. Near the beginning of his third year in 1877, he decided that he no longer wished to be compelled to attend church services. Pearson also refused to retake one of his Divinity papers because it would have interfered with studying for the Maths Tripos. The events that transpired led eventually to King's College abolishing the whole system of compulsory Divinity examinations in March 1878.

Pearson spent the rest of 1878 in preparation for the Mathematics Tripos examination, which he took in January 1879. He graduated with honors, being the Third Wrangler; subsequently, he received a fellowship from King's College, which he held for 7 years. (He was made an honorary fellow of King's in 1903.) Pearson also took the Smith's Prize examination, although he did not become a Prizeman. However, Isaac Todhunter, who had been one of Pearson's examiners for the Smith's Prize, thought that Pearson had provided a better solution to one of Barŕe De Saint-Venant's problems of elasticity than Saint-Venant had. Todhunter subsequently incorporated Pearson's solutions into his manuscript on the "Theory of Elasticity and Strength of Materials." Soon after Todhunter's death in November 1879, the Syndics of Cambridge University Press asked Pearson to finish Todhunter's manuscript. During the early 1880s, Pearson began to devote more of his time to problems of elasticity, which became his specialty in mathematical physics.

A couple of weeks after Pearson had taken his degree, he began to work in Professor James Stuart's engineering workshop and read philosophy during the Lent Term in preparation for his trip to Germany. After making arrangements with Kuno Fischer, Pearson left for Heidelberg in April 1879. His time in Germany was a period of self-discovery philosophically and professionally. The romanticist and idealist discovered positivism: Pearson thus adopted and coalesced two different philosophical traditions to fulfill two different needs. Around this time, he began to write the *New Werther*, a literary work on idealism and materialism, written in the form of letters to his fiancée from a young man wandering in Germany. For Pearson, the book was about "conflict between the ideal and the real, spirit and matter." The book was published in 1880 under the pseudonym of Loki (a mischievious Norse god).

Germany and University College London

In Heidelberg, Pearson read Berkeley, Fichte, Locke, Kant, and Spinoza and was beginning to find that his "faith in reason has been so shattered by the merely negative results to him which he found in these great philosophers that he despaired his little reason leading to anything." He subsequently abandoned philosophy because "it made him miserable and would have led to him to short-cut his career." In November, he went to Berlin to study physics under Quincke and Helmholtz and metaphysics under Kuno Fischer, and he considered becoming a mathematical physicist but decided not to pursue this since he "was not a born genius." Since "philosophy did not lead to the truth" and as he would not find success in physics, he was "determined to go to the Bar": He stayed in Berlin and attended lectures on Roman international law and philosophy by Bruns and Mommsen, and he hoped to pass the Roman Law Exam in March. A year later, he took up rooms at the Inner Temple and read law at Lincoln's Inn. He was called to the Bar at the end of 1881 and practiced the law for a very short time only. Pearson's one extant case involved setting up a partnership deed between two turnip-top sellers in Covent Garden Market; the case took him 3 days to complete, which he thought was "agony" and by then he decided he "hated the law." Still searching for some direction when he returned to London, Pearson lectured on socialism, Marx, and Lassalle at the workingmen's clubs and on Martin Luther at Hampstead from 1880 to 1881.

Three months later, however, he discovered that he was tired of the law and did not want to pursue this because it depressed him; he decided instead to "devote his time to the religious producing of German literature before 1300." Later that year, his work on "The Trinity, A Nineteenth Century Passion-Play, The Son; or Victory of Love" was published. From 1882 to 1884, he lectured on German society from the medieval period up to the 16th century. He became so competent in German that by the late spring of 1884, he was offered a post in German at Cambridge. In his pursuit of German history, Pearson consulted his friend, the Cambridge University librarian Henry Bradshaw, who taught him the meaning of thoroughness and patience in research. With Bradshaw's help, in 1887 Pearson finished *Die Fronica: Ein Beitrag zur Geschichte des Christusbildes im Mittelalter* (which involved a collection of the so-called Veronica portraits of Christ).

Despite his accomplishments with German literature, however, he "longed to be working with symbols rather than words." He then began to write papers on the theory of elastic solids and fluids as well as some mathematical

physics papers on optics and ether squirts. He deputized mathematics at King's College, London, and for Professor Rowe at UCL in 1883. Having no luck in finding a job, he thought of taking up a secretaryship in a hospital, becoming a school master, possibly emigrating to North America, or even returning to law. Between 1879 and 1884, he applied for more than six mathematical posts and he received the Goldsmid Chair of "Mechanism and Applied Mathematics" at UCL in June 1884. Thomas Archer Hirst and Alexander Kennedy had recommended him to the post.

During Pearson's first 6 years at UCL, he taught mathematical physics, hydrodynamics, magnetism, electricity, and his specialty, elasticity, to engineering students. Nearly all of his teaching on dynamics, general mechanics, and statics was based on geometrical methods. He finished editing the incomplete manuscript of William Kingdon Clifford's *The Common Sense of the Exact Sciences* in 1885, and a year later he finished Todhunter's *History of the Theory of Elasticity*.

The Gresham Lectures on Geometry and Curve Fitting

Pearson was a founding member of the Men's and Women's Club established in 1885 "for the free and unreserved discussion of all matters in any way connected with the mutual position and relation of men and women." Among the various members was Marie Sharpe, whom he married in June 1890: They had three children—Sigrid, Helga, and Egon. Six months after his marriage, he took up another teaching post in the Gresham Chair of Geometry in London, which he held for 3 years concurrently with his post at UCL. As Gresham Professor, he was responsible for giving 12 lectures a year, delivered on 4 consecutive days from Tuesdays to Fridays, during the Michaelmas, Easter, and Hilary terms. The lectures, which were free to the public, began at 6 PM and lasted for 1 hour. Between February 1891 and November 1893, Pearson delivered 38 lectures. His first 8 lectures formed the basis of his book, *Grammar of Science*, which was published in several languages, and it was in his last 12 lectures where he provided the framework to the modern theory of mathematical statistics.

Pearson's earliest teaching of statistics can thus be found in his lecture of November 18, 1891, when he discussed graphical statistics and the mathematical theory of probability with a particular interest in actuarial methods. Two days later, he introduced the histogram—a term he coined to designate a "time-diagram" to be used for historical purposes. He introduced the standard deviation in his Gresham lecture of January 31, 1893. Pearson's early Gresham lectures on statistics were influenced by the work of Francis Ysidro Edgeworth, Stanley Jevons, and John Venn. Until November 1893, these lectures covered fairly conventional statistical and probability methods. Although the material in these lectures was not original in content, Pearson's approach in teaching was highly innovative. In one of his lectures, he scattered 10,000 pennies over the lecture room floor and asked his students to count the number of heads or tails: "The result was very nearly half heads and half tails, thus proving the law of average and probability." After a lecture on experimental deductions that involved the use of 16,718 throws of the ball at the Monte Carlo Roulette Table, teetotums, and 2138 tickets from lotteries, one of his students remarked that the lecture was like "an opera without a last act." It is, perhaps, not surprising that the number of students increased 5- to 10-fold in the first couple of years; by 1893, nearly 300 students were attending his lectures.

Pearson's last 12 Gresham lectures signified a turning point in his career owing, in particular, to his relationship with Weldon, who was the first biologist Pearson met who was interested in using a statistical approach for problems of Darwinian evolution. Their emphasis on Darwinian population of species not only implied the necessity of systematically measuring variation but also prompted the reconceptualization of statistical populations. Moreover, it was this mathematization of Darwin that led to a paradigmatic shift for Pearson from the Aristotelian essentialism underpinning the earlier use and development of social and vital statistics. Weldon's questions not only provided the impetus for Pearson's seminal statistical work but also led eventually to the creation of the Biometric School at UCL.

In Pearson's first published statistical paper of October 26, 1893, he introduced the method of moments as a means of curve fitting asymmetrical distributions. One of his aims in developing the method of moments was to provide a general method for determining the values of the parameters of a frequency distribution (i.e., central tendency, variation, skewness, and kurtosis). In 1895, Pearson developed a general formula to use for subsets of various types of frequency curves and defined the following curves: type I (asymmetric beta density curve), type II (symmetric beta curve), type III (gamma curve), type IV (family of asymmetric curves), and type V (normal curve). In his first supplement to his family of curves in 1901, he defined types VI and VII (type VII is now known as "Student's" distribution), and he defined types VIII and IX in his second supplement in 1916. Many of his curves were J-shaped, U-shaped, and skewed. Pearson derived all of his curves from a differential equation whose parameters were found from the moments of the distribution. As Churchill Eisenhart remarked in 1974, "Pearson's family of curves did much to dispel the almost religious acceptance of the normal distribution as the mathematical model of variation of biological, physical, and social phenomena."

Although the method of moments is not widely used by biostatisticians today, it remains a very powerful tool in econometrics.

The Biometric School

Although Pearson's success in attracting such large audiences in his Gresham lectures may have played a role in encouraging him to further develop his work in biometry, he resigned from the Gresham Lectureship due to his doctor's recommendation. Following the success of his Gresham lectures, Pearson began to teach statistics to students at UCL in October 1894. Not only did Galton's work on his law of ancestral heredity enable Pearson to devise the mathematical properties of the product–moment correlation coefficient (which measures the relationship between two continuous variables) and simple regression (used for the linear prediction between two continuous variables) but also Galton's ideas led to Pearson's introduction of multiple correlation and part correlation coefficients, multiple regression and the standard error of estimate (for regression), and the coefficient of variation. By then, Galton had determined graphically the idea of correlation and regression for the normal distribution only. Because Galton's procedure for measuring correlation involved measuring the slope of the regression line (which was a measure of regression instead), Pearson kept Galton's "r" to symbolize correlation. Pearson later used the letter b (from the equation for a straight line) to symbolize regression. After Weldon had seen a copy of Pearson's 1896 paper on correlation, he suggested to Pearson that he should extend the range for correlation from 0 to +1 (as used by Galton) so that it would include all values from −1 to +1.

Pearson achieved a mathematical resolution of multiple correlation and multiple regression, adumbrated in Galton's law of ancestral heredity in 1885, in his seminal paper *Regression, Heredity, and Panmixia* in 1896, when he introduced matrix algebra into statistical theory. (Arthur Cayley, who taught at Cambridge when Pearson was a student, created matrix algebra by his discovery of the theory of invariants during the mid-19th century.) Pearson's theory of multiple regression became important to his work on Mendel in 1904 when he advocated a synthesis of Mendelism and biometry. In the same paper, Pearson also introduced the following statistical methods: eta (η) as a measure for a curvilinear relationship, the standard error of estimate, multiple regression, and multiple and part correlation. He also devised the coefficient of variation as a measure of the ratio of a standard deviation to the corresponding mean expressed as a percentage.

By the end of the 19th century, he began to consider the relationship between two discrete variables, and from 1896 to 1911 Pearson devised more than 18 methods of correlation. In 1900, he devised the tetrachoric correlation and the phi coefficient for dichotomous variables. The tetrachoric correlation requires that both X and Y represent continuous, normally distributed, and linearly related variables, whereas the phi coefficient was designed for so-called point distributions, which implies that the two classes have two point values or merely represent some qualitative attribute. Nine years later, he devised the biserial correlation, where one variable is continuous and the other is discontinuous. With his son Egon, he devised the polychoric correlation in 1922 (which is very similar to canonical correlation today). Although not all of Pearson's correlational methods have survived him, a number of these methods are still the principal tools used by psychometricians for test construction. Following the publication of his first three statistical papers in *Philosophical Transactions of the Royal Society,* Pearson was elected a fellow of the Royal Society in 1896. He was awarded the Darwin Medal from the Royal Society in 1898.

Pearson's Chi-Square Tests

At the turn of the century, Pearson reached a fundamental breakthrough in his development of a modern theory of statistics when he found the exact chi-square distribution from the family of gamma distributions and devised the chi-square (χ^2, P) goodness-of-fit test. The test was constructed to compare observed frequencies in an empirical distribution with expected frequencies in a theoretical distribution to determine "whether a reasonable graduation had been achieved" (i.e., one with an acceptable probability). This landmark achievement was the outcome of the previous 8 years of curve fitting for asymmetrical distributions and, in particular, of Pearson's attempts to find an empirical measure of a goodness-of-fit test for asymmetrical curves.

Four years later, he extended this to the analysis of manifold contingency tables and introduced the "mean square contingency coefficient," which he also termed the chi-square test of independence (which R. A. Fisher termed the chi-square statistic in 1923). Although Pearson used $n - 1$ for his degrees of freedom for the chi-square goodness-of-fit test, Fisher claimed in 1924 that Pearson also used the same degrees of freedom for his chi-square test of independence. However, in 1913 Pearson introduced what he termed a "correction" (rather than degrees of freedom) for his chi-square test of independence of 1904. Thus, he wrote, if $x =$ number of rows and $\lambda =$ number of columns, then on average the correction for the number of cells is $(x - 1)(\lambda - 1)/N$. [As may be seen, Fisher's degrees of freedom for the chi-square statistic as $(r - 1)(c - 1)$ is very similar to that used by Pearson in 1913.]

Pearson's conception of contingency led at once to the generalization of the notion of the association of two attributes developed by his former student, G. Udny Yule. Individuals could now be classed into more than two alternate groups or into many groups with exclusive attributes. The contingency coefficient and the chi-square test of independence could then be used to determine the extent to which two such systems were contingent or noncontingent. This was accomplished by using a generalized theory of association along with the mathematical theory of independent probability.

Pearson's Four Laboratories

In the 20th century, Pearson established and ran four laboratories. He set up the Drapers' Biometric Laboratory in 1903 following a grant from the Worshipful Drapers' Company (which funded Pearson annually for work in this laboratory until his retirement in 1933). The methodology incorporated in the Drapers' Biometric Laboratory was twofold: The first was mathematical, and included the use of Pearson's statistical methods, matrix algebra, and analytical solid geometry. The second involved the use of such instruments as integrators, analyzers, curve-potters, the cranial coordinatograph, silhouettes, and cameras. The problems investigated by the biometricians included natural selection, Mendelian genetics and Galton's law of ancestral inheritance, craniometry, physical anthropology, and theoretical aspects of mathematical statistics. By 1915, Pearson had established the first-degree course in mathematical statistics in Britain.

Although Pearson did not accept the generality of Mendelism, he did not reject it completely as is commonly believed. When William Bateson published his fiercely polemical attack on Weldon in 1902, Bateson saw Mendelism as a tool for discontinuous variation only. As a biometrician, most of the variables that Pearson and coworkers analyzed were continuous and only occasionally did they examine discontinuous variables. Although Pearson and Weldon used Galton's law of ancestral inheritance for continuous variables, they used Mendelism for discontinuous variables. Indeed, Pearson argued that his chi-square test of independence was the most appropriate statistical tool for the analysis of Mendel's discrete data for dominant and recessive alleles (such as color of eyes, where brown is dominant and blue is recessive). Even today, Pearson's chi-square test is used for analyzing Mendelian data.

A year after Pearson established the Biometric Laboratory, the Worshipful Drapers' Company gave him a grant so that he could establish an Astronomical Laboratory equipped with a transit circle and a 4-in. equatorial refractor. Hence, he also referred to his two observatories as the Transit House and the Equatorial House. Pearson was interested in determining the correlations of stellar rotations and the variability in stellar parallax. One of his larger projects involved taking 132 photographs of the eclipse of the sun on June 28, 1908. He joined the Royal Astronomical Society in 1904 and resigned in 1917 following a row he had with Lord Rayleigh and H. C. Plummer on matters relating to differences in methods when calculating these variables. Pearson was also instrumental in setting up a degree course in astronomy in 1914 at UCL.

In the autumn of 1907 when Francis Galton was 85 years old, he asked Pearson if he could take "control of the Eugenics Office as a branch of the Biometric Laboratory." Pearson had, by then, been doing the work of at least three different people: He was editing the four journals he had thus far founded, writing papers, giving lectures, and managing three laboratories aided largely by various teaching assistants, an assistant professor, human computers, calculators, and computators in addition to numerous voluntary workers and visiting scholars. Given Pearson's ongoing professional commitments, he was very reluctant to take up the directorship of another laboratory. Pearson wrote a letter to Galton saying he did not think that his biometric methods were amenable for the quantitative treatment of eugenics problems.

He explained to Galton that many of the biometric problems that were undertaken in the Biometric Laboratory took up to 5 or 6 years and required the manpower from two or three generations of students plus an additional 2 years to reduce the data before they could be analyzed statistically. His methods were clearly not suitable for the quick and easy solutions that Galton favored. Galton admitted that he "had been under the false idea that the Eugenics Laboratory would have aided the Biometric Laboratory rather than hindered it." Nevertheless, Pearson felt a great deal of affection for Galton, which had increased since Weldon's death in 1906; moreover, given the frailty of Galton at his advanced age, it is clear that Pearson would not have refused to help Galton—at least not in Galton's lifetime.

As a personal favor to Galton, Pearson agreed, albeit reluctantly, to take control of the Eugenics Record Office on February 1, 1907, and renamed it the Eugenics Laboratory when he became its director. After he had been director for 1 year, he was ready to step down whenever Galton felt another man could achieve more in the particular direction desired by Galton. Pearson's overriding professional objective was unequivocally linked to the promulgation and further development of his statistical ideas and theory, which he did not think had as much application to eugenics as they did to biology and astronomy. Because no replacement was found, Pearson stayed on as director. Nonetheless, he had long been aware of the limitations of his biometric methods for problems in

the Eugenics Laboratory, and since his methods were of limited use in the Eugenics Laboratory, his only option was to devise a new methodology for problems being investigated in this laboratory. This methodology was underpinned by the use of actuarial death rates and by a very highly specialized use of family pedigrees assembled in an attempt to discover the inheritance of various diseases (which included alcoholism, cancer, diabetes, epilepsy, paralysis, and pulmonary tuberculosis).

To a large extent, Pearson stood outside many of the debates on the implementation of various social policies for eugenics, and he never endorsed sterilization or any other dysgenic state policies. The projects he initiated in the Eugenics Laboratory were long-term quantitative studies dealing with such issues as alcoholism, insanity, and tuberculosis. He received assistance from the coworkers who worked in the laboratory and from the medical community, which acquired as much family data as possible for his family pedigrees for calculating actuarial death rates. The eminent biologist, J. B. S. Haldane, regarded Pearson's family pedigrees, most of which were published in his 21-volume "The Treasury of Human Inheritance," as his most valuable contribution to biology.

Although Pearson never developed his own theory of inheritance, he considered different theories of heredity, including blending, "exclusive" and mosaic inheritance, Galton's law of ancestral inheritance, homotyposis, and Mendelian genetics; thus, he did not reject Mendelism. Pearson's view of the world was shaped by the Cambridge Mathematics Tripos, which emphasized using applied mathematics as a pedagogical tool for obtaining the truth. To this end, he was not interested in a physiological mechanism of heredity; instead, he attempted to make sense of various hereditarian models by placing them in a mathematical context.

Family Pedigrees and the Medical Community

Pearson's family pedigrees served as the vehicle through which he could communicate statistical ideas to the medical community by stressing the importance of using quantitative methods for medical research. This tool enabled doctors to move away from concentrating on individual pathological cases or "types" and to see, instead, a wide range of pathological variation of the disease (or condition) of the doctors' specialty. Major Greenwood, who was the first medically qualified person to take an interest in Pearson's statistics in 1902, became Reader in Medical Statistics at the University of London in 1922 (the first such position to be held at a university in Britain). Pearson's and Greenwood's statistical work was further promulgated by their student Austin Bradford Hill, who had the greatest impact on the successful adoption of mathematical statistics in the medical community. In 1924, Pearson set up the Anthropometric Laboratory, which was made possible by a gift from one of Pearson's students, Ethel Eldeteron. The laboratory was open to the public and used to collect and display statistics related to problems of heredity.

In the spring of 1909, Galton was discussing the future of the Eugenics Laboratory with Pearson. Although Galton thought that Pearson would have been "the most suitable man for the first Galton Professor," Pearson let Galton know that he was "wholly unwilling to give up superintendence of the Biometric Laboratory [he] had founded and confine [his] work to Eugenics Research." A month later, Galton added a codicil to his will stating that he desired that the first professor of the post should be offered to Pearson on such condition that Pearson could continue to run his Biometric Laboratory. Six months after Galton's death in January 1911, Pearson first learned about Galton's codicil to his will. He then relinquished the Goldsmid Chair of Applied Mathematics after 27 years of tenure to take up the Galton Chair. The Drapers' Biometric and the Galton Eugenics laboratories, which continued to receive separate funding, then became incorporated into the Department of Applied Statistics. The essential aim in combining both laboratories was to enable Pearson to give up his undergraduate teaching of applied mathematics and to devote himself "solely to what had been for many years the main element of [his] research: the advancement of the modern theory of statistics."

Statistical Charts and Gunnery Computations

Pearson proceeded to raise funding for a new building for his Department of Applied Statistics. Adequate funding had been raised by 1914, and contracts for the fittings had been made. In the early summer of 1914, the new laboratory was complete and preparations were under way for the occupation and fitting up of the public museum and the Anthropometric Laboratory. It was hoped that the building would be occupied by October 1915. These developments and further biometric work were shattered by the onset of World War I. The new laboratory building was taken over by the government to be used as a military hospital. Pearson and coworkers took on special war duties. They produced statistical charts for the Board of Trade's Labour Department as well as for its census production. Pearson was also involved with elaborate calculations of anti-aircraft guns and bomb trajectories both through air and water. By June 1919, Pearson was in possession of his building and plans were under way

for the opening in October 1920. It was not until December 4, 1922, that the work had been completed and the building was occupied.

His wife, Marie Sharpe, died in 1928, and in 1929 he married Margaret Victoria Child, a coworker in the Biometric Laboratory. Pearson was made emeritus professor in 1933 and was given a room in the zoology department at UCL that he used as the office of *Biometrika.* From his retirement until his death in 1936, he published 34 articles and notes and continued to edit *Biometrika.* Pearson was offered an OBE in 1920 and a knighthood in 1933, but he refused both honors. He also declined the Royal Statistical Society Guy Medal in its centenary year in 1934. Pearson believed that "all medals and honors should be given to young men, they encourage them when they begin to doubt whether their work was of value." Pearson accepted the honorary DSc from the University of London in 1934 because if he had refused, he "would have hurt the executive of the university where he had worked" for nearly half a century.

Pearson's statistical achievement not only provided continuity from the mathematical and statistical work that preceded him (including that of Francis Ysidro Edgeworth, Francis Galton, Adolphe Quetelet, and John Venn) or that of contemporaries (such as W. F. R. Weldon and George Udny Yule) but also his work engendered the modern theory of mathematical statistics in the 20th century, which in turn provided the foundation for such statisticians as R. A. Fisher, who went on to make further advancements for a modern theory of statistics.

See Also the Following Articles

Correlations • Eugenics

Further Reading

Eisenhart, C. (1974). Karl Pearson. In *Dictionary of Scientific Biography,* Vol. 10, pp. 447–473. Scribner's, New York.

Hilts, V. (1981). *Statist and Statistician.* Arno, New York (Reprint of PhD thesis, Harvard University, 1967).

Mackenzie, D. (1981). *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge.* Edinburgh University Press, Edinburgh, UK.

Magnello, M. E. (1993). Karl Pearson: Evolutionary biology and the emergence of a modern theory of statistics. DPhil thesis, University of Oxford, Oxford, UK.

Magnello, M. E. (1996). Karl Pearson's Gresham Lectures: W. F. R. Weldon, speciation and the origins of Pearsonian statistics. *Br. J. History Sci.* **29,** 43–64.

Magnello, M. E. (1998). Karl Pearson's mathematization of inheritance: From ancestral heredity to Mendelian genetics (1895–1909). *Ann. Sci.* **55,** 35–94.

Magnello, M. E. (1999). The non-correlation of biometrics and eugenics: Rival forms of laboratory work in Karl Pearson's career at University College London. *History Sci.* **37,** 79–106, 123–150.

Magnello, M. E. (2002). The introduction of mathematical statistics in medical research: The roles of Karl Pearson, Major Greenwood and Austin Bradford Hill. In *The Road to Medical Statistics* (E. Magnello and A. Hardy, eds.), Rodopi Press, Amsterdam.

Norton, B. (1978a). Karl Pearson and the Galtonian tradition: Studies in the rise of quantitative social biology. PhD thesis, University of London, London.

Norton, B. (1978b). Karl Pearson and statistics: The social origin of scientific innovation. *Soc. Stud. Sci.* **8,** 3–34.

Pearson, E. (1936–1938). Karl Pearson: An appreciation of some aspects of his life and work. Part 1, 1857–1905, *Biometrika*, 193–257 (1936); Part 2, 1906–1936, *Biometrika*, 161–248 (1938). (Reprinted by Cambridge University Press, 1938.)

Pearson, K. (1893). Asymmetrical frequency curves. *Nature* **48,** 615–616.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. A* **185,** 71–110.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philos. Trans. R. Soc. A* **186,** 343–414.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. R. Soc. A* **187,** 253–318.

Pearson, K. (with Filon, L. N. G.) (1898). Mathematical contributions to the theory of evolution. IV. On the probable error of frequency constants and on the influence of random selection on variation and correlation. *Philos. Trans. R. Soc. A* **191,** 229–311.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. A* **195,** 1–47.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London Edinburgh Dublin Philos. Mag. J. Sc.* **50** (Fifth series), 157–175.

Pearson, K. (1904). Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation. *Drapers' Company Res. Mem. Biometric Ser. 1*, 1–37.

Pearson, K. (1914–1930). *The Life, Letters and Labours of Francis Galton* (3 vols. in 4 parts). Cambridge University Press, Cambridge, UK.

Porter, T. M. (1986). *The Rise of Statistical Thinking. 1820–1900.* Princeton University Press, Princeton, NJ.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Belknap Press of Harvard University Press, Cambridge, MA.

Performance Prism

Andy Neely
Cranfield School of Management, Bedfordshire, United Kingdom

Chris Adams
Cranfield School of Management, Bedfordshire, United Kingdom

Glossary

balanced scorecard A scorecard consists of the vital defined measures that executive teams use to manage their responsibilities in business and not-for-profit organizations. A balanced scorecard is one that contains not only an appropriate mix of both financial and non-financial measures, but also a balance of internal and external plus input and output measures. The term has been popularized since 1992 by the writings and activities of R. Kaplan and D. Norton of the Harvard Business School.

business excellence model A nine-faceted framework, consisting of five "enablers" and four "results," launched in 1992 by the European Foundation for Quality Management (EFQM) as an assessment and self-assessment framework for the European Quality Award (similar to the U.S. Malcolm Baldrige Award), sometimes adapted by organizations as a performance measurement framework.

Malcolm Baldrige Award A popular U.S. annual quality award, established in 1987, named after an American secretary of commerce who was a strong supporter of the award's aims but who, unfortunately, was killed in an accident shortly before its launch. The assessment criteria for the award address seven categories of performance excellence.

shareholder value The sum of a listed corporation's change in market capitalization (number of shares in issue multiplied by its quoted share price) plus its dividend distributions over a given period.

stakeholders All the parties who have a particular interest in, and legitimate requirements of, an organization's performance—typically, investors, customers, employees, suppliers, regulators, and the communities in which it operates.

Three fundamental premises underpin the concept of the performance prism. First, it is no longer acceptable (or even feasible) for organizations to focus solely on the needs of one or two of their stakeholders—typically shareholders and customers—if they wish to survive and prosper over the long term. Second, an organization's strategies, processes, and capabilities have to be aligned and integrated with one another if the organization is to be best positioned to deliver real value to all of its stakeholders. Third, organizations and their stakeholders have to recognize that their relationships are reciprocal—stakeholders have to contribute to organizations as well as receive something from them. These three fundamental premises underpin the holistic performance measurement and management framework—the performance prism—that this article describes.

Introduction

The performance prism is a second-generation performance management framework that consists of five interrelated perspectives on performance that pose specific vital questions:

1. *Stakeholder satisfaction*: Who are our key stakeholders and what do they want and need?
2. *Stakeholder contribution*: What do we want and need from our stakeholders on a reciprocal basis?
3. *Strategies*: What strategies do we need to put in place to satisfy the wants and needs of our stakeholders while satisfying our own requirements as well?
4. *Processes*: What processes do we need to put in place to enable us to execute our strategies?

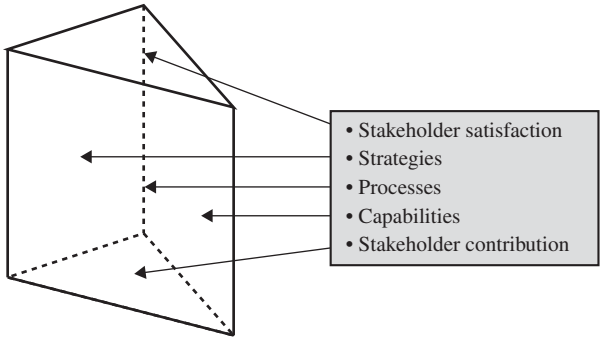


Figure 1

5. *Capabilities*: What capabilities do we need to put in place to allow us to operate our processes?

Together, these five perspectives provide a comprehensive and integrated framework for thinking about organizational performance in today's operating environment (see Fig. 1).

The performance prism also seeks to address the shortcomings of the first-generation measurement frameworks and methodologies, such as the balanced scorecard, the work on shareholder value, and the various self-assessment frameworks, such as the Malcolm Baldrige Award criteria and the business excellence model of the European Foundation for Quality Management (EFQM).

The Nature of the Measurement Problem

Why is this new performance measurement framework needed? After all, everyone knows that "you can't manage what you don't measure." And given that people have been managing organizations for years, then surely by now they must have perfected their measurement systems.

Sadly, as in so many walks of life, theory does not reflect practice. The number of organizations with weak performance measures and measurement systems in place is immense. Examples abound of organizations that have introduced performance measures that quite simply drive entirely the wrong behaviors. There must be a better way.

There has been a revolution in performance measurement and management during the last few years. Various frameworks and methodologies—such as the balanced scorecard, shareholder value added, activity-based costing, cost of quality, and competitive benchmarking—have each generated vast interest, activity, and consulting revenues, but not always success. Yet therein lies a paradox. It might reasonably be asked: how can multiple, and seemingly inconsistent, business performance frameworks and measurement methodologies exist? Each claims to be unique and comprehensive, yet each offers a different perspective on performance.

Kaplan and Norton's balanced scorecard, with its four perspectives, focuses on financials (shareholders), customers, internal processes, plus innovation and learning. In doing so, it downplays the importance of other stakeholders, such as employees, suppliers, regulators, and communities. The business excellence model combines results, which are readily measurable, with enablers, some of which are not. Shareholder value frameworks incorporate the cost of capital into the equation, but ignore everything (and everyone) else. Both the activity-based costing and the cost of quality frameworks, on the other hand, focus on the identification and control of cost drivers (non-value-adding activities and failures/non-conformances, respectively), which are themselves often embedded in the business processes. But this highly process-focused view ignores any other perspectives on performance, such as the opinions of investors, customers, and employees. Conversely, benchmarking tends to involve taking a largely external perspective, often comparing performance with that of competitors or sometimes other "best practitioners" of business processes or capabilities. However, this kind of activity is frequently pursued as a one-off exercise toward generating ideas for—or gaining commitment to—short-term improvement initiatives, rather than the design of a formalized ongoing performance measurement system.

How can this be? How can multiple, seemingly conflicting, measurement frameworks and methodologies exist? The answer is simple: they can exist because they all add value. They all provide unique perspectives on performance. They all furnish managers with a different set of lenses through which they can assess the performance of their organizations. The key is to recognize that, despite the claims of some of the proponents of these various approaches, there is no one best way to address the measurement and management of business performance. The reason for this is that business performance is itself a multifaceted concept, the complexity of which the existing frameworks only partially address. Essentially, they provide valuable point solutions.

A Better Solution to the Measurement Problem

Our solution is the three-dimensional framework that we call the performance prism. This framework has been deliberately designed to be highly flexible so that it can provide either a broad or a narrow focus. If only a partial

aspect of performance management is required, such as a single stakeholder focus or a particular business process agenda, then the performance prism can be applied to designing a measurement system and appropriate measures (and their attendant metrics) that address that context. Conversely, if a broad corporate or business unit performance management improvement initiative is required, the performance prism is equally capable of supporting that, too. How does it help to achieve these aims?

The performance prism has five perspectives. The top and bottom perspectives are stakeholder satisfaction and stakeholder contribution. The three side perspectives are the organization's strategies, processes, and capabilities for addressing those sets of wants and needs. Figure 2 illustrates these five basic perspectives of performance measurement and management.

Why does the framework look like this and why does it consist of these constituent components? It is clear that those organizations aspiring to be successful in the long term within today's business environment need to have an exceptionally clear picture of who their key stakeholders are and what they want or need. But having a clear picture is not enough. In order to satisfy their own wants and needs, organizations have to access contributions from their stakeholders—usually capital and credit from investors, loyalty and profit from customers, ideas and skills from employees, materials and services from suppliers, and so on. They also need to have defined what strategies they will pursue to ensure that value is delivered to their stakeholders. In order to implement these strategies, they have to understand what processes the enterprise requires and must operate both effectively and efficiently. Processes, in themselves, can only be executed if the organization has the right capabilities in place—the right combination of people skill sets, best practices, leading technologies, and physical infrastructure.

In essence, then, the performance prism provides a comprehensive yet easily comprehensible framework that can be used to articulate a given business's operating model. Its components are described in the following sections.

Stakeholder Satisfaction

Where should the measurement design process begin? One of the great myths (or fallacies) of measurement design is that performance measures should be derived from strategy. Listen to any conference speaker on the subject. Read any management text written about it. Nine times out of ten the statement "Derive your measures from your strategy" will be made. This is such a conceptually appealing notion that nobody stops to question it. Yet to derive measures from strategy is to misunderstand fundamentally the purpose of measurement and the role of strategy. Performance measures are designed to help people track whether they are moving in the intended direction. They help managers establish whether they are going to reach the destination they set out to reach. Strategy, however, is not about destination. Instead, it is about the route that is chosen—*how* to reach the desired destination.

At one level, this is a semantic argument. Indeed, the original work on strategy, carried out in the 1970s by Andrews, Ansoff, and Mintzberg, asserted that a strategy should explain both the goals of the organization and a plan of action to achieve these goals. Today,

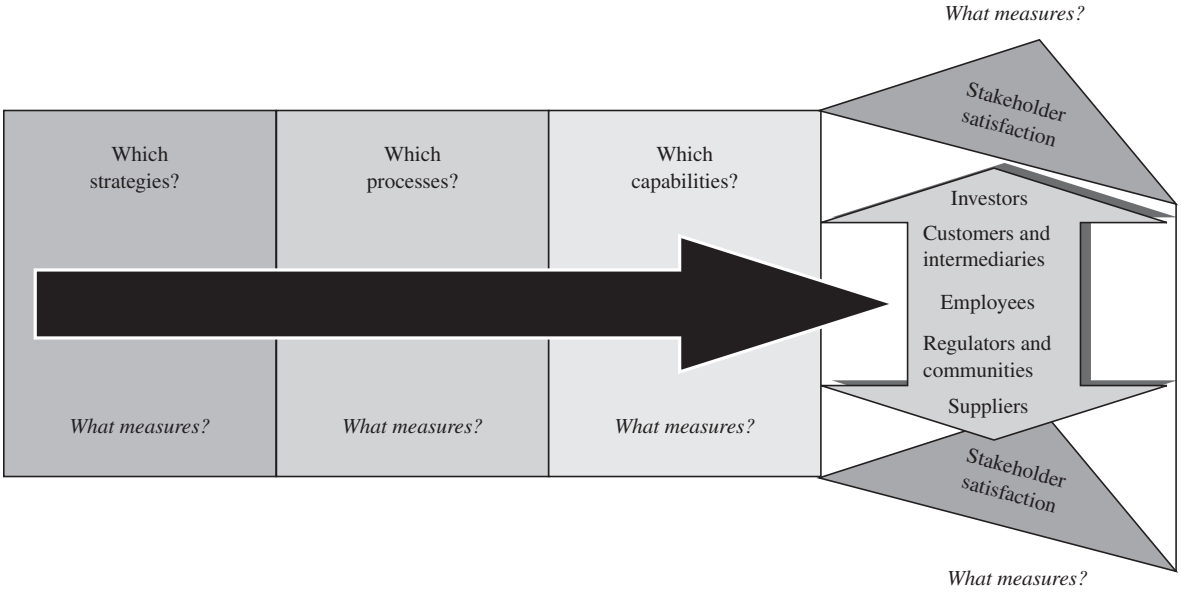


Figure 2

however, the vast majority of organizations have strategies that are dominated by lists of improvement activities and management initiatives, e.g., grow market share, extend the product range, and seek new distribution channels. While these are undoubtedly of value, they are not the end-goal. These initiatives and activities are pursued in the belief that, when implemented, they will enable the organization to deliver better value to its multiple stakeholders, all of whom will have varying importance to the organization.

An organization's key stakeholders are likely to be a combination of a number of the following:

- Investors (principally shareholders, but other capital providers too)
- Customers and intermediaries
- Employees and labor unions
- Suppliers and alliance partners
- Regulators, pressure groups, and communities.

Organizations can chose to give more attention to one stakeholder group over another, not because that particular stakeholder is implicitly more important than the others, but because that stakeholder has not received the attention it should have in the past. Executives must decide which stakeholders' wants and needs their strategies must satisfy.

So, the starting point for deciding what to measure should not be "What is the organization's strategy?" but instead, "Who are the organization's stakeholders and what do they want and need?" Hence, the first perspective on performance embedded in the performance prism is that of stakeholder satisfaction.

Stakeholder Contribution

The second perspective on performance is a subtle but critical twist on the first. Take, for example, customers as stakeholders. In the early 1980s, organizations began to measure customer satisfaction by tracking the number of customer complaints they received. When research evidence started to show that only about 10% of dissatisfied customers complained, organizations turned to more sophisticated measures, such as customer satisfaction. In the late 1980s and early 1990s, people began to question whether customer satisfaction was enough. Research data gathered by Xerox showed that customers who were "very satisfied" were five times more likely to repeat their purchase in the next 18 months than those who were just "satisfied." This and similar observations resulted in the development of the concept known as customer loyalty. The aim of this concept was to track whether customers (1) returned to buy more from the same organization, and (2) recommended the organization to others.

More recently, research data from a variety of industries has demonstrated that many customers are not profitable for organizations. For example, it has been suggested that in retail banking, 20% of customers generate 130% of profits. Other data illustrate that increased levels of customer satisfaction can result in reduced levels of organizational profitability because of the high costs of squeezing out the final few percentage points of customer satisfaction. Perhaps; but most of the evidence suggests that very few substantial companies are anywhere near running into that dilemma. Nevertheless, the reaction has been increasing interest in the notion of customer profitability. Sometimes the customer profitability data produces surprises for the organization, indicating that a group of customers thought to be quite profitable are in fact loss-makers and that other customer groups are far more profitable than generally believed by the organization's executives. Performance data allow assumptions to be challenged.

The important point, and where a subtle twist comes into play, is that customers do not necessarily want to be loyal or profitable. Customers want great products and services at a reasonable cost. They want satisfaction from the organizations they choose to use. It is the organizations themselves that want loyal and profitable customers. So it is with employee satisfaction and supplier performance. For the most part, organizations want loyal employees as well as loyal customers, and they want their workforce to do their jobs with high productivity levels. Many organizations grade their employees based on their contribution, and this grading often has a very direct bearing on their remuneration (an employee want and need, of course).

For years, managers have struggled to measure the performance of suppliers. Do they deliver on time? Do they send the right quantity of goods? And, especially, is their quality good? But these are all dimensions of performance that the organization requires of its suppliers. They encapsulate the suppliers' contribution to the organization. Supplier satisfaction is a completely different concept. If a manager wanted to assess supplier satisfaction, then he or she would have to ask the following questions: Do we pay on time? Do we provide adequate notice when our requirements change? Do we offer suppliers forward schedule visibility? Do our pricing structures allow our suppliers sufficient cash flows for future investment and, therefore, ongoing productivity improvement? Could we be making better use of our vendors' core capabilities and outsource more to them? Again, supplier satisfaction is different to supplier contribution.

The key message here is that for every stakeholder there is a *quid pro quo*: what the organization wants and needs from them as well as what the stakeholder wants and needs from the organization; the right-hand side of Fig. 2 illustrates this. We have found from experience that gaining a clear understanding of the dynamic tension that exists between what stakeholders

want and need from the organization and what the organization wants and needs from its stakeholders can be an extremely valuable learning exercise for the vast majority of organizations.

Strategies

The key question underlying this perspective is: What strategies should the organization adopt to ensure that the wants and needs of its stakeholders are satisfied (while ensuring that its own requirements are satisfied as well)? In this context, the role of measurement is fourfold. First, measures are required so that managers can track whether the strategies they have chosen are actually being implemented. Second, measures can be used to communicate these strategies within the organization. Third, measures can be applied to encourage and provide incentives for implementation of strategy. Fourth, once available, the measurement data can be analyzed and used to challenge whether the strategies are working as planned (and, if not, why not).

Strategies can be applied at different levels within an organization. Typically, corporate strategies will deal with questions such as: What businesses do we want to be in? And how shall we be successful building them? Business unit strategies will usually consider the following: What markets do we want to be in? And how shall we be successful serving them? Brands, products, and services strategies address problems such as: What brands, products and services shall we offer to these markets? And how shall we be successful offering them? Finally, operating strategies tend to ask: What processes and capabilities must we develop in order to serve these markets and provide these products and services effectively and efficiently? And how shall we successfully implement and achieve them?

The adages "you get what you measure" and "you get what you inspect, not what you expect" contain an important message. People in organizations respond to measures. Horror stories abound of how individuals and teams appear to be performing well, yet are actually damaging the business. For example, when the length of time it takes call center staff to deal with customer calls is monitored, it is not uncommon to find them cutting people off mid-call—just so the data suggest that they have handled the call within their 60 second target. Malevolently or not, employees tend to adopt "gaming tactics" in order to achieve the target performance levels they have been given. Measures send people messages about what matters and how they should behave. When the measures are consistent with the organization's strategies, they encourage behaviors that are consistent with strategy. The right measures then offer not only a means of tracking whether strategy is being implemented, but also a means of communicating strategy and encouraging implementation.

Many of the existing measurement frameworks and methodologies appear to stop at this point. Once the strategies have been identified and the right "leading and lagging" measures established, it is assumed that everything will be fine. Yet studies suggest that some 90% of managers fail to implement and deliver their organization's strategies. Why? There are multiple reasons, but a key one is that strategies also contain inherent assumptions about the drivers of improved business performance. Clearly, if the assumptions are false, then the expected benefits will not be achieved. Without the critical data to enable these assumptions to be challenged, strategy formulation (and revision) is largely predicated on "gut feel" and management theory. Furthermore, strategies can be blown off course by external dependencies that are beyond the control of the organization. Measurement data and its analysis will never replace executive intuition, but it can be used to greatly enhance the making of judgments and decisions. A critical judgment is, of course, whether an organization's strategy and business model remain valid.

A further key reason for strategic failure is that the organization's processes are not aligned with its strategies. And even if its processes are aligned, perhaps the capabilities required to operate these processes are not. Hence, the next two perspectives on performance are the processes and capabilities perspectives. Again, measurement plays a crucial role by allowing managers to track whether the right processes and capabilities are in place, to communicate which processes and capabilities matter, and to encourage people within the organization to maintain or proactively nurture these processes and capabilities as appropriate. This may involve gaining an understanding of which particular business processes and capabilities must be competitively distinctive ("winners"), and which merely need to be improved or maintained at industry standard levels ("qualifiers")—clearly, these are vital strategic considerations.

Processes

Business processes received a good deal of attention in the 1990s with the advent of business process re-engineering. Business processes run horizontally across an enterprise's functional organization until they reach the ultimate recipient of the product or service offered—the customer. Michael Hammer, the re-engineering guru, advocates measuring processes from the customer's point of view—the customer wants it fast, right, cheap, and easy (to do business with). But is it really as simple as that? There are often many stages in a process. If the final output is slow, wrong, expensive, and unfriendly, how will we know which components of the process are letting it down? What needs to be improved? In the quest for data (and accountability), it is easy to end up measuring

everything that moves but learning little about what is important. That is one reason why processes need owners: to decide what measures are important, which metrics will apply, and how frequently they shall be measured and by whom, so that judgments can be made upon analysis of the data and actions taken.

Typically, organizations consider their business processes in four separate categories:

1. Develop products and services
2. Generate demand
3. Fulfill demand
4. Plan and manage the enterprise.

Within these categories, there are various sub-processes, which tend to be more functional in nature.

Processes are what make the organization work (or not, as the case may be). They are the blueprints for what work is done where and when, and how it will be executed. The aspects or features it will be critical to measure can normally be categorized as follows:

1. Quality (consistency, reliability, conformance, durability, accuracy, dependability)
2. Quantity (volume, throughput, completeness)
3. Time (speed, delivery, availability, promptness, timeliness, schedule)
4. Ease of use (flexibility, convenience, accessibility, clarity, support)
5. Money (cost, price, value).

These five categories will help to quantify the measurement criteria for the process issues that we identify as critical to success, i.e., How good? How many? How quickly? How easily? How expensive?

We should note, however, that not all critical processes are performed continuously or even regularly. Contingency processes such as disaster recovery, product recall, or various types of system failure (e.g., power outage, labor strike) will be executed rarely, if ever, but nevertheless need to be prepared for rapid deployment at any time, with formal procedures established. The key measurement issue here will normally be the level of readiness for action.

Additionally, when measuring processes, we need to consider the component parts of the individual process itself. All processes have four common characteristics: inputs, actions, outputs, and outcomes. Starting with the process outcomes and outputs first, we need to identify measures of the effectiveness of the process:

- Does the process deliver or produce what it is supposed to do [output]?
- How well does the process output perform for the recipient [outcome]?

Next, we need to consider the actions and inputs of the process by seeking measures of the efficiency of the process, for example:

- How long does it take to execute the process?
- How frequently does the process have to be executed?
- How much volume is processed (versus capacity)?
- What does it cost to perform the process?
- What are the levels of variability in—and into—the process?
- What is the level of wastage in the process?
- How flexible (e.g., multipurpose) is the process?
- How simple/complex (e.g., transactional/knowledge-based) is the process?
- How ready for deployment is a (e.g., rarely required but essential) process?

One further feature of process measurement is its ability to be measured at either a macro or a micro level. As mentioned previously, process owners may indeed want to see the big picture, but they will also need to be able to pinpoint exactly where quality, cycle time, and bottleneck problems are occurring.

Capabilities

Even the most brilliantly designed process needs people with certain skills, policies, and procedures about the way things are done, a physical infrastructure for it to happen, and, more than likely, some technology to enable or enhance it. Capabilities are bundles of people, practices, technologies, and infrastructure. Indeed, capabilities can be defined as the combination of an organization's people, practices, technology, and infrastructure that collectively represents that organization's ability to create value for its stakeholders through a distinct part of its operations.

Measurement of a capability usually focuses on those critical component elements that make it distinctive and also allow it to remain distinctive in the future. But that does not necessarily mean that those capabilities that only need to be as good as those of competitors do not need to be measured. How would an organization know that it is not beginning to fall behind in these areas if it does not measure them? Either way, competitive benchmarks will likely be needed in order to understand the size of the gap. Competitors will also be seeking ways to create value for a very similar set of stakeholders. The goalposts seldom remain static for very long.

Lest there be any confusion about what we mean here by capabilities, consider a common business process, such as the order-to-cash fulfillment process in an electronic products business. The customer places an order, the company makes and delivers it, and then gets paid for it. It is a single process with multiple components and implies the presence of at least six different capabilities.

These capabilities are the following:

1. Customer order handling capability
2. Planning and scheduling capability
3. Procurement capability
4. Manufacturing capability
5. Distribution capability
6. Credit management capability.

Each of these capabilities requires different skill sets, practices, technologies (although some IT systems will likely be multifunctional and integrated), and physical infrastructures, such as offices, factories, and warehouses.

Linking Strategies, Processes, and Capabilities

The performance prism thus helps to identify the critical components of strategies, processes, and capabilities that need to be addressed, from a performance measurement and management point of view, in order to satisfy the wants and needs of the various stakeholders and the organization. Obviously, organizations can and should choose which elements of these three perspectives or facets of the performance prism framework they should focus their performance management attentions on at any given time in their evolution. The performance prism is a tool flexible enough to allow that selection process to be adapted in the initiatives and focus that organizations elect to pursue. Figure 3 summarizes the application of these three perspectives.

An essential element is that these three facets or perspectives of the performance prism must be linked to each other in order to understand how they fit together toward satisfying the stakeholders' and the organization's wants and needs. Bob Kaplan and David Norton talk extensively about strategy maps in their latest book on the balanced scorecard, *The Strategy-Focused Organization*. But one of the advantages of the performance prism framework is that it makes explicit what elements should be covered in a strategy map. In traditional, balanced scorecard terms, a strategy map simply covers the four perspectives on the balanced scorecard—shareholders, customers, internal processes, and innovation and learning. But, in our view, this is too narrow. Success maps, as we prefer to call them, should cover all five facets of the performance prism, using the five vital questions as prompts.

An alternative method, or perhaps most appropriately as a means of validating the outputs from the success mapping process, is to apply what we call a failure mode map (or risk map). Failure mapping helps to check whether all the critical aspects of performance measurement have been properly addressed. In essence, this technique takes the reverse approach of a success map by identifying particular scenarios that describe the opposite of success—failure. By examining each key potential failure mode, a check can be made on the strategies, processes, and capabilities that relate to this risk and whether the measures identified are sufficient to enable mitigation of the risk's occurrence or its malevolence. To be warned should be to be prepared.

Kaplan and Norton promote the application of strategy maps, but they do not go far enough since they fail to break them down into their vital components—the potential for success and the potential for failure. Organizations have many opportunities but they also face several threats; their measurement systems must be able to capture both so that executives can manage the business with a clear view of both scenarios.

Applying the Performance Prism to Measures Design

To summarize then, we have identified five distinct but logically interlinked perspectives on performance

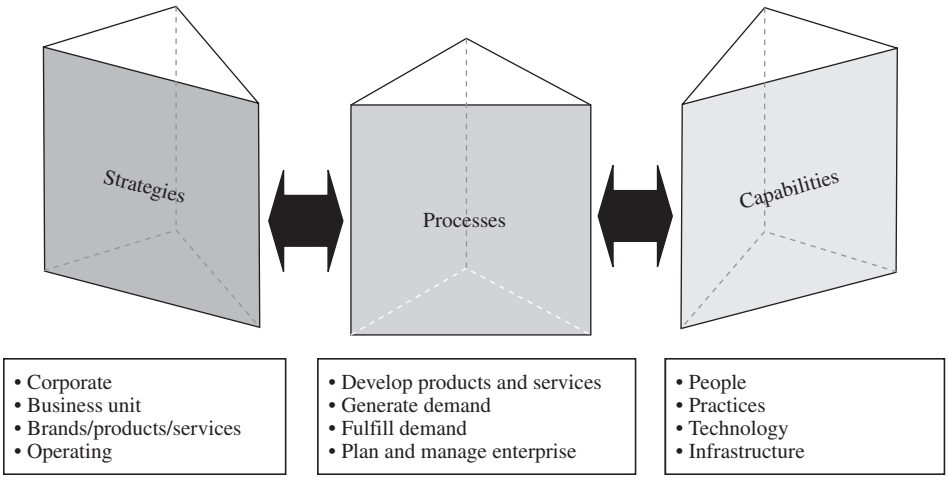


Figure 3

together with five vital questions to apply that will aid measurement design:

1. *Stakeholder satisfaction*: Who are the key stakeholders and what do they want and need?
2. *Stakeholder contribution*: What contributions do we require from our key stakeholders?
3. *Strategies*: What strategies do we have to put in place to satisfy these two sets of wants and needs?
4. *Processes*: What critical processes do we require if we are to execute these strategies?
5. *Capabilities*: What capabilities do we need to operate and enhance these processes?

As we have seen, these five perspectives on performance can be represented in the form of a prism. A prism refracts light. It illustrates the hidden complexity of something as apparently simple as white light. So it is with the performance prism. It illustrates the true complexity of performance measurement and management. Single dimensional, traditional frameworks pick up elements of this complexity. While each of them offers a unique perspective on performance, it is essential to recognize that this is all that they offer—a one-dimensional perspective on performance. Performance, however, is not one-dimensional. To understand it in its entirety, it is essential to view it from the multiple and interlinked perspectives offered by the performance prism (see Fig. 4).

The performance prism is not a cure-all tool. It will not solve all of the problems of performance measurement, and it needs to be used intelligently to optimize its potential. However, we believe that it does provide a robust and comprehensive framework through which the real problems and practical challenges of managing organizational performance can be viewed and addresses. This belief is born out of our experiences of successfully applying the performance prism within a wide variety of organizations.

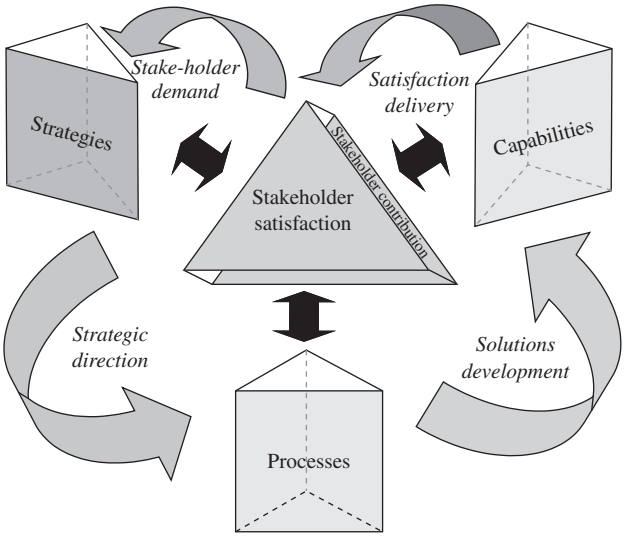


Figure 4

See Also the Following Articles

Business, Social Science Methods Used in • Critical Views of Performance Measurement • History of Business Performance Measurement

Further Reading

Centre for Business Performance Web site. http://www.cranfield.ac.uk/som/cbp

Neely, A., and Adams, C. (2001). The Performance Prism perspective. *J. Cost Manage.* **15,** 7–15.

Neely, A., Adams, C., and Kennerley, M. (2002). *The Performance Prism – The Scorecard for Measuring and Managing Business Success.* Financial Times, Prentice Hall.

Phenomenology

Shadd Maruna
University of Cambridge, Cambridge, United Kingdom

Michelle Butler
University of Cambridge, Cambridge, United Kingdom

Glossary

bracketing (or **Epoché**) Setting aside one's prior conceptions and experiences regarding how the world operates in order to be as open and receptive as possible to others' view of the world.

emergent themes Themes that arise out of the subjective accounts that bring to light certain concerns and motivations which may not be directly referred to by the person but which have an influence on the person.

interpretative phenomenological analysis A form of analysis concerned with how participants make sense of their social and personal worlds by examining the meanings that particular events, experiences and states hold for the participants.

intersubjectivity Mutual understanding or the overlap of agreement between different individuals' subjective experiences.

life world (or **Lebenswelt**) Each individual's subjective construction and understanding of the world around them, a personal narrative created with the help of building blocks and methods offered to him by others.

phenomenological reduction When one's life world is reduced by bracketing to that of a world of pure phenomena, the researcher's own viewpoint is suspended, and an effort is made to experience the world anew.

understanding (or **Verstehen**) The grasping of the subjectively intended meanings and symbolizing activities of others through empathy and deep listening.

Generally speaking, phenomenology is the study of lived experience and the subjective perceptions of human actors. Phenomenology seeks to understand how humans experience, make sense of and create meaning out of their existence. In a more specialized sense, phenomenology is also the name of a philosophical movement associated with the German philosopher Edmund Husserl (1859–1938) and the numerous social theorists influenced by him. This philosophical method concentrates on the careful analysis of conscious experience, without making assumptions about explanation, metaphysics, or other traditional philosophical issues. In the social sciences, phenomenological research involves detailed descriptions of consciousness and inner experiences from a first-person perspective. In other words, the "world" of concern to phenomenology is the "life world" or world as experienced and made meaningful in consciousness. In what follows, we will provide a closer look at the What, Who, When, Where, How, and Why of phenomenology, and conclude with an example of phenomenology in practice.

What Is Phenomenology?

Considerable definitional confusion surrounds the subject of phenomenology and this is not always aided by the primary sources of the phenomenological perspective. Phenomenological theorist Maurice Merleau-Ponty, for instance, answered the question of "What is phenomenology?" by arguing, "We shall find in ourselves, and nowhere else, the unity and true meaning of phenomenology."

Indeed, the term "phenomenology" is sometimes avoided, even by researchers engaged in phenomenological research. The word is not easy to pronounce quickly and can sometimes seem a pretentious way to describe "walking in another person's shoes." Moreover, some of

the theoretical work on phenomenology is found to be impenetrable for non-philosophers (and indeed some philosophers). Most importantly, though the term phenomenology seems to be avoided because it can mean so many different things in different contexts. The word phenomenology, after all, can refer to a philosophy, a paradigm, a method, or even an intellectual movement.

Yet, phenomenology can also simply denote an understanding of subjective experience or the study of human consciousness. The word phenomenon is rooted in the Greek term *phaenesthai*, to show itself, to flare up, to appear, and in its most basic form, phenomenology is simply about illuminating the perspectives of others. As such, one need not prescribe to the philosophical tenets of phenomenology as outlined by Edmund Hussrl or be a trained "phenomenologist" to be interested in the phenomenology of individuals in some situation. That is, one can speak legitimately about the "phenomenology of working class experience" (from Charlesworth) or the "phenomenology of criminal behavior" (from Katz) without being a philosopher or accepting the theory of mind posited by Husserl. Phenomenology in this sense simply refers to the description and understanding of lived, human experience through observable forms of immediate cognitive experience and reflective analysis.

In some ways, it is easiest to define phenomenology by first describing what it is not. Often the term phenomenology is used to indicate a contrast to traditional social science methodology, and indeed, phenomenology first emerged as a critique of paradigmatic approaches to ways of knowing in philosophy. Phenomenologists are opposed to both grand, over-arching social theories based on speculation as well as to the abstracted empiricism associated with some forms of positivist research (see also Mills). Moreover, whereas the positivist approach to social science is to seek the causes and explanations of social phenomena apart from the subjective states of the individual actors, phenomenologists are committed to understanding social phenomena from the actor's own perspective, and they prioritize description over explanation.

Indeed, one of the chief accomplishments of phenomenological research is to restore subjectivity from the derogatory status assigned to it in traditional positivist social science (e.g., the critique that something is "merely subjective"). Phenomenological theory insists that the subjective matters. It is not simply "measurement error" or idiosyncratic noise too variable or inaccessible to be studied. The most important "reality" to phenomenologists is the reality that people perceive and construct (their "life world" or "Lebenswelt") not some objective world. Phenomenology views human behavior, what people say and do, as a product of these definitions of reality. The ultimate (and probably unattainable) goal of interpretative phenomenological analysis is to experience the reality of another's consciousness through the achievement of "verstehen" or understanding. Particular attention is paid to the everyday lived experience of individuals, and subjects are typically treated as a whole and not reduced to variables.

Despite its attention to lived experience, however, phenomenological enquiry is not limited to understanding microsociological phenomena. Alfred Schutz, for instance, who is probably most responsible for introducing phenomenological philosophy to the social sciences, paid considerable attention to the role of historical, structural, and cultural factors in shaping consciousness. Likewise, although phenomenological psychologists tend to focus on individual persons as the creators of meanings in their lives, there is a universal recognition that these meanings are socially and culturally bounded.

History of Phenomenology: Who, When, and Where

The philosophy of phenomenology is over a century old, emerging in the mid-1890s in Germany and quickly spreading to Russia, Spain, and Japan prior to the First World War. As a method of social inquiry, phenomenology is most closely associated with the German philosopher Edmund Husserl (1859–1938), who wrestled with Kantian issues of epistemology (the philosophy of knowledge) in books like *The Idea of Phenomenology*. Other key proponents of phenomenology in philosophy include Merleau-Ponty, Martin Heidigger, and Jean-Paul Sartre.

This work was later popularized to the social sciences by Husserl's disciple Alfred Schutz (1899–1959) with his book *Phenomenology of the Social World*, first published in 1932, but then reprinted in 1967. It was with this latter edition, released at the height of an experimental and critical era in social science research, that phenomenology gained the most widespread acceptance among social science researchers. Unlike Husserl, Schutz focused on understanding the common sense reality of the everyday world, in particular the concept of "intersubjectivity" or the taken-for-granted assumptions and understandings that bind individuals together in a given culture or subculture. Although he was primarily a theorist rather than a sociological researcher, Schutz wrote two important, sociological treatments that apply a phenomenological perspective to sociological issues. For instance, written soon after he emigrated to the United States himself, his essay "The Stranger" focuses on the ways outsiders interpret and make sense out of a particular culture and attempt to orient themselves within it.

Yet, an interest in phenomenological questions is most certainly not limited to self-proclaimed philosophical phenomenologists. Phenomenology is closely tied to, if not wholly synonymous with, numerous other

approaches to social sciences including interpretivist sociology, symbolic interactionism, ethnomethodology, conversation analysis, social constructionism, existentialism, and humanistic-existential psychology. Practitioners from all of these perspectives share an interest in the subjective understandings of individual actors. Indeed, according to one of the founders of the movement, this interest in the construction of personal meanings was central to the "cognitive revolution" that swept throughout the social sciences during the latter part of the 20th Century.

How to "Do Phenomenology"

There is no single, "correct" method of conducting phenomenological research. Indeed, inherent in the philosophy itself is the suggestion that there should be no rule books or step-by-step guides for "doing" phenomenology. Still, in general, learning about another person's subjective reality tends to involve deep and empathetic listening to another, taking seriously his or her understandings of reality. As such, there is an emphasis on utilizing qualitative data collection techniques, such as ethnography and open-ended interviewing, that allow the researcher to remain close to the phenomena themselves and to convey in thick detail how it feels to live through some experience of reality. In this sense, phenomenological research tends to stress validity over reliability. However, empirical phenomenology sometimes employs independent raters to check intersubjectivity, sometimes utilizing a prior coding mechanism in the systematic content analysis of interview transcripts.

Crucially, the phenomenological researcher is urged to listen to others' accounts in their own terms, as free as possible of any of the researchers' own preconceptions and interferences. The maxim of philosophical phenomenology is "To the Things Themselves" (see Moustakas) and phenomenological understanding involves first "encountering" the perspective of others as if starting from a blank slate. This process of setting aside one's prior conceptions and experiences is called "bracketing" (or "epoché"). The aim is to achieve phenomenological reduction whereby the world is experienced anew. In this way, phenomenology is said to "begin in silence."

Finally, the analytic process in phenomenological work tends to be inductive, with patterns developing from deep immersion in the data. Subjects are usually sampled for theoretic reasons. Small groups of individuals who share common experiences (e.g., people who have experienced extended heroin use) will be interviewed to better understand the subjective experience (e.g., "phenomenology of heroin use"). Phenomenological analysis focuses on concrete descriptions of experiences and aims to reveal structures that are common to the group. At the same time, some of the identified patterns might be "emergent" or implicit themes that are inferred from what is said by subjects rather than that which is directly identified.

Why Phenomenology?

There are numerous reasons why introspective descriptions of individuals' conscious, inner worlds might be interesting and important to social scientists. Most pragmatically, understanding how individuals perceive the social world may help social scientists better explain and predict their behavior. Fundamental to the phenomenological approach is the symbolic interactionist truism, "If [people] define situations as real, they are real in their consequences." In the original passage containing this famous phrase, Thomas and Thomas are attempting to account for the unusual behavior of a convict at Dannemora prison who had murdered several persons because they "had the unfortunate habit" of talking to themselves on the street. They write, "From the movement on their lips he imagined that they were calling him vile names, and he behaved as if this were true." Even beyond such extreme cases, phenomenologists argue that "each individual extracts a subjective psychological meaning from the objective surroundings and that subjective environment shapes both personality and subsequent interaction" (see Caspi and Moffitt).

Phenomenology's usefulness in enhancing the predictive powers of causal explanation is not its only strength, however. Phenomenological theorists suggest that reaching verstehen (or phenomenological understanding) is "the key to understanding what is unique about the human sciences" (see Schwandt). Victor Frankl, for instance, argues that the human being's search for meaning is "a primary force in his life and not a "secondary rationalisation" of instinctual drives." Arguing against the determinism of positivist social science, phenomenologists like Frankl argue that human beings are ultimately self-determining. "Man does not simply exist but always decides what his existence will be, what he will become in the next moment." As such, from this perspective, a social science that does not include some account of this process of meaning construction and self-determination in human existence can hardly be considered a human science at all.

On the other hand, there are numerous reasons why phenomenological research is avoided by social scientists. Critics contend that phenomenological work cannot be empirically verified and is therefore antiscientific. Additionally, the practical relevance of largely descriptive phenomenological enquiry for the applied world of policy formulation is not always clear, compared, for instance, to correlational and variable-oriented research.

Phenomenology in Action: The Seductions of Crime

A recent example of the potential contribution of phenomenological research to the social sciences is Jack Katz's critically acclaimed *Seductions of Crime*. Prior to Katz's work, the study of crime had been overwhelmingly preoccupied with the background characteristics of individuals who commit crimes. Criminological research tends to ask whether some combination of variables (neighborhood of origin, IQ, personality traits, and so forth) can predict who will or will not become involved in crime. Such research, and criminological theory in general, tends to portray offenders as unfortunate, passive, over-determined products of their circumstances. As Katz and others point out, this image seems a long way from the excitement, danger and, sometimes, sheer joy of the experience of offending itself. Moreover, Katz argues:

> *The statistical and correlational findings of positivist criminology provide the following irritations to inquiry: (1) whatever the validity of the hereditary, psychological, and social-ecological conditions of crime, many of those in the supposedly causal categories do not commit the crime at issue, (2) many who do commit the crime do not fit the causal categories, and (3) what is most provocative, many who do fit the background categories and later commit the predicted crime go for long stretches without committing the crimes to which theory directs them. Why are people who were not determined to commit a crime one moment determined to do so the next?*

Katz argues for a shift from the social and psychological background factors in the criminal equation to a focus on the phenomenological foreground of crime, the seductive qualities of crime or the aspects that make crime a compelling, sensual, even morally transcendant pursuit. Utilizing analytic induction, Katz seeks to construct phenomenological understandings of a variety of criminal acts from shoplifting, to vandalism, to cold-blooded murder, in each case asking, "What are people trying to do when they commit a crime?" He suggests that this shift from background to foreground factors will explain more variation in criminality than background correlations allow.

Just as importantly though, Katz's work, unlike the majority of criminological research, provides an intimate and at times disturbing insight into the emotional and cognitive experience of crime. Getting "inside the minds" of serial killers, rapists and other offenders is a favorite pursuit of pop psychologists, novelists and scriptwriters, of course, and the public seems to hunger for the understanding such works provide. Yet, for the most part these sensationalized accounts are not based on even the barest research evidence. *Seductions of Crime*, on the other hand, is the product of a systematic analysis of what amounts to a mountain of research evidence. Katz draws on a wide array of published and unpublished ethnographic data, autobiographies, and collections of first-person accounts in constructing each of his phenomenological accounts.

Far from the last word on the phenomenology of any of these criminal pursuits, Katz's work has opened the door for a new type of criminology, focused not on prediction, but on understanding. The result is a much deeper understanding of the criminal as a person, rather than crime as a product of multiple variables interacting. Katz does not offer this portrait as a critique of criminological research on the background factors associated with crime, but rather as an essential complement to this literature. Phenomenology, in this case and others, seeks to put the "human" in the human sciences.

See Also the Following Article

Criminology

Further Reading

Bruner, J. (1990). *Acts of Meaning.* Harvard University Press, Cambridge, MA.

Caspi, A., and Moffitt, T. (1995). The continuity of maladaptive behaviour: from description to understanding in the study of antisocial behavior. In *Developmental Psychopathology, Vol 2: Risk, Disorder and Adaptation* (D. Cicchetti and D. J. Cohen, eds.). Wiley, New York.

Charlesworth, S. J. (2000). *A Phenomenology of Working-Class Experience.* Cambridge University Press, Cambridge, UK.

Frankl, V. E. (1964). *Man's Search for Meaning: An Introduction to Logotherapy.* Hodder and Stoughton, London.

Hein, S. F., and Austin, W. J. (2001). Empirical and hermeneutic approaches to phenomenological research in psychology: A comparison. *Psychol. Methods* **6,** 3–17.

Katz, J. (1988). *Seductions of Crime.* Basic Books, New York.

Maruna, S. (2001). *Making Good: How Ex-Convicts Reform and Rebuild Their Lives.* American Psychological Assoc., Washington, DC.

Merleau-Ponty, M. (1962). *Phenomenology of Perception* (trans. C. Smith). Routledge and Kegan Paul, London.

Mills, C. W. (1959). *The Sociological Imagination.* Oxford University Press, London.

Moustakas, C. E. (1994). *Phenomenological Research Methods.* Sage, Thousand Oaks, CA.

Psathas, G. (1973). Introduction. In *Phenomenological Sociology.* (G. Psathas, ed.). Wiley, New York.

Schutz, A. (1964). *Collected Papers II.* Nijhoff, The Hague.

Schutz, A. (1970). *On Phenomenology and Social Relations.* Selected writings. University of Chicago, Chicago.

Schwandt, T. A. (1998). Constructionist, interpretivist approaches to human inquiry. In *The Landscape of Qualitative Research: Theories and Issues* (N. K. Denzin and Y. S. Lincoln, eds.), pp. 221–259. Sage, Thousand Oaks, CA.

Thomas, W. I., and Thomas, D. S. (1928). *The Child in America: Behavior Problems and Programs.* Knopf, New York.

Phrase Completion Scales

David R. Hodge
University of Pennsylvania, Philadelphia, Pennsylvania, USA

David F. Gillespie
Washington University, St. Louis, Missouri, USA

Glossary

coarse data Coarse data has units of information that cannot be specified as equal. Ordinal-level data, for example, are more coarse than interval-level data because the units of information that comprise ordinal data cannot be assumed equal, whereas the units of information that comprise interval-level data are assumed equal.
confounding Confounding occurs when it is impossible to separate the effects of two or more variables from each other. In measurement, this happens when two or more dimensions are present in a single question.
construct An idea that identifies a phenomenon that is not directly observable, such as attitudes.
Likert items Represent a questionnaire format developed by Rensis Likert that is widely used to measure attitudes. Likert items have an introductory stem consisting of a positively or negatively stated proposition followed by a graduated response key composed of adverbs (e.g., "strongly") and verbs (e.g., "agree").
superimposition A dimension implied by the response key that differs from, and consequently is imposed over, the dimension implied by the question's introductory stem.
unidimensional Concepts that have only one source of variation. More concretely, a measure of a single attribute of a construct.
univocal stimulus A concise, unambiguous question or statement that conveys a single meaning.

Phrase completion scales are concise, unidimensional measures designed to tap ordinal-level data in a manner that approximates interval-level data. The phrase completion method of constructing scales was developed as an alternative to Likert scales. Although the Likert method is widely used, a number of disadvantages are associated with this approach to measuring sentiments. After introducing the phrase completion method and the assumptions underlying its development, the disadvantages associated with the Likert method are discussed. The delineation of these disadvantages—namely multidimensionality, multivocal stimuli and unsubstantive responses, and coarse data—highlights the comparative advantages provided by the phrase completion method. The article concludes with a brief summarization of the limitations and strengths of the phrase completion method.

Introduction

The phrase completion method was developed to improve the measurement of personality, social, and psychological sentiments. Specifically, the phrase completion method represents an attempt to either meet or more precisely approximate the assumptions underlying measurement and statistical methodologies. Three assumptions in particular animated the development of the phrase completions. The first two—unidimensionality and univocal stimulus—are basic to all measurement. The third—the assumption of normally distributed interval- or ratio-level data—is unique to parametric statistics.

Unidimensionality

Unidimensionality is an underlying assumption shared across all levels of measurement. The individual items that comprise measurement scales are commonly classified as nominal, ordinal, interval, or ratio. Nominal data

refers to discrete, categorical units of information that are not ordered in any hierarchical format (e.g., race, gender, and ethnicity). Ordinal-, interval-, and ratio-level data all rank units of information in a hierarchical arrangement. Unit *C* is greater in magnitude than unit *B*, which in turn is greater in magnitude than unit *A*. The distinguishing characteristic between ordinal and interval data concerns the magnitude or the distance between the units of information. With interval data, the units of information are identical; the distances between *A*, *B*, and *C* are the same. With ordinal data, the units of information are typically unequal; there is no guide as to the distance between *A*, *B*, and *C*. In other words, ordinal-level data consists of ordered categories of varying magnitude, whereas interval data consists of a series of equally spaced units.

All four levels of measurement, however, should be characterized by unidimensionality. In other words, each individual item in a scale should grade a population only on a single, given quality. Units *A*, *B*, and *C*, must reflect information about a distinct, discrete entity.

This is a particularly significant assumption since the existence of multidimensionality in an item raises the issue of confounded dimensions. If an item taps more than one dimension, it is not a valid measure. If an item purports to assess the level of dimension *X* but in actuality assesses some level of dimension *X* and some level of dimension *Y*, then it is impossible to distinguish the effects of *X* from the effects of *Y*. Consequently, it is impossible to establish a relationship between constructs since the researcher cannot ascertain the extent to which each item is measured.

It is important to note that even among what are commonly referred to as multidimensional measures, each theorized dimension or domain must exhibit unidimensionality. Multidimensional measures designed to assess the dimensions *A*, *B*, and *C* of a particular construct must not assess additional dimensions nor overlap with each other. In other words, the items designed to assess the level of dimension *A* must not concurrently assess the level of an additional dimension *D* nor be confounded with the items designed to assess *B* and *C*. Each individual item must be unidimensional.

Univocal Stimulus

Measurement is based on the premise that a univocal or single stimulus in the form of a set of questions can tap a given construct. A critical assumption is that all individuals understand the stimulus in the same way so that different responses are due to real differences in the amount of the construct among individuals. In short, the stimulus must be univocal rather than multivocal. If individuals encounter multiple stimuli, the internal logic of measurement is compromised since any apparent difference may be due to the different stimuli rather than a reflection of real differences in the construct of interest.

Questions must be clear and concise if they are to produce substantively meaningful results. If a question is vague and ambiguous, understandings of the question can vary. Because individuals receive multiple stimuli, different respondents can answer the same question differently due to varying interpretations. Similarly, if items are too complex, respondents that are apathetic, hurried, tired, from different cultures, or who have lower levels of education may understand stimuli differently. In such cases, variation in the construct of interest may be due to the complexity of the questions.

It is important to note that, at least to some extent, multivocalilty is a problem that can never be completely eliminated. Each individual receives a slightly different stimulus from any given question due to differences in age, class, gender, spiritual orientation, etc. Given these realities, the intent should be to produce items that come as close as possible to approximating univocal stimulus. To achieve substantively meaningful results, it is important to construct concise, unambiguous questions that reduce cognitive complexity.

Normally Distributed Interval- or Ratio-Level Data

Concise unidimensional items are fundamental to all statistical procedures. The most widely used procedures, however, are parametric statistics, such as regression, analysis of variance, or *t* tests. Additional assumptions are associated with the use of these statistics.

Specifically, parametric statistics are based on the assumption that interval- or ratio-level data with a normal distribution are used. In other words, parametric statistics require the use of data that are at least interval level. Due to the subjective nature of human attitudes, it is difficult to obtain interval-level data on sentiments. Consequently, in practice, ordinal-level data are commonly used with parametric statistics. This practice is based on the widespread recognition that many parametric statistics are robust and can withstand some violation in the underlying assumptions. Nevertheless, the violation of assumptions can result in errors, such as the failure to detect significant effects (e.g., type II errors).

The phrase completion method was developed as a concise, unidimensional approach that provides a closer approximation of interval-level data. Although from a technical standpoint the phrase completion method produces ordinal measures, it is important to note that variation exists in ordinal measurement. This variation can be understood as a continuum. At one end of the continuum, the data are very coarse. The difference between the units of information varies substantially. At the other end

of the continuum, the data are quite fine. The difference between the units of information varies minimally. At the latter end of the continuum, the units of information approximate the equal spacing that characterizes interval-level data. It is this type of relatively fine, ordered data that phrase completions were designed to capture.

Construction of Phrase Completions

Phrase completions consist of a phrase followed by an 11-point response key. The phrase introduces part of a concept. Designating a reply on the response key completes the concept initiated by the phrase.

The response key represents the underlying continuum of the construct that the completed phrase is designed to tap. In theory, most constructs exist along a continuum, ranging from the absence of the construct in question to some theorized maximum amount of the construct in question. This continuum is operationalized and specified in the 11-point response key. Specifically, verbal anchors are typically placed on the ends of the response key. The number 0 is used to signify the absence of the construct and 10 is used to signify the maximum amount of the construct.

Presented with the underlying continuum of the construct in question, the respondent selects a choice along the continuum that completes the phrase. This approach produces a concise, unidimensional measure. By delineating the underlying continuum of the construct in question, respondents can select the point along the continuum that best reflects their views. Validity is enhanced since the theoretical continuum is presented to the respondent in the measure. Similarly, the delineation of the continuum helps to ensure that the measure is unidimensional.

When numbers are presented in response keys, respondents tend to use them to guide their thinking. The presence of integers helps respondents view the range of available options. By delineating the theoretical continuum as a series of integers, which are by nature equally spaced, respondents may be more inclined to view their sentiments in terms of a continuous variable. The integers work in concert with the anchor phrases to directly imply the degree of the attribute: Zero equals the absence of the construct, 10 the maximum, and the intervening numbers signify increasing amounts of the construct. As discussed later, this type of formatting engenders data that approximate interval-level data.

Table I provides an example of the phrase completion method. Standard orienting material is provided along with three questions that tap various aspects of spirituality. As can be seen, the underlying theoretical continuum has been specified in the response keys. For example, with the first question the responses range from the absence of the attribute—never aware of God's presence—to a maximum of the attribute—continually aware of God's presence—in a unidimensional manner. The reader is informed in the orientation that the ends of the continuum depict extremes and the middle values represent the more frequently occurring amounts of the variable in question. In some cases, it may be helpful to use a third anchor phrase in the center of the response key to accent the fact that middle values represent moderate response options.

The response key is reversed for every second question to encourage thoughtful responses and to reduce any response set bias that may exist. The scores of the questions that comprise the scale are summed to create

Table I Illustrating the Phrase Completion Method

The following questions use a sentence completion format to measure various attributes. An incomplete sentence fragment is provided, followed directly below by two phrases that are linked to a scale ranging from 0 to 10. The phrases, which complete the sentence fragment, anchor each end of the scale. The 0 to 10 range provides you with a continuum on which to reply, with 0 corresponding to absence or zero amount of the attribute and 10 corresponding to the maximum amount of the attribute. In other words, the end points represent extreme values, whereas 5 corresponds to a medium, or moderate amount of the attribute. Please circle the *number* along the continuum that seems to best reflect your intial feeling.

1. I am aware of the presence of God or the Divine

never									**continually**
0	1	2	3	4	5	6	7	8	9	**10**

2. I spend periods of time in private spiritual thought and meditation

every day, without fail									**never**
10	9	8	7	6	5	4	3	2	1	**0**

3. In terms of the questions I have about life, my spirituality answers

no questions									**absolutely all my questions**
0	1	2	3	4	5	6	7	8	9	**10**

a total score for the attribute in question. This basic approach is used with all phrase competitions.

As previously implied, phrase completions were developed in response to the disadvantages associated with traditional ordinal measures, most notably Likert scales. Consequently, the advantages of phrase completions are particularly evident when they are compared with the Likert method. As a precursor to this comparison, an overview of the Likert approach is provided.

Overview of Likert Scales

Likert or Likert-type scales are perhaps the most popular method for measuring attitudes. They were originally introduced by Rensis Likert as an alternative to the more time-intensive and judgment-based Thurstone approach to attitude scaling. Indeed, their ease of construction, intuitive appeal, adaptability, and usually good reliability have been key factors in fostering their widespread use.

In contemporary usage, the Likert method consists of a declarative stem followed by a response key. The stem clearly expresses a positive or negative opinion (e.g., "Quite often I have been aware of the presence of God or the Divine"). Typically, individuals are asked to indicate their level of disagreement or agreement regarding the stated opinion on a 5-point response key ranging from "strongly disagree" to "disagree," "neither disagree or agree," "agree," and "strongly agree."

Agreement with a positively stated proposition is considered to indicate the presence of the underlying construct. The responses are equated with integers (e.g., strongly disagree = 1 and strongly agree = 5), and negatively worded items are reverse scored. The items are summed, creating an index, which is hypothesized to indicate the degree to which the respondent exhibits the trait in question.

Although the agree/disagree format is perhaps the most common form of Likert scale, other types of response keys are also frequently employed. For instance, Likert originally used a "strongly disapprove/strongly approve" continuum. However, the basic approach—a positively or negatively stated proposition followed by a graduated response key composed of adverbs and verbs—is commonly viewed as the distinguishing hallmark of Likert scales.

Although Likert scales are widely used, a number of little known disadvantages are associated with their use, perhaps most prominently multidimensionality, multivocal stimuli and unsubstantive responses, and coarse data. The various facets of these disadvantages are delineated next along with some of the approaches that have been proposed to address the difficulties. Also highlighted is the manner in which phrase completions circumvent these disadvantages.

Multidimensionality

Perhaps the most prominent disadvantage associated with Likert scales is multidimensionality. The multidimensionality associated with superimposition and the confounding of the direction and intensity dimensions exist in all Likert scales. However, an additional set of difficulties is evident in odd-numbered scales, such as the widely used 5-point scales that employ a midpoint. All three areas are discussed next, and an exposition of the advantages of the phrase completion method follows the discussion of the second and third areas.

Superimposition

As noted previously, Likert stems express opinions. In effect, the stem asks respondents to think along a given dimension. For example, a stem in a spirituality survey that states, "Quite often I have been aware of the presence of God" asks individuals to conceptualize along a dimension that might be called awareness of God's presence.

In place of a response key that taps into the dimension suggested by the stem, the Likert response key superimposes an alternative dimension (e.g., agree/disagree) over the dimension suggested by the stem. Respondents must indicate their response on this alternative dimension that is imposed over the original dimension suggested by the stem. To follow up on the previous example, no opportunity exists to directly indicate the extent to which one experiences the awareness of God's presence. Multidimensionality exists because the response key superimposes a dimension that differs from the dimension suggested by the stem.

The existence of multidimensionality raises a number of concerns. Most prominently, the assumption of unidimensionality in measurement is violated. In essence, the measurement process is contaminated by the existence of multiple dimensions. However, in addition to superimposing a different dimension on the stem, Likert response keys also confound two distinct dimensions in the response key.

Confounding Direction and Intensity

Attitudes can be expressed in different ways. For example, two different attitudinal dimensions are direction or content and intensity or strength. Direction pertains to having either a positive or a negative view toward an entity, whereas intensity refers to the degree of conviction with which the view is held. The distinction between the two dimensions can be seen in the fact that a person may disagree with a certain proposition with a great deal of conviction or only mildly.

In addition to superimposing a dimension on the stem, Likert scales confound the direction and intensity dimensions. The response key on Likert items incorporates both direction (agree or disagree) and intensity (strongly or not strongly). Consequently, by virtue of their design, Likert scales ask respondents to think across three dimensions: the dimension suggested by the stem, direction, and intensity.

To address the problem of confounding direction and intensity, researchers have attempted to separate the two dimensions by using a two-question format. Individuals are first asked whether they agree or disagree with a given proposition, and then a separate follow-up question is used to assess the level of intensity. This procedure has produced mixed results, at least in part because the response key continues to superimpose a different dimension on the stem. Some researchers report little improvement, suggesting that confounding remains a problem because the use of a sequential format continues to confound the two dimensions. However, even in instances in which the findings have been more positive, the two-question approach has limitations. The number of questions is doubled, for example, resulting in larger expenditures of resources for both researchers and potential respondents. Finally, it should be noted that the multidimensionality associated with superimposition remains an issue.

Phrase completions circumvent these disadvantages. As can be seen in Table I, the multidimensionality associated with superimposition is avoided by the phrase completion method. To follow up on the previous example, as is evident with item 1 in Table I, respondents are asked to conceptualize along a single dimension—awareness of God's presence. Accordingly, the confounding of direction and intensity is eliminated by asking individuals to think along a single dimension. Similarly, the process of operationalizing the underlying continuum in the response key helps to provide substantively meaningful options while simultaneously upholding the assumption of unidimensionality in measurement items.

The Confounded Midpoint

The confounding of direction and intensity is also a relevant issue with the use of the midpoint. Likert scales commonly employ the use of an option, such as "undecided" or "neither disagree or agree," that is placed in the center of the response key. When numbers are assigned and the items are summed, midpoint responses are accorded a value of 3 on a 5-point scale.

The use of a midpoint is not universal, however, due in part to confusion over exactly what is being signified by its selection. Based on the value that a midpoint response receives during the summation process, it is designed to represent a neutral reply. In other words, it denotes a midpoint along the intensity or strength dimension, halfway between 1 and 5.

Many respondents, however, understand the midpoint in terms of the direction or content dimension. In other words, it represents an alternative to expressing agreement or disagreement. For individuals unable to decide whether they agree or disagree with the declarative stem, the midpoint functions as a means to express an option such as "don't know," "no opinion," or "haven't thought about it."

When the midpoint is used as part of the direction dimension, it should be given no score and removed from the summation process. However, since it is impossible to determine which dimension respondents are basing their response on, due to the confounding of direction and intensity in the midpoint, it is impossible to remove their scores. Thus, in many instances, individuals are being assigned a value that indicates that they have a medium amount of the trait being assessed when in realty they exhibit no amount of the trait.

In recognition of this problem, it has been suggested that a separate, distinct, "no opinion" response be added to filter out the appropriate respondents. This approach, however, has produced mix results. If a discrete "no opinion" response option is offered, a number of respondents who are unsure about the degree of their intensity appear to select the "no opinion" response as a means of avoiding the mental work associated with making a decision. In lieu of such an option, respondents expend the required mental effort to ascertain the exact point along the intensity continuum that best reflects their views. Consequently, a number of researchers believe that the costs of incorporating a "no opinion" response outweigh the benefits, unless there is reason to believe that a large number of respondents will be completely unfamiliar with the topic under investigation, in which case a "no opinion" filter is advisable.

In contrast, the phrase completion method circumvents the confounding of direction and intensity that occurs with the use of a midpoint. As illustrated in Table I, phrase completions employ no midpoint. Consequently, no confusion exits in the respondent's mind concerning which dimension the midpoint reflects. A clear zero option is presented that represents the absence of the trait in question. In addition, researchers still have the option of incorporating a "not applicable" option when they believe it is warranted. Consequently, the measurement error that results from inappropriately attributing values to respondents is minimized.

Multivocal Stimuli and Unsubstantive Responses

As noted previously, questions should be clear and concise with the goal being the production of substantively meaningful results. At least to some extent, Likert scales

fail to meet these criteria. Specifically, Likert scales are associated with increased cognitive complexity and the production of results of questionable validity, particularly when negative wording is used. The next two sections address the former concern, and the following section addresses the latter issue. Then, an illustration of how the phrase completion method addresses the disadvantages associated with negative phrasing is presented.

The Complexity Innate in Multidimensionality

The multidimensionality inherent in Likert items increases the level of cognitive complexity for respondents. Individuals are asked to simultaneously conceptualize along three dimensions. Individuals have to keep in mind the stated opinion while simultaneously evaluating whether they agree or disagree with the proposition and the strength with which they hold their views. The added cognitive load that occurs when respondents are asked to think across three dimensions can result in measurement error.

Negatively Worded Stems and Added Confusion

An added level of cognitive complexity is related to the use of negatively worded stems (e.g., stems that incorporate the word "not" and similar variations). Likert scales commonly incorporate the use of negatively worded items, sometimes referred to as item reversals or simply reversals, to guard against response set or acquiescent bias—the tendency to agree with a set of positively worded items. These items are interspersed with positively stated propositions and then reverse scored before summing all the items to achieve a total score.

The use of negatively worded stems, however, increases the level of cognitive complexity. Consider, for example, the following Likert item intended to assess the level of respondents' spirituality: "I never spend periods of time in private spiritual thought and meditation." Respondents must disagree with this statement to record scores indicating an elevated level of spirituality. Thus, in addition to having to think across three dimensions, individuals have the added cognitive load of having to conceptually synthesize a double negative (never spend time) and disagree in order to be coded as exhibiting spirituality.

In short, many individuals have difficulty expressing agreement with a given construct by disagreeing with a negatively worded item. The effects on validity and reliability may be most pronounced among children, the elderly, immigrants from different cultures, people with low levels of education, and situations in which respondents exhibit a low level of engagement with the questions.

Validity Issues Associated with Negative Stems

As noted previously, acquiescent bias provides the rational for using negatively worded stems that are interspersed with positively worded stems. If the number of negative and positive stems is identical, individuals who indiscriminately agree with every item end up, at least in theory, in the middle of the scale, with an average value of 3 on a scale of 1–5.

It is debatable, however, if individuals who respond in such a manner warrant a score in the middle of the scale. If these people could be induced to respond in a more thoughtful manner, it seems probable that they would often record scores that place them in other positions on the scale. In other words, individuals would record a variety of positions. However, by virtue of their design, item reversals reduce this variety to a single score.

Various methods have been used to address the disadvantages associated with the use of negative wording. Some researchers have attempted to completely eliminate item reversals and state all stems using positive wording. For example, some researchers have attempted to develop items that tap a mirror image of the construct being studied. In theory, respondents agree with the positively stated items that reflect the construct and disagree with the positive stems that reflect the hypothesized mirror image of the construct. The items in the latter group are then reverse scored. Developing such items, however, is a difficult process and questions exist about the unidimensionality of the trait and its hypothesized mirror image.

Other researchers have recommended using all positively worded stems and reversing the response key. In other words, odd-numbered questions might have a response key ranging from strongly disagree to strongly agree, whereas even-numbered items would have the response key ranging from strongly agree to strongly disagree. Still other researchers, noting the tendency of respondents to rate positively worded statements more favorably, continue to recommend the use of positive and negatively worded stems as the most optimal method, even though this approach may result in two-factor, multidimensional outcomes.

Phrase completion scales circumvent these difficulties by eliminating the use of negative wording in both the stem and the response key. Consequently, the confusion associated with having to synthesize a double negative (e.g., never spend time and disagree) is eliminated. Following up on the spirituality item discussed previously,

item 2 in Table I illustrates how this Likert item might be stated using the phrase completion method. As can be seen, by presenting a concise, unidimensional item to individuals, the level of cognitive complexity is reduced.

Regarding the validity issues associated with negative stems discussed previously, having respondents indicate where they fall on the underlying continuum likely fosters responses of greater validity. By removing the option to agree with a positively worded statement, obviously respondents can no longer indiscriminately agree with such statements. Instead, thoughtful interaction with each item is encouraged, a dynamic that is abetted by reversing the response key for alternating questions. By directly presenting individuals with the underlying construct, researchers are, in a manner of speaking, placing their theoretical cards on the table. This encourages respondents to directly select the option that corresponds to their views, fostering more substantively meaningful responses.

Coarse Data

Although the use of Likert data with nonparametric statistics is appropriate, the use of Likert scales with parametric statistics is widely considered problematic. The coarse nature of Likert data results in a violation of the assumptions that inform the use of parametric statistics. The coarse nature of Likert data is apparent in at least three areas: significantly unequal units, limited number of units, and restricted range. The advantages offered by the phrase completion method are noted at the end of each section.

Unequal Units

By definition, ordinal-level data are composed of unequal data units. As noted previously, however, variation exists in terms of the equality of data units. Some ordinal data are very coarse with grossly unequal data units. Conversely, other ordinal data may approximate interval data with measurement units that are relatively uniform. Consequently, measurement units should be as similar as possible to collect the highest amount of information. Indeed, the concept of normality of distribution becomes meaningless when the distances between scale points are unknown.

A number of studies have been conducted to examine the difference between units with verbal labels that have some relationship with probabilities. Typically, these studies present individuals with a series of verbal labels and ask them to assign a numerical value ranging from 0 to 100 that best reflects the magnitude of the label. Since the studies use labels that have some type of link with probabilities, they represent a more favorable scenario than exists with the widely used disagree/agree format.

Table II presents a series of these studies collected by Krosnick and Fabrigar that examined the numerical values attributed to a continuum ranging from excellent to very poor. These studies were conducted over the course of approximately 20 years, spanning a time period from 1968 to 1991. As can be seen, the actual numerical values assigned vary across studies in most cases, a fact that emphasizes the imprecise nature of such labels. Furthermore, with the possible exception of study 4, the difference between verbal labels varies considerably. In no study are all the differences between units equivalent. In a number of instances, the difference between the smallest unit and the largest differs by a factor of two and ranges as high as a factor of eight. In study 3, for example, the difference between "good" and "poor" was eight times larger than the difference between "poor" and "very poor." Similarly, the difference between "very good" and "good" was four times larger than the difference between "very poor" and "poor."

Likert scales may yield coarser data than are obtained with labels that are regularly associated with probabilities in common usage. A sentiment such as agreement may be more difficult to parse into equally spaced units due to its amorphous nature. Feelings of intensity would seem to

Table II Difference in Units for Likert-Type Labels

	Study 1		*Study 2*		*Study 3*		*Study 4*	
Verbal label	*Assigned value*	*Unit difference*	*Assigned value*	*Unit difference*	*Assigned value*	*Unit difference*	*Assigned value*	*Unit difference*
Excellent	93		91		91		99	
		15		9		11		26
Very good	78		82		80			
		11		12		22		
Good	67		70		58		76	
		24		21				25
Fair	43		49			42	48	
		22		32				25
Poor	21		17		16		23	
		9		12		5		22
Very poor	12		5		11		1	

be more difficult to quantify than labels that are regularly used in conjunction with numerical values. Many individuals, for instance, have received test scores in academic environments accompanied with labels ranging from excellent to very poor. Conversely, few forums exist in which respondents would have gained experience quantifying sentiments such as agreement.

In addition, the problems of multidimensionality and multivocal stimuli likely accent the difference between Likert units. As previously implied, individuals generally respond differently to negative and positive wording, which in turn complicates attempts at quantitative equivalence. Consequently, it is doubtful that units of agreement are equivalent to units of disagreement. Similarly, the confounding of direction and intensity dimensions suggests that a smooth continuum of intensity is nonexistent in many cases since the direction dimension may be more salient with many respondents. This is particularly the case with the midpoint response option that respondents may view as a "haven't thought about it" option instead of a true neutral option on the intensity dimension. Additionally, the added cognitive complexity associated with the use of negatively worded items suggests that such items are quantified differently than positively stated items. These factors suggest that the units in Likert response keys may vary greatly.

In contrast, phrase completions were designed so the units of measurement are quite similar. The concise, unidimensional nature of the phrase completion method skirts the problems of multidimensionality and multivocal stimuli. By using equally spaced integers for the majority of the response key, respondents are guided toward thinking about the entity of interest in the form of a series of uniform units. The impreciseness associated with verbal labels is replaced with quantitatively similar integers. This measurement approach likely provides a closer approximation of interval-level data.

Limited Number of Units

Another facet of coarseness is the number of units. Interval-level data are commonly understood to be composed of a series of equally spaced units. Scales that consist of few units do not allow respondents to make distinctions along a continuum. Furthermore, in order for a normal distribution to exist, there must be a sufficient number of units in the scale to comprise a meaningful distribution.

The five points commonly used in Likert scales represent a relatively coarse response key. In other words, relatively few units of information exist. Although a number of researchers have attempted to subdivide Likert keys into a greater number of units (e.g., strongly disagree, disagree, and slightly disagree), this approach has met with limited success.

The response options in a given key should correspond to something that exists in respondents' actual experience. For the collected data to have substantive meaning, the options must coincide to respondents' phenomenological reality. Although the agreement/disagreement sections of the Likert scale could be further divided, it is questionable to what extent respondents are able to parse their intensity of agreement about a given opinion into fine discrete categories that have substantive meaning. Indeed, one of the reasons the standard five-point key is so widely used is because many researchers believe its response options are consistent with individuals' actual experiences.

Consequently, further sectioning of the intensity dimension may actually increase measurement error because the added options have no anchor points in respondents' experience. Respondents may skip over response options that hold little meaning or use finely graded terms interchangeably. In other words, terms such as agree and slightly agree may be selected at random with no real distinction made in the respondent's mind about the two terms.

In short, Likert response keys offer only a limited number of units. Although the data are mixed, a number of studies that have explored the validity and reliability of Likert response keys suggest that either 5- or 7-point scales represent the limit of the method. Respondents can only parse their views toward a given declarative statement into so many substantively meaningful units.

Phrase completions take a different route in terms of developing a response key with more units. In place of attempting to parse sentiment about a given statement into increasingly finely ordered categories, the underlying continuum is operationalized in the response key using a 0—10 scale. The units of this continuum are likely to hold more substantive merit to respondents than units of agreement concerning a given declarative statement. The phrase completion method is likely to produce more substantively meaningful options for respondents, at least in part due to the restricted range associated with Likert items.

Restricted Range

Another manifestation of coarseness is data that only tap a portion of the available range. As noted previously, most constructs exist along a continuum from the absence of the attribute in question to some maximum amount of the attribute. Ideally, measurement items should assess the full extent of this continuum.

However, by virtue of their design, Likert scales only tap a portion of the extant continuum. Likert stems express either a clearly positive or negative opinion. Using positively stated propositions as an example, individuals are expected to agree with positively stated stems

to the degree to which their attitudes are more positive than the opinion reflected in the stems. Agreement with the expressed opinions is posited to indicate the presence of the underlying construct.

Consequently, stems are crafted to reflect a moderate opinion. For example, the stem "Spirituality is especially important to me because it answers many questions about the meaning of life" contains the word "many" to moderate the opinion. If the opinion stated in the stem is too strong (e.g., "Spirituality answers *all* my questions"), then relatively few respondents will express agreement with the proposition. Opinions must be of a moderate nature in order for sufficient numbers of respondents to feel comfortable selecting the various agreement options. In other words, the opinion must be moderate enough so that the item is able to adequately distinguish individuals across the hypothesized continuum.

The moderate nature of the opinion, however, results in the loss of sentiment at the ends of the continuum. Moderately stated stems do not distinguish individuals who hold relatively extreme attitudes. To follow up with the previous spirituality stem, for example, devout individuals for whom spirituality does answer all their questions about life would select "strongly agree" just as respondents whose spirituality answers many questions about life would select "strongly agree."

The Likert method offers no means of tapping the ends of the attitude continuum. As previously implied, using items that express a strongly stated opinion will result in scores that fail to vary sufficiently across a population. A stem that articulates a strong opinion, however, is the only way to capture relatively extreme sentiments (e.g., to distinguish those individuals for whom spirituality answers all questions). The Likert method is only able to tap a restricted range of the underlying continuum.

In contrast, the phrase completion method operationalizes the underlying continuum in the response key. As can be seen with item 3 in Table I, this approach captures extreme sentiments that traditional Likert scales fail to assess. Since the complete continuum is presented to respondents, individuals who hold views along all points of the continuum can select the option that reflects their position. The full range is assessed.

As a final point, it is important to note that summing individual Likert items into an index does not alleviate the problems discussed previously. Although the resulting product of the summation process is often treated as if it possessed interval-level properties, in actuality the final index does not represent interval-level data. Individuals do not select a value from the index. Rather, individuals respond to single items. When these individual items are summed, the limitations associated with the individual items remain. Summation cannot eliminate these limitations or transform coarse ordinal-level data into interval-level data.

The summation process can also be understood in terms of information. The units that comprise interval-level scales contain more information than do ordinal-level units. In addition to rank ordering units, interval-level data contain information about the distances between units. By discarding some of the existing information, it is possible to collapse higher levels of information into lower levels. Interval data, for example, may be reduced to ordinal data. The rank ordering can be retained while the information concerning the distance between units is discarded.

It is impossible, however, to transform ordinal-level data into interval-level data. Information that was not originally collected cannot be added at a later date. Adding together a number of units with no information about the distance between units provides no guideline concerning how the distance between any two units is to be understood. There is nothing intrinsic in the summation process that imparts information about the distance between units. Although summed Likert data may be treated as interval-level data, the resulting index retains the limitations associated with coarse ordinal data.

Since the summation process neither adds nor subtracts information, the battle to achieve a high level of information is fought and won or lost at the item level. Each question should be designed to gather as much information as possible. Consequently, phrase completions were designed to tap information in a manner that approximates interval-level data. In turn, when phrase completion items are summed, the resulting index also approximates interval data.

Limitations and Strengths of the Phrase Completion Method

The phrase completion approach to attitude scaling is characterized by at least two disadvantages: the difficulty of operationalizing the underlying theoretical continuum and cultural limitations related to scaling.

Operationalizing the Continuum

Perhaps the most prominent disadvantage of the phrase completion method is the difficulty of operationalizing the underlying theoretical continuum. As the examples in Table I suggest, in many situations the operationalization of this continuum is a relatively straightforward process. In these cases, the original Likert items were easily transformed into similar phrase completion items. However, in some cases, it may be difficult to identify verbal anchors that operationalize the ends of the continuum.

Ideally, the ends of the scale should be anchored with simple, behaviorally based labels that are readily understood among the sampling frame of interest to the researcher. Developing such anchors may be a time-consuming process or even impossible in some cases. As noted previously, in some situations it may be helpful to employ a third verbal anchor in the center of the response key to highlight the fact that middle values represent moderate response options.

In contrast, it is important to note that creating Likert scales is a relatively simple process. Although the phrase completion approach of specifying the underlying continuum offers a number of advantages, any additional time spent constructing phrase competitions functions as a disadvantage. Furthermore, as previously implied, in some cases it may be essentially impossible to operationalize the underlying continuum in a scientifically useful manner.

Cultural Limitations

Cultural limitations may exist regarding the use of graded response keys. Specifically, some cultural groups may have difficulty expressing their views on a continuum. For example, certain populations raised outside the Western ideological context may be more inclined to use affirmative or negative, yes or no expressions. Consequently, methods such as Likert or phrase completions scales that employ responses on a continuum may lack validity with such groups. Furthermore, because the phrase completion approach is predicated on an 11-point response key, among populations for whom this limitation is a salient factor, the phrase completion method is likely to be especially disadvantageous.

Review of the Strengths of the Phrase Completion Method

The primary strength of the phrase completion method is that it represents an attempt to better meet three important assumptions. Unidimensionality and univocal stimuli are fundamental assumptions of measurement. The phrase completion method represents a unidimensional measurement procedure. The multidimensionality associated with superimposition, the confounding of direction and intensity, and the use of a midpoint are eliminated with the phrase completion approach.

Similarly, phrase completion items represent relatively clear, univocal stimuli. The cognitive complexity associated with multidimensionality and negative wording is eliminated. Specifically, the phrase completion approach avoids the use of reverse coding, items reversals, and mirror image items. By operationalizing the underlying theoretical continuum, the phrase completion approach yields relatively clear, concise, unidimensional items that are designed to yield substantively meaningful results.

With widely used parametric statistics, a central assumption is the use of interval- or ratio-level data with a normal distribution. The phrase completion method approximates interval-level data by using equally spaced integers to guide respondents' thinking along a continuum that reflects the full range of the substantively meaningful response options. Similarly, the use of an 11-point response key composed of equally spaced integers is congruent with the concept of a normal distribution.

Phrase completions were designed as an alternative to the Likert method. The limitations of the phrase completion method preclude their use in some situations, and the advantages of the Likert approach ensures the continued use of Likert scales. However, in a number of situations the phrase completion method provides a superior approach to measurement.

See Also the Following Articles

Likert Scale Analysis • Multidimensional Scaling (MDS) • Surveys

Further Reading

Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational Psychol. Measurement* **60**(3), 361–370.

Brody, C. J., and Dietz, J. (1997). On the dimensionality of two-question format Likert attitude scales. *Social Sci. Res.* **26**(2), 197–204.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Appl. Psychol. Measurement* **18**(3), 205–215.

Garg, R. K. (1996). The influence of positive and negative wording and issue involvement on responses to Likert scales in marketing research. *J. Market Res. Soc.* **38**(3), 235–256.

Krosnick, J. A., and Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In *Survey Measurement and Process Quality* (L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, eds.), pp. 141–164. Wiley, New York.

Roberts, J. S., Laughlin, J. E., and Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational Psychol. Measurement* **59**(2), 211–233.

Russell, C. J., and Bobko, P. (1992). Moderated regression analysis and Likert scales: Too coarse for comfort. *J. Appl. Psychol.* **77**(3), 336–342.

Pilot Study

J. Rodney Turner
Groupe ESC Lille, Lille, France

Glossary

business risk Two-sided risk due to the uncertainty of estimates.
insurable risk One-sided risk due to an unplanned event (which may or may not be foreseeable).
pilot study An element of a project or program used to gather data to reduce risk or uncertainty on the project or program, especially in the definition of the product to be produced or the method of producing that product; often used to prove technical or commercial feasibility.
program A temporary organization used as an agency for managing a group of related projects to achieve a higher order strategic objective not delivered by any of the projects on their own.
project A temporary organization used as an agency to which resources are assigned to undertake a unique novel and transient endeavor, and to manage the inherent risk, to deliver beneficial objectives of change.
prototype A sample product produced by a research project to gather data about the product to be produced by a larger project or program or about the method of producing it; often prepared as a precursor to a pilot project.
risk A threat or opportunity that can adversely or favorably affect the achievement of the objectives of an investment.

A pilot study is an element of work of a larger project or program undertaken to gather data to reduce risk or uncertainty. A pilot study can be undertaken to help in the selection of the appropriate risk mitigation strategy or in the application of the chosen method. The uncertainty in a project or program will usually lie in the definition of the product to be produced or in the method of producing that product. A pilot study can be used to gather data about either or both to facilitate project choices, particularly in proving the technical or commercial feasibility of options being considered. In the process, the pilot study may also contribute to organizational learning. In the context of social measurement, a pilot study may be undertaken for all of these purposes. It is assumed that a project or program is being undertaken to gather data to make a measurement. The pilot study may be used to test the feasibility of the data-gathering method and whether it will deliver the required data (testing the process) or whether the data gathered is a true measure of the item under investigation (testing the product). In the process, it will contribute to an understanding of the measurement methodology (organizational learning). There may also be risks, such as the measurement methodology will not work or there will be an adverse reaction from the population under investigation. A pilot study can help mitigate these risk. This article describes how pilot studies can be used to aid risk mitigation, to assist in the process of project definition, and to facilitate organizational learning. It also describes how to prepare a brief for a pilot study to aid its successful definition.

Introduction

A pilot study is an element of work undertaken as part of a larger project or program of projects to reduce risk or uncertainty. The pilot study may be a subproject of a program or an area of work or work package of a larger project. By generating additional data or information, the pilot study will help to reduce uncertainty about the product to be delivered by the project or program or about the method by which the product will be delivered. The information created from the data can also contribute to organizational learning.

Pilot studies are used extensively throughout research, technology, engineering, and business (despite the dearth of literature on the subject). In the context of the "Encyclopedia of Social Measurement," we are mainly

interested in projects for research and measurement. However, I discuss generically pilot studies and then try to relate the topics under discussion to research and measurement projects and programs.

I focus on the role of pilot studies in risk management on projects and how to structure a pilot study as a project for its successful completion. I do not engage in the debate about the difference between data, information, knowledge, or wisdom. I also do not engage in the debate about the difference between risk, uncertainty, or opportunity management. Suffice it to say that pilot studies can be equally useful in helping organizations to reduce risk and uncertainty associated with projects or to be better positioned to exploit opportunity. I also do not describe how to select and structure the data-gathering methodology used in a pilot study or how to analyze the data gathered.

Definitions

A pilot study is part of a larger project or program undertaken to improve understanding of the product being delivered by the project or program or the method of delivering that product. A project is an agency for undertaking a novel piece of work within an organization, created as a temporary organization to which resources can be assigned to do that work and to manage the inherent risk associated with it. A program is a collection of several projects with a common strategic objective. A pilot study can be a package of work in a larger project or, if sufficiently large and complex, a project in its own right, part of a larger program. Projects and programs are essentially unique, novel, and transient:

- Because a project (or program) is doing novel work, an organization will never have done a project exactly like the current one (although I do accept that some projects are more familiar than others).
- Being unique, they require novel processes for their delivery, involving novel teams of people, the human resources assigned to the agency.
- They are transient, being disbanded when the objective has been achieved.

Because they are unique and novel, projects and programs entail uncertainty. There is uncertainty about the product the project will deliver and whether it will achieve the objectives set for it, and there is uncertainty about whether the process adopted will deliver the desired objectives. The management of risk is therefore an essential part of the management of projects. Strategies for mitigating risks on projects include

1. Reducing the uncertainty associated with the definition of the product to be produced or with the method by which it is to be produced.
2. Avoiding the risk by finding a different way of doing the project.
3. Abandoning the project.
4. Reducing the likelihood of the risk occurring or the impact on the project if it does occur.
5. Transferring the risk to other parties, such as contractors or insurance companies.
6. Accepting the risk.
7. Creating a contingency plan to deal with the risk if it does occurs.

Pilot studies provide data to help in the selection of the appropriate risk mitigation strategy and so that the strategy selected can be better implemented. They also produce data to reduce uncertainty in the definition of the product to be produced by the project, or the process by which it will be produced, and therefore facilitate choices about the design of the project's product or process, further reducing risk and uncertainty. In providing this information, they also contribute to the learning of the organization.

Pilot Studies for Reducing Risk

Pilot studies can be used to help choose the appropriate risk mitigation strategy or to implement the chosen strategy.

Reducing Uncertainty of Estimates

There are two types of risk on projects: business risk and insurable risk. Business risk is two-sided risk, mainly due to the uncertainty of estimates. The estimate of the outturn value of an activity is some midrange value, usually assumed to be the most likely out-turn. The activity can turn out better or worse. A project made up of many such activities will also have a spread of possible outturns.

Insurable risk is one-sided risk, due to the occurrence of an unplanned event, which will affect the project unfavorably (or favorably). As a result, the project will turn out worse (or better) than expected.

Uncertainty of estimates leads to business risk, which can impact a project in two ways. First, the spread of possible outcomes for each activity causes a spread of possible outcomes for the project. It is suggested (based on a heuristic application of probability theory) that if ε is the average error in the estimate of each activity, E is the error in the estimate of the total project, and N is the number of activities that make up the project, then

$$\frac{\varepsilon}{E} = \sqrt{N}. \tag{1}$$

One way of reducing the error in the project estimate is to break the plan into smaller activities. However, this requires greater effort. Equation (1) implies that to

reduce the error in the project estimate by half requires four times as much planning effort as has been expended so far. This relationship has been observed empirically. Furthermore, it may not be possible to subdivide some activities, or there may be no estimating data available. Thus, it may be preferable to conduct a pilot study to reduce the uncertainly of estimates either because the pilot study is cheaper than the planning effort or because there is no other source of estimating data. In the context of social measurement, the pilot study may help determine how long the research project will take, the required sample size, or the likely response rates.

Second, if the estimated outturn of each activity has a uniform distribution (the mean, median, and mode are the same), the predicted outturn of the project will also have a uniform distribution. However, in reality the expected outcome of most activities is skewed to the worse. It does not require much of a skew in the estimates before the predicted outcome for the project becomes heavily skewed. It is often the case that the raw estimate of the project (obtained by summing the most likely outturn of each activity) has a very low chance of being achieved, and the expected outcome for the project (obtained by summing the expected outcome for each activity) is somewhat larger. The difference is contingency. Sometimes, there may be no information about how skewed an activity is, and so it may be worthwhile to conduct a pilot study, undertaking trials of specific activities, with one more of the following objectives: to determine how skewed an activity is, to find ways of reducing the extent of the skew, and to determine whether the activity should be avoided if possible (i.e., the risk should be avoided). In the context of social measurement, the measurement of the item under investigation may be skewed by an unknown amount, and a pilot study can help identify that in the selection of an appropriate measurement methodology.

Avoiding Risk

Sometimes, a risk is so severe that it has to be avoided. The potential impact on the project means that the project is very likely to fail. In order to know which risks should be avoided an organization needs to know its risk tolerance. An insurable risk should be avoided if it has both a high likelihood of occurring and a high impact if it does occur, and a business risk should be avoided if the expected outturn is very much greater than the most likely outturn. Sometimes, some of these parameters will not be known, and a pilot study is one way of determining them. Therefore, a pilot study may be conducted to determine the likelihood of an insurable risk occurring, the impact if it does occur, and the spread of possible outturns of a business risk.

In the context of social measurement, insurable risk might be that the required data cannot be gathered or an adverse reaction from the population under investigation. A pilot study can test these in advance so an appropriate response can be determined.

Abandoning the Project

One way of avoiding risk is to abandon the project, and this may often be done as a result of a pilot study; indeed, it may be the most common result. This may be very common in the context of social measurement.

Reducing the Likelihood or Impact of Risk

Similarly, if the likelihood or impact of an insurable risk is known, then a pilot study can be conducted to find ways of reducing the likelihood of it occurring or reducing the impact if it does occur. This is similar to conducting a pilot study to find ways of reducing the tail in the estimates of a business risk.

Transferring Risk

Determining the likelihood or impact of a risk is necessary if the risk mitigation strategy is to transfer the risk to other parties. Risks can be transferred in two ways. First, risks with a high chance of occurring, but with a small impact if they do occur, are often transferred to a contractor to manage. The contractor will accept the risk, allow a contingency for it, and add a profit margin for managing it. In order to do this commercially, the contractor needs to know the likelihood and impact of occurrence. If these are unknown, a pilot study can be used to help determine them. In the context of social measurement, this would entail asking a research organization to gather data on behalf of the client and carry the associated risk with the sample size, for instance. The research organization may want to conduct a pilot study to understand the risk before taking it on.

Second, risks with a very low chance of occurring, but a very large impact if they do occur, are best insured. The insurance company accepts the risk on payment of a premium and spreads its exposure over a large number of such risks. Again, to determine the premium and whether or not it wants to accept the risk, the insurance company needs to know the likelihood and impact of occurrence. A pilot study may help reduce the premium or even make an insurance company willing to accept a risk in the first place in the case in which a risk is not well understood. It might be necessary for a researcher to obtain professional indemnity insurance. If the insurance

company has no data on which to base the premium, a pilot study can help provide that data.

Accepting Risk

Sometimes, the best strategy is just to accept the risk. This is usually the case for risks with a small to medium chance of occurring and a small to medium impact if they do occur. Again, an organization needs to know its risk tolerance and what it means by small to medium in both instances. For risks that are not well understood, a pilot study can help determine whether or not a risk can be accepted.

Creating a Contingency Plan

For risks with a medium to high chance of occurrence and a medium to high impact if they do occur, allowing a contingency may be the best way of dealing with them. A contingency can be of three forms:

1. Pure contingency is a plan for how to respond if the risk occurs, with no prior action to reduce the likelihood or impact of the risk. An example with regard to a research project is to gather more data if necessary.
2. Contingency with essential prior action is a plan for how to respond with action taken to reduce the impact of the risk while the response is implemented. The essential prior action usually makes the response faster. On a research project there may be a cost and delay associated with copying the research instrument. If there is a high chance that additional data will be required, and a large cost associated with delay, it may be worthwhile making the additional copies in advance.
3. Contingency with mitigating prior action is a plan to respond with action taken to reduce the likelihood of the risk occurring. In a research project, this would entail spending money to try to increase the response rate.

With essential and mitigating prior action, the likelihood and impact of the risk without the action are compared to the likelihood and impact with the action and the cost of the action. If L is the likelihood of the risk and C the cost of its occurrence without action, l is the likelihood and c the cost with the action, and K is the cost of the action, then the action is taken if

$$L \times C > l \times c + K.$$

Pilot studies can be conducted to determine the efficacy of a contingency plan, or of the proposed courses of action, or to determine the likelihood and/or impact of a risk with and without the proposed action.

Pilot Studies for Proving Product or Process

The uncertainty in a project often lies in the definition of the product to be produced or in the method or process by which it will be produced. This is particularly the case with research projects in which the research instrument may be unproven or it is not known what results it will produce. Figure 1 illustrates a typology for projects, whereby projects are judged by the uncertainty of the product and process. In type 2, 3, and 4 projects, there is uncertainty about the product to be delivered, the method of delivering that product, or both, respectively. The risk lies in the lack of definition, the fact that the end product of the project may not work, or the fact that the method of producing that product may not be feasible; thus, the organization undertakes a pilot study to reduce uncertainty before committing to full implementation. In undertaking the pilot study, the organization may produce a prototype of the final product to prove the design of the product or the method by which it will be produced. The risk mitigation strategy adopted may be any one of those described previously. Some examples are described next.

Commercial Bank

During a research project on project categorization, a major high street (commercial) bank in the United Kingdom told me that it always conducts a pilot study for any new products or change in organization being introduced into its branch network. Innovations are introduced in a three-step process:

Step 1: A pilot in six branches
Step 2: A trial in 50 branches
Step 3: Roll out to the remaining branches

Methods well defined	Goals well defined: Yes	Goals well defined: No
No	**Type 2** Development Milestone planning using known deliverables	**Type 4** Research Soft systems planning to define deliverables
Yes	**Type 1** Engineering Activity-based planning	**Type 3** Computer systems Milestone planning using life-cycle stages

Figure 1 Goals and methods matrix.

The products of these projects may be a new product to be sold in the branches, a new method of working in the branches, a new computer system, or a new way of interfacing with customers. The uncertainty here mainly lies with the products; they are type 3 projects according to the classification in Fig. 1. The pilot study is undertaken to prove the product and make changes before trial and roll out. In this case, the change project is being undertaken in a social context, and the pilot study measures the reaction of that social context to the change.

Testing a Research Project

Huemann and Winkler conducted a pilot project as part of a larger research project on benchmarking project processes. Since they adopted a social constructivist approach to their research, they were making measurements in a social context. They used the pilot to test their instrument before roll out to a larger sample. In this case, the product of the project, the desired outcome of the research, is well defined. The uncertainty lies in the method of achieving that product, the design of the questionnaire or research instrument. It is a type 2 project. The pilot study is used to reduce the uncertainty of the project process before roll out. Van Teilingen and Hundley confirm that this is good practice. Table I provides reasons for conducting such a pilot study to test a research instrument. Many of these are familiar. Table II lists procedures for improving the validity of the instrument. However, they caution that to avoid contamination, data gathered in the pilot study should not be included in the main results, and people surveyed in the pilot study should not be resurveyed because they will not give independent answers.

Table I Reasons for Conducting a Pilot of a Research Instrument[a]

Developing and testing adequacy.
Assessing feasibility.
Designing a protocol.
Establishing and testing efficacy of the sampling technique.
Assessing efficacy of recruitment processes.
Identifying logistical problems.
Estimating variability to determine optimum sample size.
Planning and estimating resources need for the main study.
Assessing the data analysis techniques.
Training researchers in research techniques.
Convincing funding bodies that the research team is competent.
Convincing funding bodies that the study is feasible and worth funding.
Convincing other stakeholders to give support.

[a] Adapted from Van Teilingen and Hundley (2003).

Electrical Goods Manufacturer

An electrical goods manufacturer planned to develop a new model of coffee maker. It wanted to reduce the price of its coffee makers to gain a wider market share. To do so, it planned to make the new model out of a cheaper plastic, polypropylene rather than polycarbonate. There was a problem in that polypropylene is more difficult to cast than polycarbonate and can suffer shrinkage, causing rippling, especially in a casting as large as the water tank. To overcome this problem, the company planned to design the water tank with a ripple effect to disguise the ripples caused during casting. There was uncertainty about the feasibility of the design and the casting process. This was a type 4 project during the initial research and feasibility stage of the project. A pilot study was conducted, and prototypes were produced to test various options. As a result, when released to market the coffee maker was very successful. (This is an example of a technical use of pilot studies. The social element of this project was the market research to test the response to the proposed new design, conducted as part of the main pilot study.)

Prototype

This last case describes the production of a prototype, which can be the result of many pilot studies. There is very little literature on pilot studies, and in the 2.5 million hits in Google and Yahoo there is little agreement about what is meant by a pilot study. However, Field and Keller state that there is a difference between a prototype and a pilot study. They are mainly discussing information systems (IS) projects, but they state that a prototype precedes a pilot study. A prototype is produced in the laboratory as part of the research, feasibility, or design stage of a project, and on an IS project will operate on trial data. A pilot study is conducted during the implementation stage of a project and is a limited implementation

Table II Improving the Internal Validity of a Research Instrument in a Pilot Study[a]

Administer the pilot in the same way as the main study.
Ask subjects for feedback to identify flaws.
Record time taken to compete the instrument and decide if it is reasonable.
Discard unnecessary, difficult, or ambiguous questions.
Assess whether questions give an adequate range of responses.
Establish whether responses can be properly interpreted to give information required.
Check that all questions are answered.
Reword and resale questions as necessary.
Shorten and revise the instrument.

[a] Adapted from Van Teilingen and Hundley (2003).

using real data in a live operating environment to generate information to reduce risk during roll out. Hayes *et al*. differentiate between prototype and pilot production run in the same way, but van Teilingen and Hundley do not.

Pilot Studies for Learning in Organizations

Another purpose or side effect of pilot studies often emphasized is the learning they provide for organizations. All the purposes discussed previously have learning as a side effect:

- Learning how to mitigate risk
- Learning how to reduce uncertainty in product or process of a project
- Learning what will work or not work in the design of a new product
- Learning by testing the efficacy of a research instrument

Van Teilingen and Hundley describe the training of research staff as a possible reason for conducting a pilot study of a research instrument. However, sometimes pilot projects are more overtly established with the express purpose of generating data to aid learning or knowledge management in organizations. Ayas describes a pilot study to demonstrate the effectiveness of the learning instruments to be adopted in a development program in an aircraft manufacturing company. Project managers often manage on short time horizons and need to see quick results. The pilot study demonstrated the efficacy of the learning instruments adopted and showed that they were likely to deliver the desired results. Ayas was describing the measurement of learning in a social context.

Preparing a Brief for a Pilot Study

A pilot study is part of a larger project or a project in a larger program. Therefore, good project management practices should be applied to its management. In particular, all pilot studies should be initiated by a project brief. Here, I describe the contents of a project brief. The brief may be 12 pages for a reasonably complex pilot project, but it may be as short as 1 or 2 pages for a study conducted as a package of work for a larger project. The brief defines the pilot study and contains the following sections:

Background: The context of the pilot study is defined; what is being done; and the difficulty encountered that has created the risk that needs to be mitigated.

Purpose: The reason for undertaking the study is defined; why it is being done. The risk to be mitigated is stated more specifically as a problem to be solved.

Objectives: The outcomes of the study are stated; the desired results or deliverables that will enable the purpose to be achieved and that will be used to solve the problem.

Scope: The work to be done to achieve the objectives is described. This gives an overview of how the data are to be gathered and analyzed. Any constraints also need to be stated, along with interfaces with other elements of work in the project or program.

Plan: This shows how the work is to be done and how long it is expected to take. For a more complex project, this may be a milestone plan as described by Turner, and on a larger pilot project each milestone may be further broken down into a list of activities. On a smaller project or a package of work the plan may just be a list of activities.

Responsibility chart: The roles and responsibilities of people involved in the study are defined. The use of responsibility charts is described by Turner.

Stakeholders: This identifies the parties who have an interest in or are affected by the study and their possible (adverse) responses. Plans to deal with any adverse responses are briefly stated.

Quality requirements: Quality standards to be met by the pilot study are defined.

Cost and schedule: Rough estimates of the cost of the study and its duration are given.

Acceptance criteria: Standards to be met by the pilot study to enable full roll out are defined.

Known risks: The pilot study is undertaken to mitigate risk but may itself entail risk. Any known risks of the pilot study are identified, and mitigation strategies for them are described.

Epilogue

I was surprised when I was asked to write a chapter on pilot study. Although I am well-known in the project management research community, and pilot studies are a common element of projects, I assumed there were many people with more experience in writing about pilot studies than myself.

However, when I did research for this article, I could not find any reference to pilot studies in any of the books in my library. There was no entry in the index under "pilot study" in books that I examined from the general management literature, project management literature (even in my own books), operations management literature, change management literature, innovation management literature, or strategic management literature. I did a search in Google and Yahoo. In both cases, I got approximately 2.5 million hits. However, in both cases the first 60 or so primarily concerned people reporting the outcomes of their pilot studies; only in one instance was somebody giving advice on how to use them.

In the literature, I found three references on conducting pilots. There was an index entry for "pilot project" in Lundin and Midler to a paper by Ayas, but like the hits in Google and Yahoo, Ayas was reporting the results of her own pilot project. There is also an entry for "pilot production run" in Hayes *et al*., but they only suggest that it is something that might be done. They give no guidance on how to do it. The only index reference to "pilot study" I found was in Field and Keller, who devote one line to the difference between pilot studies and prototypes.

Pilot studies are very important in the undertaking of research projects to ensure that the correct methodology is adopted and implemented properly to achieve the required results, yet they appear to be written about nowhere. Pilot studies are like the person who was not there upon the stair or like Bishop Berkley's tree in the quad: is it still there when there is nobody in the quad to observe it?

There once was a man who said, "God
Must think it exceedingly odd
If he finds that this tree
Continues to be
When there's nobody about in the quad."

Robert Knox (1924)

So, somewhat to my surprise, this article seems to be almost the first attempt to set out the purposes and functions of pilot studies and give guidance on how to manage them. Van Teilingen and Hundley suggest that the reason pilot studies are not written about is because data produced by them have no academic validity. Academic research papers need to be based on testable, verifiable data, which are obtained from a full survey, not from the pilot study. Thus, they are very important but not written about in research journals. Therefore, like Bishop Berkley's tree, they are there even though they are not often written about.

Dear Sir, Your astonishment's odd
I am always about in the quad
And that's why the tree
Will continue to be
Since observed by yours faithfully, God

Anonymous

See Also the Following Articles

Organizational Behavior • Risk and Needs Assessments

Further Reading

Ayas, K. (1998). Learning through projects: Meeting the implementation challenge. In *Projects as Arenas for Renewal and Learning Processes* (R. A. Lundin and C. Middler, eds.). Kluwer, Boston.

Crawford, L., Hobbs, J. B., and Turner, J. R. (2002). Investigation of potential classification systems of projects. In *Proceedings of the PMI Research Conference 2002, Seattle, June* (D. I. Cleland, D. P. Slevin, and J. K. Pinto, eds.). Project Management Institute, Philadelphia.

Field, M., and Keller, L. (1998). *Project Management*. International Thompson Business Press, London.

Hayes, R. H., Wheelwright, S. C., and Clark, K. B. (1988). *Dynamic Manufacturing: Creating the Learning Organization*. Free Press, New York.

Huemann, M., and Winkler, G. (1996). Project management benchmarking: An instrument for learning. In *Projects as Arenas for Renewal and Learning Processes* (R. A. Lundin and C. Middler, eds.). Kluwer, Boston.

Lundin, R. A., and Middler, C. (eds.) (1998). *Projects as Arenas for Renewal and Learning Processes*. Kluwer, Boston.

Turner, J. R. (1999). *The Handbook of Project Based Management*, 2nd Ed. McGraw-Hill, London.

Van Teilingen, E. R., and Hundley, V. (2003). The importance of pilot studies. *Social Sci. Res. Update* **35,** (Available at www.soc.surrey.ac.uk/sru35.html)

Playfair, William

Ian Spence

University of Toronto, Toronto, Ontario, Canada

Howard Wainer

National Board of Medical Examiners, Philadelphia, Pennsylvania, USA

Glossary

Biderman, Albert (1922–) American sociologist trained at the University of Chicago. He was instrumental in founding the National Science Foundation (NSF) working group on social graphics, and his NSF project on graphics can be credited with instigating the work of Edward Tufte and Howard Wainer in statistical graphics.

Euler, Leonhard (1707–1783) Swiss mathematician who trained under Jean Bernoulli. He published over 800 books and papers on every aspect of pure and applied mathematics, physics, and astronomy. In 1738, when he was professor of mathematics at St. Petersburg Academy, he lost sight in one eye. In 1741 he moved to Berlin, but returned to St. Petersburg in 1766, where he soon lost sight in the other eye; however, his prodigious memory allowed him to continue his work while totally blind. For the princess of Anhalt-Dessau he wrote *Lettres à Une Princesse D'Allemagne* (1768–1772) in which he gave a clear, nontechnical outline of the principal physical theories of the time. His *Introductio in Analysin Infinitorum* (1748) and later treatises on calculus and algebra remained the standard texts for more than a century.

Funkhouser, Howard Gray (1898–1984) American mathematician and educator who was born in Winchester, Virginia. He was a 1921 graduate of Washington and Lee and received his Ph.D. from Columbia. He taught mathematics at Washington and Lee from 1924 to 1930 and spent 1931 on the mathematics faculty at Columbia. In 1932, he accepted a position on the faculty at Phillips Exeter Academy, where he remained until his retirement in 1966. His groundbreaking paper "Historical development of the graphical representation of statistical data," published in *Osiris* in 1937, joined two other papers, "Playfair and his charts" (1935) and "A note on a tenth century graph" (1936), as the jumping off point for subsequent researchers in the history of graphics.

Louis XVI (1754–1793) King of France (1774–1793), born in Versailles, France, the third son of the dauphin Louis and Maria Joseph of Saxony, and grandson of Louis XV, whom he succeeded. He was married in 1970 to Marie Antoinette, daughter of the Hapsburg Empress Maria Theresa. He made a number of unfortunate decisions (e.g., failing to support reform of financial and social structures, involvement in the American Revolution), which exacerbated the national debt. In August of 1792 the monarchy was abolished; he was then tried by the revolutionary government, and in 1793 he and his queen were guillotined in Paris.

Meikle, Andrew (1719–1811) Millwright and inventor who was born in East Lothian, Scotland. He inherited his father's mill, and to improve production invented the fantail (1750), a machine for dressing grain (1768), and the spring sail (1772). His most important invention was a drum threshing machine (patented in 1788) that could be driven by wind, water, horse, or (some years later) steam power.

Minard, Charles Joseph (1781–1870) He was first a civil engineer and then an instructor at the École Nationale des Ponts et Chaussées (ENPC). He later was an Inspector General of the Council des Ponts et Chaussées, but his lasting fame derived from his development of thematic maps in which he overlaid statistical information on a geographic background. The originality, quality and quantity of this work led some to call him "the Playfair of France." His intellectual leadership led to the publication of a series of graphic reports by the Bureau de la Statistique Graphique of France's Ministry of Public Works. The series (*l'Album de Statistique Graphique*) continued annually from 1879 until 1899 and contained important data on commerce that the Ministry was responsible for gathering.

In 1846, he developed a graphical metaphor of a river, whose width was proportional to the amount of materials being depicted (e.g., freight, immigrants), flowing from one geographic region to another. He used this almost exclusively to portray the transport of goods by water or land. This metaphor was employed to perfection in his 1869 graphic, through which, by using the substitution of soldiers for merchandise, he was able to show the catastrophic loss of life in Napoleon's ill-fated Russian campaign. The rushing river of 422,000 men that crossed into Russia when compared with the returning trickle of 10,000 "seemed to defy the pen of the historian by its brutal eloquence." This now-famous display has been called "the best graph ever produced."

Playfair, John (1748–1819) Minister, geologist, and mathematician. Born in Dundee, Scotland, he was the older brother of William Playfair. He studied at St. Andrews and became professor of mathematics (1785) and natural philosophy (1805) at Edinburgh University. In addition to influential writings on geometry, including a widely used textbook, he also investigated glaciation and the formation of river valleys. He was responsible for clarifying and amplifying the revolutionary geological theories of James Hutton, which anticipated modern scientific ideas such as evolution, natural selection, plate tectonics, and asteroid strikes.

Playfair, William (1759–1823) Scottish engineer, writer on political and economic topics, and the father of modern graphical methods.

Priestley, Joseph (1733–1804) Chemist and clergyman who was born in Fieldhead, West Yorkshire, England. He became a Presbyterian minister in Suffolk in 1755 but returned to Leeds in 1767 where he continued his scientific and philosophical studies. He is best known for his research on the chemistry of gases and for his discovery of oxygen. Some of his most productive scientific years were spent in Birmingham, where he also wrote books on education and politics. His political activities and his support of the French Revolution were controversial, making his continued presence in England more than uncomfortable, and in 1794 he emigrated to America, where he was well received.

Rennie, John (1761–1821) Born in East Linton, Scotland, he apprenticed with Andrew Meikle and later studied at Edinburgh University. He worked briefly at Boulton & Watt, and in 1791 started his own engineering company in London. He built docks at Hull, Liverpool, Greenock, Leith, Portsmouth, Chatham, and Plymouth. He is best known for his bridges, and his successes include Leeds Bridge, Southwark Bridge, Waterloo Bridge, and London Bridge, which was dismantled in 1967 and re-assembled in Arizona as a tourist attraction.

Tufte, Edward R. (1942–) American political scientist and graphics expert. He was born in California and trained at Yale and Stanford. His seven books include *The Visual Display of Quantitative Information*, *Envisioning Information,* and *Visual Explanations*, which have received unprecedented attention garnering among them more than 40 awards for content and design. They have sold, in aggregate, more than a half million copies and their author is in constant demand as an influential critic of graphical design.

Von Humboldt, Alexander (1769–1859) Naturalist and geographer who was born in Berlin and educated in Frankfurt, Berlin, Göttingen, and Freiberg. He spent 1799 through 1804 exploring South America with Aimé Bonpland (1773–1858). When he was 58, he spent three years traveling throughout central Asia. The Pacific current of the coast of South America is named for him. His principal book, *Kosmos,* tries to provide a comprehensive characterization of the universe.

Watt, James (1736–1819) Inventor. He was born in Greenock, Scotland, and in 1754, he apprenticed as an instrument maker in Glasgow, where he stayed and set up a business. As part of his trade he carried out surveys for canals and began to study the use of steam as an energy source. In 1763, he repaired a model Newcomen steam engine and found he could improve its efficiency through the use of a separate steam condenser. In 1774, he joined with Mathew Boulton in an enterprise to manufacture an improved steam engine in Birmingham. He subsequently made numerous other inventions, including the double-acting engine, parallel motion linkage, centrifugal governor for automatic speed control, and the pressure gauge. He is credited with coining the term "horse-power." The standard unit of electrical power is named for him.

A ubiquitous practice in modern science is the atheoretical plotting of data points with the goal of looking for suggestive patterns. This practice was initiated by William Playfair, an 18th century Scot, who not only invented most of the graphical forms used today but also showed in numerous publications how they could profitably be used.

Introduction

Today, with illustration at the heart of communication, we see pie, bar, and line charts everywhere—in the press, on television, on computer screens, on boardroom desks, on blackboards, whiteboards, and greenboards, in video presentations, handouts, flyers, and so forth. Aircraft control panels and nuclear power station monitors contain displays that look like bar charts in motion, and video games often keep track of players' scores in graphical form. It is difficult to determine how many graphs are created each day, but more than a decade ago, one noted commentator estimated the number at more than 5 million. However large the true number is, you can be sure it will be larger tomorrow, and the multitude of users of all these charts, graphs, and displays will grow apace. Statisticians employ graphs but so do scientists of every stripe, and unnumbered professionals in

business and commerce make use of graphs every single day. Economists, sociologists, psychologists, social workers, medical professionals, and even historians are just a few of the occupations for which graphs are the stuff of everyday life.

Graphs convey comparative information in ways that no table or description ever could. Trends, differences, and associations are effortlessly seen in the literal blink of an eye. The eye discerns immediately what the mind would take seconds or minutes to deduce from a table of numbers. This, of course, is what makes graphs so appealing—they allow the numbers to speak clearly; they release them from their mute state. Graphs and charts not only show what numbers tell, they also help scientists—numerical detectives, if you will—tease out the critical clues from their data. Graphs transcend national boundaries—a Chinese can read the same graph that a Russian draws. There is no other form of communication that more appropriately deserves the description "universal language."

Who invented these ubiquitous and versatile objects? Have they been around for millennia, rather like the wheel or fire, the work of inventors unknown? The truth is that statistical graphs were not created in some distant past; their inventor lived only two centuries ago. He was a man of such unusual talents and background that had he not introduced his charts at the end of the 18th century, during the Age of Enlightenment, we might have waited until the 20th century for statistical graphs. The inventor was not a cloistered academic, although he was deeply knowledgeable on many subjects and wrote more prolifically than many in the ivory towers. He was a man of several careers and varied experience. He dearly wanted to be rich, but none of his many schemes realized this desire. He was something of a rogue, but oddly enough this rascally aspect may have helped bring his graphical inventions to the world.

Though born and raised in Scotland, he lived most of his life in London and Paris during turbulent times. William Playfair (1759–1823) was so convinced that he had found the best way to display economic data that he spent almost 40 years of his life trying to influence others to follow his example. He made notable converts, including the doomed Louis XVI of France, but he was unsuccessful in persuading the academic establishment and thus his inventions waited almost a century before widespread adoption.

William Playfair is the principal inventor of statistical graphs. Although one may point to isolated instances of rudimentary line graphs that precede Playfair's work,

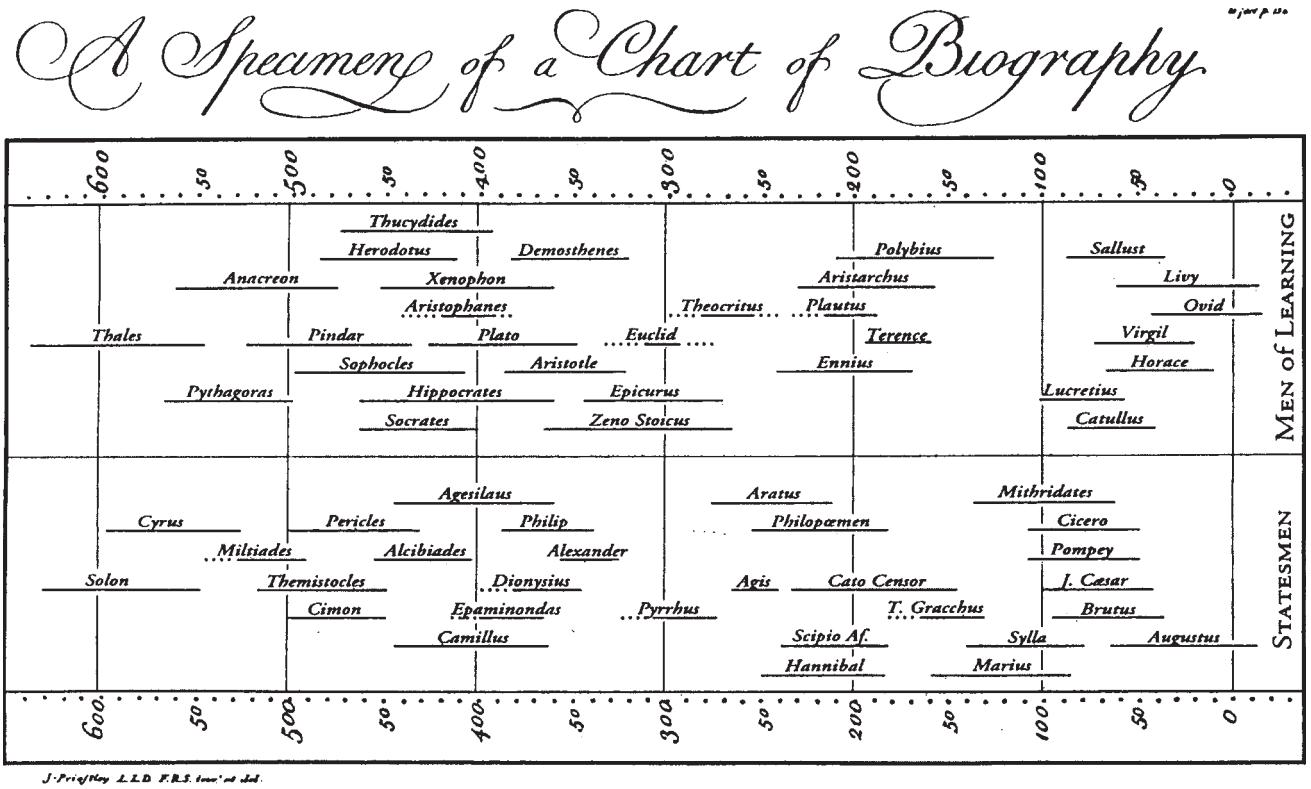


Figure 1 Life spans of 59 famous people in the six centuries before Christ. Its principal innovation is the use of the horizontal axis to depict time. It also uses dots to show the lack of precise information on the birth and/or death of the individual shown. From Priestley (1765).

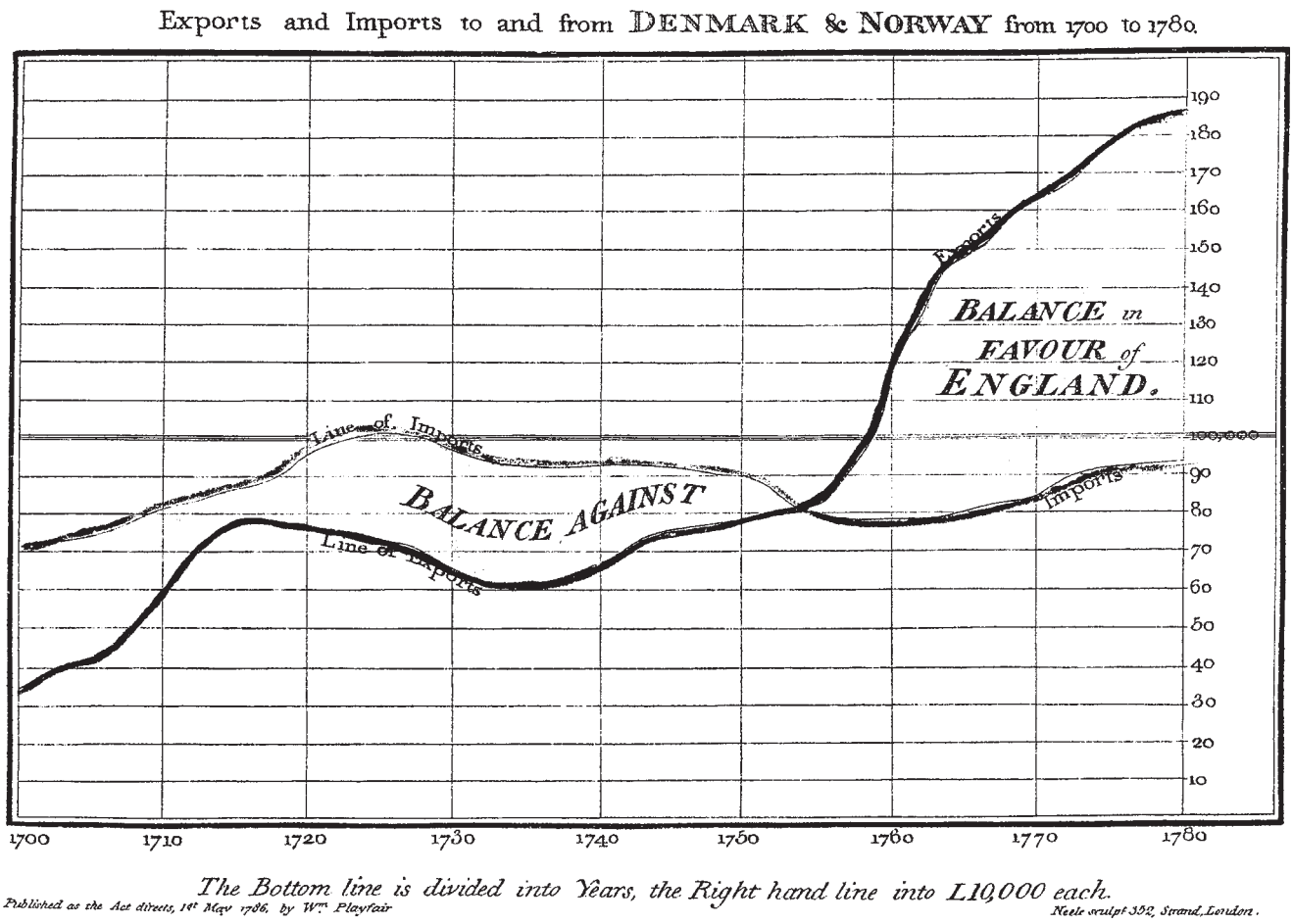


Figure 2 A typical line chart from Playfair's 1786 *Commercial and Political Atlas*, which uses one line to show imports and another for exports over the 80-year span from 1700 until 1780. The balance of trade between England and Denmark and Norway is the area between the two curves, which was shaded red (prior to 1765) when it was against England, and green thereafter when it was in England's favor.

such examples generally lack sophistication and, without exception, failed to influence others. In contrast, Playfair's graphs were elaborate and well constructed: they appeared regularly in several publications over a period of more than 30 years and they introduced a surprising variety of devices and techniques that are in use to this day. He invented three of the four basic forms; the statistical line graph, the bar chart, and the pie chart. The other important basic form—the scatter plot—did not appear until almost a century later. Playfair also invented other graphical elements that are still in use today, for example, the circle diagram and statistical Venn diagram, but these innovations were less effective and are less widely used.

William Playfair was born in 1759 in Scotland during the Enlightenment, on the cusp of the Industrial Revolution, to which he contributed as a young draftsman in the employ of Boulton & Watt. Upon his death in 1823—after a controversial and unconventional life—Playfair's obituaries were united in the conclusion that his talent had been squandered. But all published tributes missed the key achievement of his life, and a century would pass before the value of his work was fully recognized. As Funkhouser (1937) has noted, Playfair invented a universal language useful to science and commerce, and though his contemporaries had failed to grasp the significance, Playfair never doubted that he had changed the way we would look at data. Very few shared his enthusiasm for pictorial display, and it is a curiosity of history that one of those who did appreciate the inventions was the ill-fated King of France, Louis XVI. Playfair noted that, after receiving a copy of his pioneering volume *The Commercial and Political Atlas*, Louis XVI said, "[The charts] spoke all languages and were very clear and easily understood."

William Playfair was the fourth son of the Reverend James Playfair of the parish of Liff and Benvie near the city of Dundee, Scotland. His father died in 1772, leaving the eldest brother John to care for the family. John was subsequently to become one of Britain's foremost

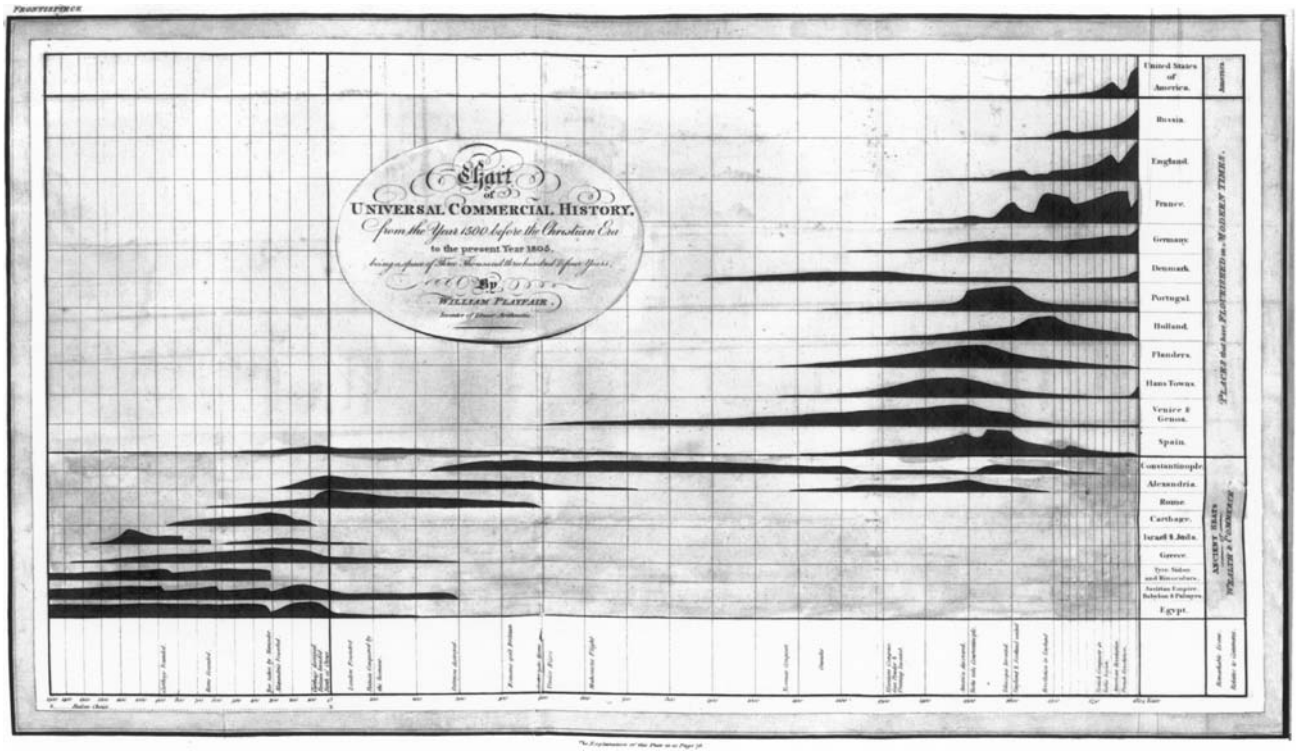


Figure 3 Playfair's only silhouette chart found as the frontispiece in *An Inquiry* (1805, 1807). This is a diagrammatic chart with no quantitative values represented. It depicts the rise and fall of 20 economies over a period of more than 3000 years. The original uses various colors in the background.

mathematicians and scientists as professor of natural philosophy, mathematics, and geology at Edinburgh University. After an apprenticeship with Andrew Meikle, the inventor of the threshing machine, William became draftsman and personal assistant to the great James Watt at the steam engine manufactory of Boulton & Watt at Birmingham in 1777. Thus, William's scientific and engineering training was at the hands of the leading figures of the Enlightenment and Industrial Revolution. On leaving Boulton & Watt in 1782, Playfair set up a silversmithing business and shop in London, but the venture failed. Seeking his fortune and hoping to apply his engineering skills to better effect in a developing French economy, Playfair moved to Paris in 1787.

It was about this time that Playfair developed most of his graphical formats and published several examples. He already had the mathematical training from his brother John, the engineering know-how from Meikle and Watt, knowledge of business and economics from men such as Boulton, and some hard practical experience of the world of affairs from his botched enterprises in London and Paris. But his charts were not readily accepted, especially in Britain, where concerns regarding accuracy were not eased by his occasional carelessness and his less than reputable personal standing. He was received more sympathetically in Germany and France, gaining approval from the geographer Alexander von Humboldt, among others. Nevertheless, there was still considerable opposition to his ideas and there was no general adoption of his methods until the second half of the 19th century, when Minard and Bertillon incorporated some of Playfair's devices in their cartographical work.

The Time-Series Line Graph

In 1786, shortly before he left for Paris, Playfair published his *Commercial and Political Atlas*, which contained 44 charts, but no maps; all of the charts, save one, were variants of the statistical time-series line graph. Playfair credits his brother for the inspiration that led to the line graph—John had made him keep daily records of temperature and chart these data in similar fashion. As Scotland's foremost mathematician of the day, John Playfair was certainly familiar with the use of Cartesian graphs to represent functions, and would also have been aware of the work of Lambert, who superimposed empirical data points on such functions. Another influence can be found, a decade beforehand, in the work of Joseph Priestley, who had conceived of representing time geometrically (see Fig. 1). The use of a grid with time on the horizontal axis was a revolutionary idea, and the

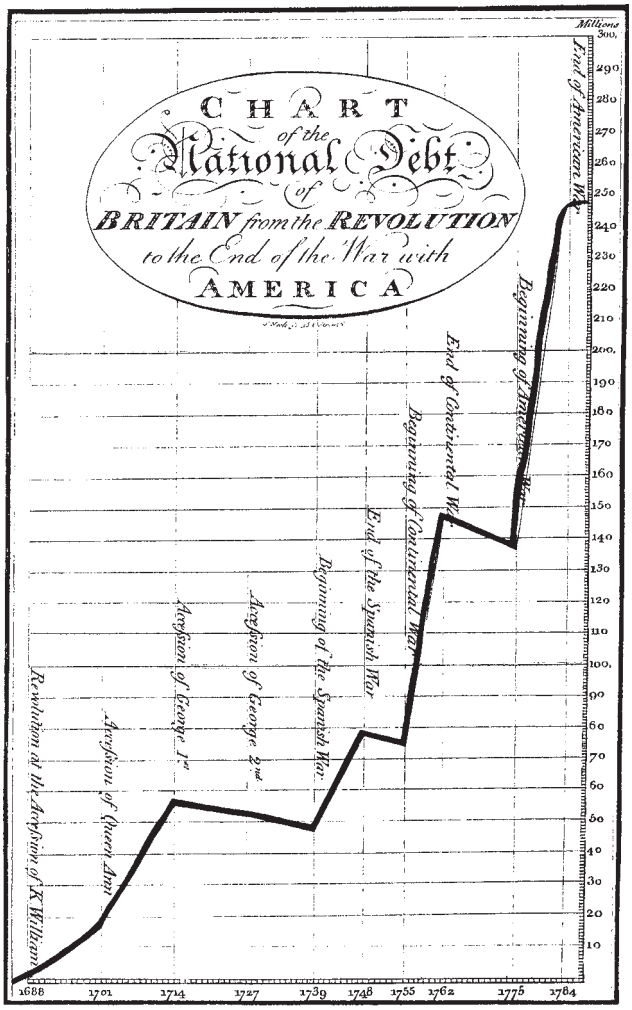


Figure 4 This remarkable "Chart of the National Debt of England" appeared as plate 20, opposite page 83 in the third edition of Playfair's *Commercial and Political Atlas* in 1801. Not only is it the first "skyrocketing government debt" chart, but it also uses the innovation of an irregularly spaced grid along the time axis to demark events of important economic consequence.

representation of the lengths of reigns of monarchs by bars of different lengths allowed immediate visual comparisons that would otherwise have required significant mental arithmetic. What is more, the relative differences in time periods and their relative position in the overall chronology could also be readily apprehended; no wonder that Priestley's device proved popular.

William Playfair was well acquainted with Priestley and his work, encountering the older man on a frequent basis in the Birmingham of the late 1770s. It was Playfair's genius to take the ideas implicit in Lambert and Priestley's charts and, with the stimulus of his brother's early instruction, to produce the first statistical time-series line graphs (see Figs. 2 and 3).

These charts introduced a large number of innovative features that remain part and parcel of the statistician's repertoire today: the use of an informative title; graduated and labeled axes; ruled gridlines, with greater line weight for major intervals; broken and solid lines, or different line colors, to distinguish time series of different kinds; hachure, solid fill, and color to indicate areas that represent accumulated amounts; the colors green and red to indicate positive and negative balances; appropriately placed labels to indicate historical events (Fig. 4); and so forth.

Playfair also introduced novelties that are still occasionally seen today: for example, in one chart, whose vertical dimension was insufficient to contain a particularly high peak in expenditures, Playfair extended the curve beyond the frame at the top and allowed it to repeat at the bottom of the graph. Although it is likely that this obvious peculiarity may have resulted from an error in planning, it turned out to be, in effect, an editorial comment. The reader is immediately drawn to the unusual nature of the sharp rise in cost—the implication is that the spike in prices is so egregious that the scope of the chart is unable to accommodate the excursion.

The Bar Chart

Playfair freely acknowledged Priestley's chronological diagrams as the source of the single bar chart that appeared in his atlas (Fig. 5). But he introduced this chart with apologies. He had insufficient data to be able to present the time-series chart that he had intended and so of necessity had to invent another form in which the horizontal axis did not represent the flow of time. He thought so little of this invention that he made no subsequent use of it, at least in its original form. He did, however, use bars in later graphs, but to display changing data over time.

The Pie Chart

Whereas both the line graph and bar chart used linear extent to represent quantity, Playfair's next inventions used area. The *Statistical Breviary* contained charts that were intended to show the areas, populations, and revenues of European states (Fig. 6). The charts also indicated which countries were maritime powers by coloring them green, while the others were stained red. The first chart of the *Breviary* shows the countries before the French Revolution of 1789, and the second chart displays the changes thereafter in 1801, the year of the Luneville peace.

Thus in two charts in a single volume, Playfair introduced three new forms of statistical graph: the circle diagram, the pie chart, and a Venn-like diagram, which is

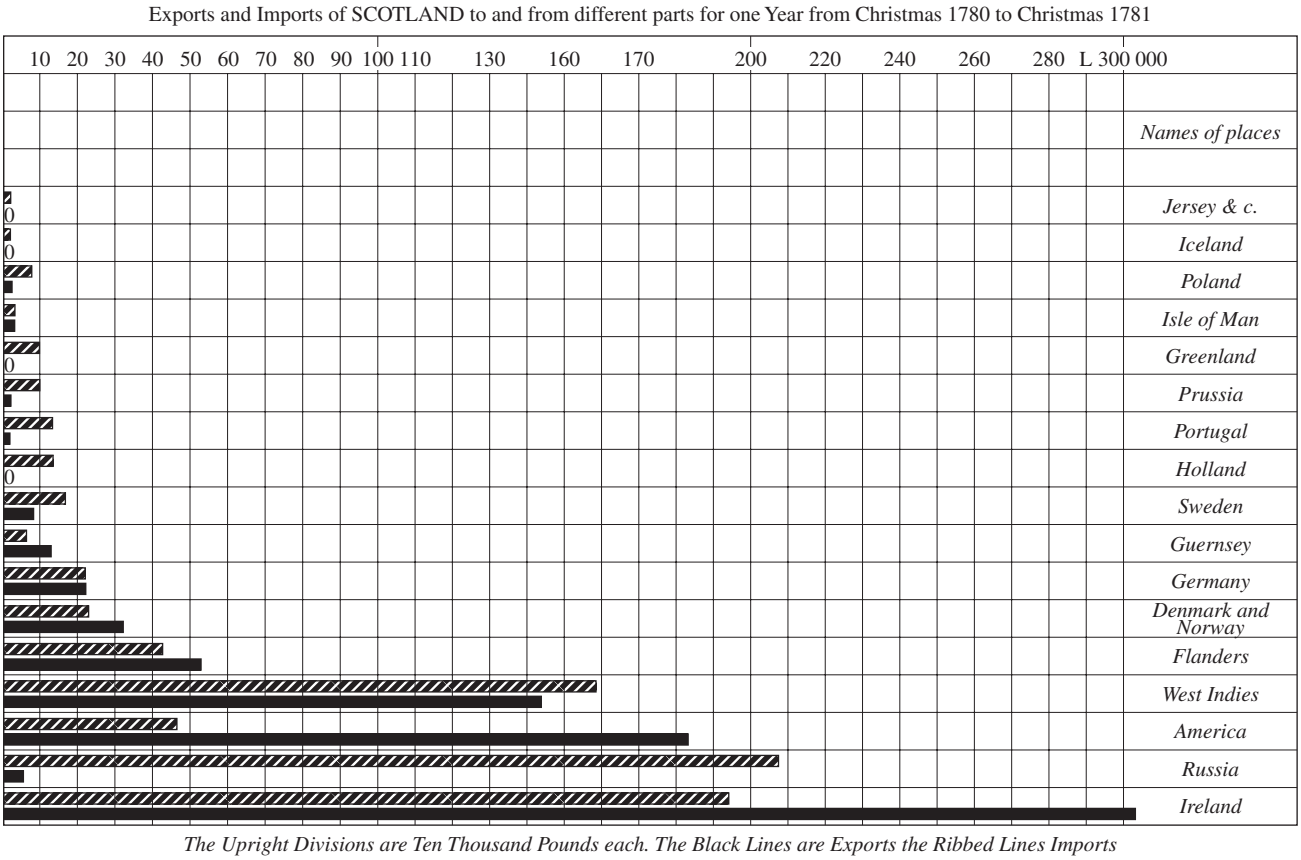


Figure 5 Imports from and exports to Scotland for 17 different places. After Playfair (1786), plate 23.

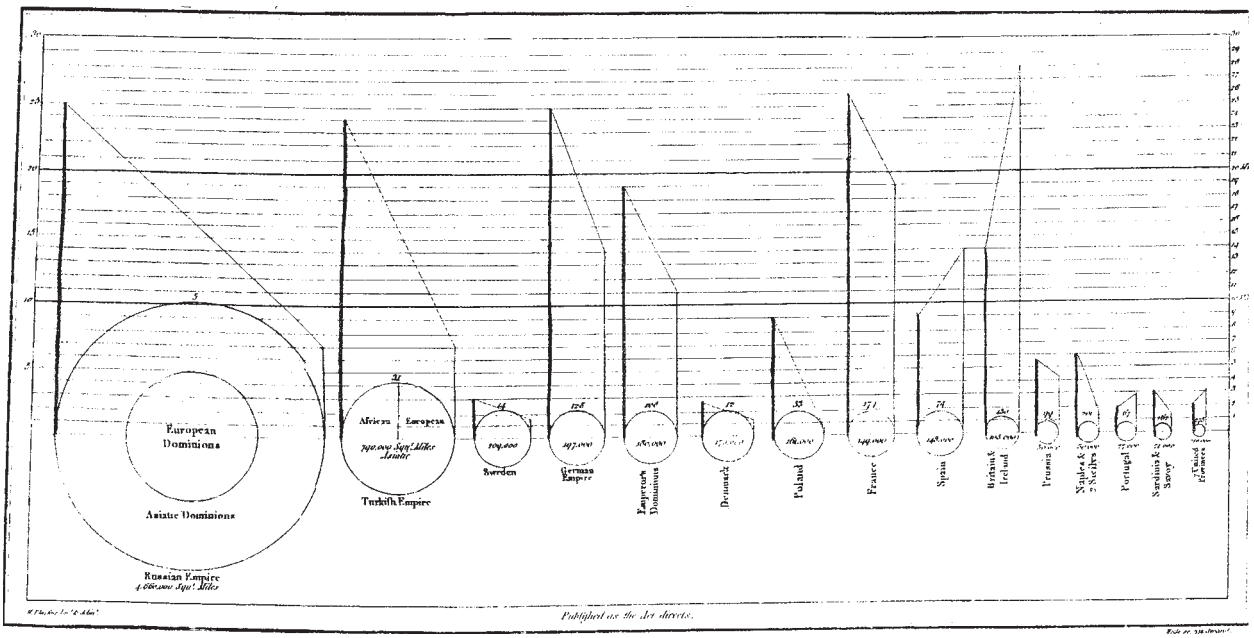


Figure 6 An innovative design from Playfair's *Statistical Breviary* (1801), which showed multivariate data using area to depict quantity for the first time. The circle represents the area of the country, the line on the left of each circle represents the size of the population, and the line at the right the tax revenues collected. The slope of the dotted line indicates the extent of the tax load. It is easy to see that the small populations of Britain and Ireland stand out as the most heavily taxed of all nations included.

used to show joint properties. As in the case of the line and bar charts that he had introduced 15 years before, his basic designs were sound and have scarcely been improved upon. The areas of circles are used to represent varying quantities, and the practice of using circles, or areas of other figures, persists to this day. In the pie chart, Playfair used angle to denote proportion, and used color and labeling to differentiate the segments that make up the whole (Fig. 7). The use of Venn-like diagrams to portray statistical quantities is less common, both today and two centuries ago, but the device is not unknown.

However, the pie chart remains a mystery—Playfair left us no indication of its inspiration. And yet it is likely that he was copying or adapting the ideas of others—his career is replete with instances of adaptation. Perhaps a clue is to be found in the intersecting and included circles. Such diagrams were used by Venn in his work on logic in 1880—but, of course, Playfair's diagrams precede Venn's. Venn (1834–1923), contrary to popular myth, was not the inventor of such logic diagrams. Euler (1707–1783) had used them for exactly the same purpose more than a century before. And before Euler, Leibniz (164–1716) devoted serious attention to the analysis of logical propositions by means of diagrams: he explored various means of representing Aristotelian syllogisms by means of geometric figures, including Venn-like diagrams as well as his own linear versions, which he considered superior. It was, however, Euler who popularized the use of circle diagrams, although he was quick to point out that the type of shape was unimportant. It is interesting to observe that the work of Euler and Leibniz was well known to John Playfair, the author of "Progress of Mathematical and Physical Science since the Revival of Letters in Europe" in the fourth edition of the *Encyclopedia Britannica*. This article has been described as the best short general history of science written during the first half of the 19th century. His concluding discourse on the genius of Leibniz and Newton was universally admired. Because of his intimate familiarity with the work of Leibniz and Euler, John Playfair could not have failed to make William aware of this work as he instructed his 12-year-old younger brother in mathematics, after the early death of their father.

After the great inventions of 1786 and 1801, Playfair introduced no further innovations of any consequence. He did, however, refine his graphs, and later publications include rather splendid examples that combined the time-series line graph, a seamless sequence of bars depicting quantities averaged over fixed time periods, and a chronological diagram in a single chart (Fig. 8).

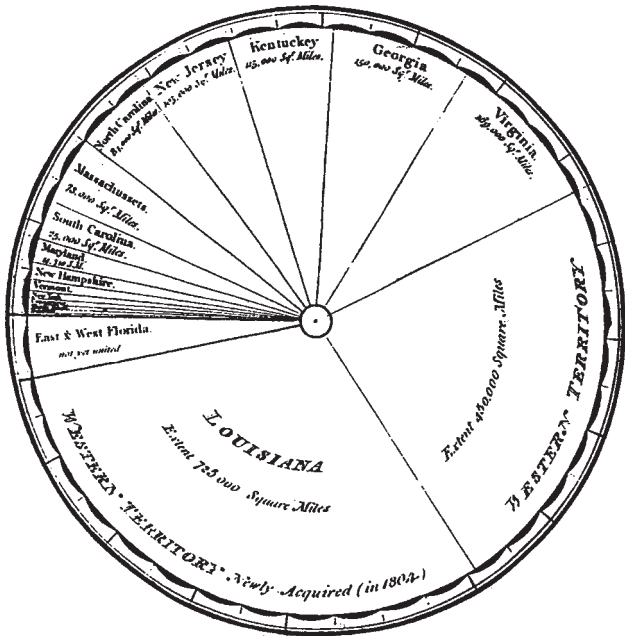


Figure 7 Playfair's pie chart using the segments of the pie to represent the relative size of each component of the United States at the beginning of the 19th century. Areas in the chart are proportional to the areas in square miles. From Donnant (1805).

Conclusion

Playfair's final two decades were not easy. He was in frequent financial difficulty, despite his involvement in a variety of schemes. These generally involved publishing, banking, and writing about economics. When he returned to London, Playfair and his partner Hartsinck opened the Security Bank, modeling its practices on schemes that he had seen in Paris during the Revolution. The London establishment, however, did not tolerate these unregulated innovations, and the venture collapsed after a conflict with the Bank of England. From the mid-1790s onward, he made his living as a writer and also as a gun carriage maker, developing the occasional new mechanical invention. He argued against the excesses of the French Revolution and commented extensively on British policy toward France. In his illustrated nine-volume *British Family Antiquity*, he catalogued the peerage and baronetage of the United Kingdom. He also edited more than one periodical, including the *Tomahawk*. After the restoration of the Bourbon monarchy, he returned to France and became editor of the periodical *Galignani's Messenger*, but his comments on a duel between a Colonel Duffay and the Comte de St. Morys were held to be libelous by the widow St. Morys and led to Playfair's prosecution. Sentenced to three months imprisonment, a fine, and damages, Playfair thought flight a better option, and he spent his remaining years in London writing. He constantly pushed the boundaries of legality and was convicted on more than one occasion.

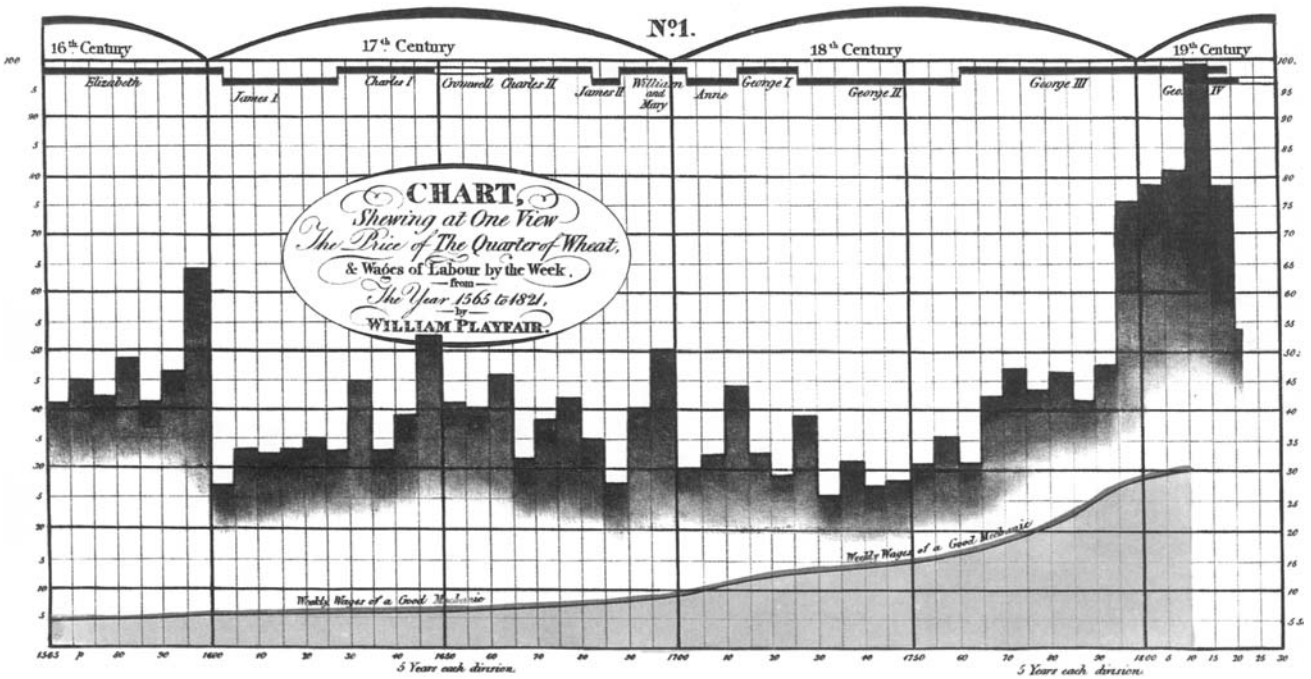


Figure 8 One of Playfair's most dramatic displays designed to show the unfair price of a quarter of wheat compared to the humble wages of a "good mechanic." In the original, the curve for the wages of the mechanic appeared in red and the area beneath was stained blue. This chart appeared in both editions of *Agricultural Distresses* (1821, 1822).

He even descended to a kind of genteel blackmail of acquaintances (for example, the famous engineer John Rennie) and more aggressively with strangers (Lord Archibald Douglas). Despite his considerable efforts, including a hopeful but brief return to France, his schemes failed to gain the fortune that he craved. He died in modest circumstances at the age of 63.

See Also the Following Articles

Graph Theory • Lazarsfeld, Paul • Time Series Analysis in Political Science

Further Reading

Biderman, A. D. (1990). The Playfair enigma: The development of the schematic representation of statistics. *Inform. Design J.* **6,** 3–25.

Donnant, D. F. (1805). *Statistical Account of the United States of America*. Translated from the French by William Playfair. Greenland & Norris, London.

Euler, L. (1761). *À une princesse d'Allemagne sur divers sujets de physique et de philosophie*, Vol. 2, Letters No. 102–108.

Funkhouser, H. G. (1937). Historical development of the graphic representation of statistical data. *Osiris* **3,** 269–404.

Lambert, J. H. (1779). *Pyrometrie.* Berlin.

Palsky, G. (1996). *Des chiffres et des cartes: Naissance et développement de la cartographie quantitative français au XIX^{eme} siécle.* Éditions de CTHS, Paris.

Playfair, W. (1786). *The Commercial and Political Atlas.* Corry, London.

Playfair, W. (1801). *The Commercial and Political Atlas,* 3rd Ed. John Stockdale, London.

Playfair, W. (1801). *The Statistical Breviary.* T. Bensley, London.

Donnant, D. F. (1805) *Statistical Account of the United States of America*. Translated from the French by William Playfair. Greenland & Norris, London.

Playfair, W. (1807). *An Inquiry into the Permanent Causes of the Decline and Fall of Powerful and Wealthy Nations.* Greenland & Norris, London.

Playfair, W. (1821). *A Letter on Our Agricultural Distresses, Their Causes and Remedies.* London: W. Sams.

Playfair, W. (1822). *Can This Continue? A Question Addressed to Whom It May Concern.* W. Sams, London.

Priestley, J. (1765). *A Chart of Biography.* London.

Priestley, J. (1769). *A New Chart of History.* London. (Reprinted 1792, Amos Doolittle, New Haven.)

Spence, I., and Wainer, H. (1997). William Playfair: A daring worthless fellow. *Chance* **10,** 31–34.

Wainer, H., and Velleman, P. (2001). Statistical graphics: Mapping the pathways of science. *Annu. Rev. Psychol.* **52,** 305–335.

Police Records and the Uniform Crime Reports

Scott Akins
Oregon State University, Oregon, USA

Glossary

classifying The Uniform Crime Reports (UCR) requirement that local law enforcement agencies report crimes to the UCR in a manner that is consistent with the crime categories employed by the UCR.

hierarchy rule The UCR requirement that in a multiple offense situation only the most serious offense should be recorded.

National Incident-Based Reporting System (NIBRS) A program for measuring crime that uses each criminal occurrence rather than summary counts and has been argued to be better than the UCR for measuring the true volume of crime.

National Crime Victimization Survey (NCVS) A self-report survey conducted by the U.S. Bureau of the Census on behalf of the Bureau of Justice every 6 months that measures respondents' experiences with rape, robbery, assault, burglary, larceny, and motor vehicle theft.

Part I (index) offenses Eight crimes monitored by the UCR for which more extensive data are available and that are intended to give a reasonable picture of crime in the United States. Part I offenses include murder and nonnegligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny, motor vehicle theft, and arson.

Part II offenses Crimes that are considered by the UCR to be either less serious or serious but relatively infrequent. The data available for Part II offenses are less extensive than for Part I offenses and are limited to arrest data alone.

scoring The process by which local law enforcement agencies "count" the number of offenses present in each crime category dependent on certain conditions stipulated by the UCR program.

self-report survey A research method in which people are asked to reveal information about their involvement in crime.

Uniform Crime Reports (UCR) The most widely used source of official data on crime in the United States. Maintained by the Federal Bureau of Investigation, the UCR collects and reports data on offenses reported to police and arrested offenders from local police agencies throughout the country.

Police records are those data collected as a consequence of the regular operation of the police. The most prominent source of official data on crime is the Uniform Crime Reports (UCR) and its developing National Incident-Based Reporting System. It is important to note that these data, like all data, suffer from some shortcomings. However, UCR data provide an invaluable contribution to our knowledge of crime, and the purpose of this article is to familiarize the reader with the strengths and shortcomings of the data so that they can be used with these issues in mind.

Uniform Crime Reports: Historical Perspectives

The Federal Bureau of Investigation's (FBI) Uniform Crime Reports (UCR) is the most widely used source of data on crime in the United States. Prior to the establishment of the UCR program in 1930, systematic data on crime across jurisdictions were unavailable. However, in the late 1920s, the International Association of Chiefs of Police (IACP) formed a committee to promote a national crime statistics program. In part, as noted by researchers such as Maltz, these efforts were aimed at preventing newspapers from manufacturing "crime waves" to boost paper sales and in part because estimates and comparisons of crime trends in the country were largely impossible due to differences in crime definitions, recording procedures, and recording accuracy. The IACP proposed seven types of crime that should be monitored for an understanding of crime in the United States: murder

and nonnegligent manslaughter, forcible rape, robbery, aggravated assault, burglary, larceny, and motor vehicle theft. According to the FBI, these particular crimes were selected on the basis of their seriousness, frequency of occurrence, pervasiveness in all geographic areas of the country, and likelihood of being reported to law enforcement. Referred to as Part I offenses or "index crimes," these offenses were and are the primary basis for the UCR program.

Although the UCR program has experienced few fundamental changes since its inception in 1930, some minor modifications have been made. The first significant revision of the UCR program came in 1958 when statutory rape, negligent manslaughter, and the larceny of less than $50 were all reclassified from Part I to Part II offenses. In this same year, the FBI began to make estimations of crime rates for the nation as a whole and to use a composite measure of all index crime, referred to as the Crime Index. In 1974, larcenies of less that $50 were once again included as a Part I offense, and a key revision of the UCR program came in 1979 when arson became the eighth index crime. As noted by Mosher *et al.*, the addition of arson was the result of a congressional mandate in October 1978 and occurred despite the protests of FBI officials that the accurate classification and recording of arsons would be particularly problematic. Although the passage of the Anti-Arson Act of 1982 permanently established arson as an index crime, arson figures are not included in Crime Index totals because these figures are not consistently available. The last significant modification of the UCR program (excepting the development of the National Incident-Based Reporting System, discussed later) came in 1990 when the UCR began to collect data on hate crime as a result of the Hate Crime Statistics Act.

Uniform Crime Reports: Data Collection

In addition to the information collected on Part I offenses, the UCR program also compiles data on crimes that are considered to be either less serious or serious but relatively infrequent (e.g., kidnapping). These are referred to as Part II or "nonindex" offenses. Part II offenses include white-collar crimes, such as fraud, embezzlement, and forgery/counterfeiting; public order crimes, such as drug offenses, gambling, and prostitution; and other crimes, such as simple assaults, minor theft, sex crimes excepting forcible rape and prostitution, disorderly conduct, and vagrancy. The reporting of data on Part II offenses is optional for those departments that participate in the UCR program, and the data available for these crimes are less extensive than for Part I offenses, being limited to information on arrests alone.

Although the FBI considers the participation of law enforcement agencies in the UCR program "strictly voluntary," Schneider and Wiersema note that many states have mandated that their law enforcement agencies participate in the UCR program. Despite its "voluntary" nature, the UCR program enjoys impressive participation rates. Indeed, as noted by Mosher *et al.*, a total of 16,788 state, county, and city law enforcement agencies, covering more than 272 million inhabitants, submitted crime reports to the UCR 1999. The total population covered by the UCR is an impressive 97%. However, coverage outside of urban areas drops slightly, falling to 90% in cities outside metropolitan areas and to 87% in rural areas.

The actual data collection process occurs each month as the FBI gathers crime statistics in the form of UCR tally sheets and report forms. These are obtained from local law enforcement agencies or state agencies designed to act as intermediaries between local law enforcement and the UCR program. Cross-checks of the reported data and on-site training and consultation are performed by the FBI in an attempt to maintain quality control in data collection and to ensure that the data are comparable across jurisdictions. This can prove challenging at times because an essential part of the UCR data collection process is the requirement that local law enforcement agencies report crimes to the UCR in a manner consistent with the crime categories employed by the UCR, referred to as classifying. Many local jurisdictions have definitions of crimes that differ from the definitions employed by the UCR, and classifying is apt to be particularly problematic in these instances. Following the classification of offenses, local agencies then score the offenses, or count the number of offenses in each crime category depending on certain conditions. Both the process of classifying and scoring are potential sources of error in UCR data and will be discussed in more detail later.

Sources of Ambiguity in Uniform Crime Reports Data

Despite the extensive use of UCR statistics, these data are subject to several well-known criticisms. These criticisms can be broadly grouped into those that address the reporting of crimes to the police, methodological problems with UCR coding procedures and measures, and questions of whether UCR data reflect the behavior of criminals or the behavior of police. These methodological and measurement problems have been the subject of extensive discussion and debate, and this article only touches on the many issues. Readers seeking a more in-depth examination are encouraged to see the excellent

reviews of this literature provided by Mosher *et al.* and Schneider and Wiersema.

Issues of Reporting

Because official data on crime are derived from the regular operation of the justice process, they incorporate all the issues and problems that are inherent in this process. One of these problems is underreporting. Policing is a very reactive process, with police depending heavily on regular citizens to report crimes. Indeed, unless a crime is witnessed and the witness decides to report it, the crime will typically remain unknown to the police. Because many crimes remain unnoticed by the public, and because even when crimes are witnessed many choose not to report them to the police, the majority of all crime that is committed never becomes known to the police and recorded.

This underreporting of criminal behavior in official police statistics, often referred to as "the dark figure of crime," is known to be substantial. Even among serious crimes levels of underreporting typically exceed 50%, although this varies by the type of crime and numerous other variables (discussed later). Information on the extent of underreporting became increasingly available in the late 1960s with the development of victimization surveys, which avoid some of the problems that characterize official data on crime. Indeed, the National Crime Victimization Survey (NCVS), the nation's largest and most comprehensive victimization study, was created in large part to address the shortcomings inherent in official sources of data on crime such as the UCR.

There are several reasons for underreporting. Often, people believe that the crime in question is not serious enough to merit police involvement, such as in a case of vandalism or a minor theft. Other times, people may believe that involving the police will not help them to remedy the situation in any way and will just be "one more hassle to deal with." Additionally, many crimes are never reported because of their "victimless" nature, being entered into consensually by all parties involved (e.g., drug use and sales and prostitution), and even when there is a victim in a crime many choose not to report their victimization out of a fear of retaliation (e.g., a battered wife or girlfriend) or because they distrust or fear the police. Underreporting also results because victims of crime who are engaged in criminal activity, particularly at the time of their victimization, may be especially unlikely to report an offense out of fear they will be punished (e.g., rape of a prostitute or robbery of a drug dealer). These people may choose to "just forget" the incident or to settle it themselves, which may result in more (potentially unreported) criminal activity.

Further complicating these problems is that rates of underreporting are likely to vary based on a number of factors, such as the type of crime, the demographic characteristics of the victim, the location of the crime, and the relationship between the victim and the offender. Perhaps most influential among these factors is the type of crime involved. Violent crimes are reported at higher rates than are property crimes (excepting motor vehicle theft, which is highly reported for insurance purposes), particularly for the crimes of homicide and robbery. However, although the seriousness of the crime plays an important role in reporting, many other factors are also important. Indeed, rape, an extremely serious crime, is among the most underreported of offenses. The race and gender of victims have also been found to influence reporting, with women and blacks being more likely than men and whites to report their victimization, as has region, with people in the southern states being more likely than those in other regions to report crimes. The relationship between the victim and the offender has been found to influence reporting. Indeed, Turner found that people are more likely to report crimes committed by strangers than by those they know. Levels of underreporting may also vary with time and circumstance. Indeed, evidence seems to indicate that rape, although still highly underreported, is probably less underreported now than a decade or two ago as public awareness and sensitivity about the crime and its victims have grown. Thus, it is particularly important to realize when using UCR data that not only does underreporting exist but also the levels of underreporting are not constant across offense type and many other variables.

After citizens report incidents of crime to police they may still be left unrecorded. The wide discretion granted to police enables them to decide, based on a number of factors, whether a particular offense necessitates an arrest or citation. Verbal warnings are an extremely common response by police, particularly in dealing with minor forms of crime, leaving many offenses unrecorded. This situation is complicated by the fact that an officer's decision to formally record a reported crime is influenced by a number of factors. Black's classic study on this topic concluded that variables such as the seriousness of the crime, the complainants' desire to see the matter taken further, the relationship of the offender and victim (found to be the most important factor), and the complainants' social class and behavior toward the officer all had significant effects on whether the crime was officially recorded. Police recording practices are also likely to vary by region and jurisdiction, and all these factors are likely to make comparisons across jurisdictions problematic.

Another potential source of underreporting (or misreporting) that may occur after the crimes have been reported to police deals with the alteration of data by law enforcement administrators for political or budgetary purposes. Articles such as "How to Cut City's Crime Rate: Don't Report It" reported in the *Philadelphia Inquirer* and "As Crime Falls, Pressure Rises to Alter Data" reported in the *New York Times* attest to the pressure

police administrators are under to show decreased crime rates. As Mosher *et al.* noted, methods of "cooking the data" to reduce overall crime counts may involve practices such as not reporting all crime incidents to the UCR, downgrading Part I offenses to Part II offenses so they are not included in national UCR summaries, and misusing the hierarchy rule to combine many separate incidents into a single (artificial) multiple-offense incident. Those such as Maltz have noted that although the pressure on police administrators to alter reported crimes has decreased somewhat in recent years, these temptations will likely always be present.

Reporting problems can also arise from the actions of police administrators even when they are acting in a completely even-handed manner. Changes in policy or standard police procedures may inadvertently have a major impact on the rate of crimes reported. These changes can be formal, such as providing an "emergency 911" number or mandating arrest for domestic abusers, but they can also be informal, such as a police chief telling his officers to arrest prostitutes as a result of an embarrassing newspaper report. Regardless of the cause, reported crimes may be substantially influenced because of a change in the official response to crime rather than any actual change in criminal behavior.

A final problem related to reporting in the UCR is the lack of attention given to particular types of crime. Specifically, UCR data do not address federal or political crimes, and they severely undercount organizational and occupational crime. The inadequacy of official data to address white-collar crime dates back to Edwin Sutherland's presidential address to the American Sociological Association in 1939, but only recently have any significant steps been taken to address this shortcoming. Measures of white-collar crime in the UCR are limited to fraud, forgery/counterfeiting, and embezzlement, all of which are Part II offenses with only limited information. However, the development of the National Incident-Based Reporting System (NIBRS) to enhance the UCR has addressed this problem somewhat by providing much more information on white-collar offenses. Although some reason that the UCR has paid little attention to white-collar crime because these crimes are much more difficult to detect and report on, others argue that the FBI is simply biased in favor of the upper class and thus chooses to focus on "street crimes" rather than "suite crimes" because the former are more likely to be committed by members of the underclass.

Methodological Problems with UCR Data

One of the most problematic aspects of the UCR data collection process is classifying, or taking the numerous offense titles employed by jurisdictions across the country and grouping the incidents of these categories into the appropriate standardized crime categories used by the UCR. For example, legal definitions for the crime of rape vary greatly by jurisdiction. Some laws require forced penile–vaginal penetration for a rape to occur, whereas others stipulate that other types of offensive intimate contact constitute rape. Jurisdictions may require some degree of physical "resistance" on the part of the victim for it to be considered a rape, and some may or may not allow males to be considered victims of rape. Finally, some jurisdictions may completely abandon the term of rape and classify these acts under categories such as sexual assault in the first, second, or third degree. Regardless of the legal strategy employed by the local jurisdiction, when reporting crime data to the UCR the local jurisdiction must conform to the UCR definition of the crime. This process of classification, discussed briefly earlier, can be difficult for police to accomplish and is certain to introduce some error into UCR data. However, the availability of technology and software developed for the automated reporting of UCR and NIBRS data may help to manage these problems somewhat in the future.

The process of scoring, the counting of offenses after they are classified, can also be problematic because offenses may not be counted if they meet certain conditions. When scoring, the UCR requires that for crimes against property (i.e., burglary, larceny/theft, motor vehicle theft, and arson) one offense is counted for each "distinct operation," defined as a criminal incident that is separate in time and place. In the instance of a property crime that victimizes more than one person at the same time and in the same place (e.g., the burglary of three housemates), the distinct operation criteria requires that each of the persons offended against is not counted as a separate crime.

Conversely, for crimes against persons (i.e., criminal homicide, forcible rape, robbery, and aggravated assault) each offense that has a victim is counted as one offense. For example, if a person were to murder two people in a robbery, each murder would be recorded by the UCR because there are two victims. However, although both of the homicides would be recorded, not even one robbery would be recorded. This is because scoring must also take into account the hierarchy rule, one of the most frequently discussed and debated issues with regard to counting procedures used by the UCR.

The hierarchy rule states that in a multiple-offense situation (e.g., robbery–homicide), after classifying all the Part I offenses, only the most serious offense should be scored and the rest ignored. Thus, in the previous example, each homicide is recorded but each robbery is ignored. This obviously results in a systematic underreporting of crime, but it also raises concerns about differential levels of compliance with the hierarchy

rule among reporting agencies. The only important exception to the hierarchy rule occurs for the crime of arson, which is always reported, even in multiple-offense situations. It is important that UCR data are used with these issues in mind because, as noted by Mosher *et al.*, it is likely that the index crimes are subject to considerable variability in counting and scoring across reporting agencies.

Additional methodological criticism of the UCR addresses the extent to which the crime measures used by the UCR are flawed. Criticisms have been lodged against the UCR for aggregating offenses that are conceptually dissimilar into a single category (e.g., shoplifting and purse-snatching), excluding important supplementary information about the crimes (e.g., the victim–offender relationship), and including attempted crimes along with completed crimes. Comment has also been made on the fact that the UCR uses estimates of population counts in noncensus years when calculating crime rates, which can be problematic particularly in times of rapid population growth or decline. Finally, some concern has been raised over the size of the sampling error and sampling bias that may be introduced when the UCR generates crime estimates for those areas where agencies were either unable or unwilling to provide data to the UCR.

Official Statistics: A Measurement of Criminal Behavior or Police Behavior?

As noted by Wolfgang, the last and perhaps most fundamental criticism of the UCR addresses the extent to which the data reflect the behavior of police rather than criminals. As discussed earlier, levels of reported crime in official data may be heavily influenced by discretionary decisions on the part of officers, changes in police procedure, or intentional tampering with the reported figures for political reasons. In addition to these concerns is the problem of race and social class biases in policing, which has been demonstrated by numerous studies. For example, Skolnick found that police often act in harsher or more confrontational manners when policing minority communities in part because they believed minority youth to have a greater potential for violence. Analogous work by Smith examined 60 neighborhoods of varying racial and economic composition and found that police were three times as likely to arrest a suspect in a poor and predominately minority neighborhood, regardless of the crime involved, compared to a higher class and predominately white neighborhood. Similarly, in his ethnographic study of urban communities, Anderson noted that police appear to engage in an informal policy of monitoring young black men as a means of controlling crime, and they often seem to go beyond the bounds of their duty. Finally, work by Chambliss on policing the underclass concludes that if police patrols paid half as much attention to the crimes of students at universities as they do to the crimes of young black males, the arrest and incarceration rate for young white males would be quite similar to those for young black males. In summary, bias on the part of police likely results in minorities and members of the lower class being more harshly treated by police, resulting in disproportionately high rates of citation and arrest. Furthermore, the degree to which the data are influenced by these factors is likely to vary by the type of crime (e.g., serious crimes appear to be less influenced by these factors) and other relevant variables. Therefore, it is important that the data be used with these issues in mind.

Despite the numerous problems with UCR data, they are a good source of data on crime and are extensively used by researchers. Not only are easily accessible and constantly updated they also provide excellent coverage over time. Additionally, the extensive data on homicide provided by the UCR's supplementary homicide reports are exceptionally good because nearly all homicides come to the attention of law enforcement and a large portion result in arrest. Although underreporting is an issue with UCR data, levels of underreporting are relatively stable across time for most crimes. Thus, UCR data allow us to effectively examine changes in crime patterns over time, even if the total amount of crime captured by the UCR is known to be inaccurate. In summary, it is important to note that the criticisms lodged against UCR data should not preclude the use of these data but rather warn researchers about the weaknesses present so that the data can be used with an awareness of their limitations.

The National Incident-Based Reporting System

The development of the NIBRS resulted partly in response to the criticisms levied at the UCR. The incident-based counting system employed by the NIBRS allows the collection of extensive, detail-rich data on crime that are superior to the summary data currently available through the UCR. The NIBRS collects data on each criminal incident, defined as one or more offenses committed by the same offender or group of offenders acting in concert, and arrest within 22 broader offense categories containing 46 specific crimes. Information is provided on the number of offenses that occur in each incident (avoiding the hierarchy rule), with extensive details on the offense(s) also provided. These details include whether the crime was only attempted or actually completed, whether the offender was suspected of using alcohol or drugs, whether the offender was suspected of using a computer in the commission of the crime (one

potential indicator of white-collar crime), whether the crime was motivated by bias, the location of the crime, and whether a weapon or force were used in the commission of the crime. Additionally, for each incident, data are collected on both the offender(s) and the victim(s) (regardless of whether an arrest has been made). Information on the offender(s) includes age, sex, and race. Information on the victim(s) includes the number of victims per incident (up to 999); the age, sex, race, ethnicity, and resident status of the victim(s); whether an injury was sustained by the victim(s) and, if so, what type; the relationship of the victim(s) to the offender(s); the circumstances surrounding the incident in the case of an aggravated assault or homicide; additional information on the incident circumstances in the case of a justifiable homicide; and an ID number identifying the offender(s) in the incident so that those data can be cross-referenced (although data on victims cannot be accessed through the offender records). Furthermore, for each incident, data are collected on any property that is involved (e.g., the type of property, the value, and whether it was destroyed or seized), as well as information on arrests, such as whether an arrest was made; the date of arrest; the arrest offense code; any weapon(s) the arrestee(s) possessed; and the age, sex, race, ethnicity, and resident status of each arrestee.

As noted by Chilton, the extensive volume of information collected by NIBRS is magnified by the possibility of multiple offenses, multiple victims, multiple offenders, and multiple arrests all within a single incident. Obviously, although these data are extremely impressive in terms of the volume and quality of information they can provide, they also involve further challenges in data coding and recording that will take time to implement at the local level. The FBI first began accepting NIBRS data from local agencies in 1989, and as of early 2002 it reported that there were 4192 law enforcement agencies contributing NIBRS data to the UCR program. This provided coverage for 17% of the U.S. population and accounted for 15% of the crime statistics collected by the UCR. The limited, though improving, national coverage provided by the NIBRS is indicative of the challenges faced by those attempting to implement it. For example, the tremendous complexity of the NIBRS coding process makes participation in the program by local law enforcement agencies problematic. Indeed, a widespread belief exists among law enforcement personnel that implementing the NIBRS will be very costly for local law enforcement agencies, and administrators may wonder whether it is the best use of their resources. Roberts points out that police officials often feel burdened by a complex data entry process that appears more valuable to researchers than to law enforcement agencies. Additionally, there is a widespread perception among law enforcement personnel that reported crime will increase with the adoption of NIBRS, partially due to the elimination of the hierarchy rule. Although the fact that this perception exists is of key importance for implementing the NIBRS at the local level, a study by Rantala and Edwards found little evidence to support it. Comparing NIBRS and UCR data for the same years and jurisdictions, the study found that when using NIBRS data rates of homicide remained the same; rape, robbery, and aggravate assault rates increased by approximately 1%; larceny rates increased by 3.4%; motor vehicle theft rates increased by 4.5%; and burglary rates decreased by approximately 0.5%.

In summary, although the NIBRS presents a tremendous opportunity for criminologists, the detailed nature of the data requires a meticulous and complex recording process that will take time and resources to implement. The data are tremendous in terms of their scope and address many of the problems associated with the UCR—although those dealing with data coding and entry may actually be greater, at least for the time being. It is also important to recognize that because local police agencies are vital to the collection and reporting of official crime data, it is essential that they be convinced of the value of the NIBRS for it to replace the UCR as the central source of official crime data in the years to come.

Official police data, such as the UCR and the developing NIBRS, are some of the most valuable and frequently used data on criminal behavior. Like all forms of data, they suffer from some shortcomings. The most salient of these deal with issues of underreporting; methodological problems, such as inappropriate classification and inadequate measures; and problems regarding the extent to which official police data reflect the behavior of police rather than criminals. However, research based on official police data has contributed immensely to our knowledge of crime, and these data can be extremely valuable if used with an appropriate awareness of their limitations.

See Also the Following Articles

Correctional Facilities and Known Offenders • Criminal Justice Records • Criminology • Experiments, Criminology • Jevons, William Stanley

Further Reading

Anderson, E. (1990). *Streetwise.* University of Chicago Press, Chicago.

Black, D. (1970). Production of crime rates. *Am. Sociol. Rev.* **35,** 733–748.

Chambliss, W. (1995). Crime control and ethnic minorities: Legitimizing racial oppression by creating moral panics. In *Ethnicity, Race and Crime* (D. Hawkins, ed.), pp. 235–258. State University of New York Press, Albany.

Chilton, R. (1998). Victims and offenders: A new UCR supplement to present incident-based data from

participating agencies. Paper presented at the annual meeting of the American Society of Criminology, Washington, DC.
Federal Bureau of Investigation (2002). *Crime in the United States.* U.S. Government Printing Office, Washington, DC.
Maltz, M. (1977). Crime statistics: A historical perspective. *Crime Delinquency* **23,** 32–40.
Maltz, M. (1999). *Bridging Gaps in Police Crime Data: A Discussion Paper from the BJS Fellows Program.* U.S. Government Printing Office, Washington, DC.
Mosher, C., Miethe, T., and Phillips, D. (2002). *The Mismeasure of Crime.* Sage, Thousand Oaks, CA.
Rantala, R., and Edwards, T. (2000). *Effects of NIBRS on Crime Statistics*. U.S. Department of Justice, Washington, DC.
Roberts, D. (1997). *Implementing the National Incident-Based Reporting System: A Project Status Report.* U.S. Department of Justice, Washington, DC.
Schneider, V., and Wiersema, B. (1990). Limits and uses of the Uniform Crime Reports. In *Measuring Crime* (D. MacKenzie, P. Baunach, and R. Roberg, eds.), pp. 21–48. State University of New York Press, Albany.
Skolnick, J. (1966). *Justice without Trial.* Wiley, New York.
Smith, D. A. (1986). The neighborhood context of police behavior. In *Communities and Crime* (A. Reiss and M. Tonry, eds.), pp. 313–341. State University of New York Press, Albany.
Turner, A. (1972). *Methodological Issues in the Development of the National Crime Survey Panel: Partial Findings.* National Criminal Justice Information and Statistics Service, Washington, DC.
Wolfgang, M. (1962). Uniform Crime Reports: A critical appraisal. *Univ. Pennsylvania Law Rev.* **3,** 708–738.

Political Conflict, Measurement of

Monty G. Marshall
University of Maryland, College Park, Maryland, USA

Glossary

authority The power to direct, manage, or otherwise control political conflict.
co-optation Neutralizing conflict either by satisfying the personal interests of opposition group members or through the inclusion of opposition group goals.
demonstration A public display of group support for a political position or of group identity, strength, and cohesion.
event A single occurrence or episode of a particular type or classification of a social phenomenon.
imposition Negating conflict through the demonstration of superior force.
mobilization Organizing, preparing, and directing a social group for political action.
negotiation Neutralizing conflict through bargaining and agreement between opposing groups.
protest A principal form of communicating opposition or grievance to an established political authority group.
revolution A sociopolitical movement to overthrow an existing government and replace it with something new; alternatively, a sudden, momentous or radical transformation of the structures and values of society and governance.
riot A violent disturbance involving an assembled group; riots may result from spontaneous action or an escalatory sequence during a nonviolent demonstration; riots may also be a planned event.
separation Neutralizing conflict by constraining interaction(s) between otherwise opposing groups; separatism is often made salient by different and opposing ethnic identity groups.
war Major, open armed conflict between disciplined militant organizations representing the interests of larger opposing groups engaged in a political conflict.

The measurement of political conflict presents tremendous challenges to social scientists engaged in the quantitative analysis of political behavior. The term "political conflict" can have many different meanings depending on the context in which it is used; the appropriate measurement of political conflict depends upon meaning and context. It is crucial that any measure of political conflict phenomena include a complete and explicit delineation of its definitional parameters. In its most general sense, political conflict refers to a public demonstration of opposition, or disagreement, between two or more social groups. Social groups themselves are quite commonly organized through the impetus, or mobilization potential, of conflict. Political conflict (conflict between groups) is usually distinguished from social conflict (conflict involving individuals in a social group context) and interpersonal conflict (conflict between individuals) by reference to notions of "sovereignty," the primacy of state authority in the management of group conflict, or the inherently political nature of social group organization. Classic definitions of political conflict derive from Machiavellian principles of "state's rights" (*raison d'Etat*) and focus on the direct behavior of the sovereign state in its relations with other states (interstate conflict), direct applications of state coercion to members of constituent groups (repression), and direct challenges to state authority by constituent groups (civil conflict). Others, arguing a human rights perspective, claim that all conflicts between groups and individuals must be considered political conflict, regardless of the role of the state in defining the conflict or the presence or absence of direct actions by state authorities. From this perspective, the legitimacy of the state may be measured by the qualities of its involvement in political conflict. One can even speak of personal conflicts as political conflict as each individual struggles with internal tensions deriving from conflict between personal desires and the political nature of the need for social acceptance or approval. As political conflict involves directed group behavior in an

interactive sequence with an opposing group, it is necessarily conditioned by established authority and governance structures; that is, governing institutions are either directly or indirectly involved in all political conflict situations.

The two principal forms that political conflict may take are protest (voice) and coercion (action). The presence of political conflict impedes, or even thwarts, the direct pursuit and attainment of a group's professed goal or goals. As such, political conflict produces a contention that must be resolved before goal-oriented progress can be restored. It is generally recognized that the crucial factor in the definition of political conflict is its potential for escalation, that is, the possibility of a "resort to force" as the "final arbiter" in a conflict interaction and the attempt to negate the resistance of an opposition group (or groups) and impose a resolution of the conflict by one or more parties to the conflict without the support of the opposition. The coercive resolution of political conflict includes the implicit or explicit threat of political violence, that is, by enforcement, repression, or armed conflict. Enforcement actions are prescribed by law (rule of law); acts of repression are directed by the political interests of governing authorities (rule of force); armed conflict occurs when established authority structures are insufficient for commanding loyalty and discipline or ineffective in resolving important disputes. As armed conflict is the most dramatic and "visible" form of political conflict, the two terms are commonly equivocated simply as "conflict." This equivocation of the most common forms with the most extreme forms of social interaction tends to prejudice our understanding of the nature of political conflict and the role of force and violence in conflict behavior.

Measuring Social Phenomena

Constraints and Limitations

Conflict is at once an essential quality and principal feature of human social behavior. Louis Coser has even argued that conflict is a necessary condition to stimulate group formation and collective activity and as a *raison d'être* for social identity and group cohesion. In this formulation, some amount (or type) of conflict is seen as beneficial, or functional, in the transformation of individual to social behavior. Yet, it appears that too little conflict or too much conflict may be detrimental to the quality of social relations and the success of collective action. In a social setting, individuals politicize their interests and learn to cooperate with others (mobilize) in order to better manage social conflict and pursue common goals. As such, conflict, in its broadest sense, is a core, common, and continual feature of human relations, as is the purely individualized quality of choice of action in response to the stimulus posed by conflict. The perception of conflict is inherently subjective and the decision to act, whether alone or in concert, is necessarily complex and steeped in individualized ideas, attitudes, and preferences. However, group behavior requires more or less open communication and a convergence of ideas and attitudes among group members that makes the process of political conflict more transparent and regular, therefore, more comprehensible, predictable, and measurable.

If we accept the concepts of "free will" and "rational choice" in human behavior then we must also accept the precept that behavior is not deterministic but, rather, strategic. Individuals have a choice in how they will act and react in conflict situations and, in making choices in the pursuit of goals, behavior tends to be both rational and strategic: options are selected according to their perceived utility in attaining preferred goals. Coser goes on to identify three conflict mobilization options: participation, innovation, or rejection. Albert Hirschman provides a similar trilogy of political relations: loyalty, voice, or exit. These three main modes of social or political interaction provide the basis for distinct measures of attributes, qualities, and dynamics of conflict in the societal context. The first mode focuses on the role of conflict behavior in the creation and maintenance of the social system, the second mode focuses on system change through adaptation, innovation, or expansion, and the third focuses on separation and the potential for contention between opposing groups. While each perspective provides crucial information and the basis for unique measures of political conflict behavior in its own right, all three perspectives taken together provide a more comprehensive examination of political conflict in the social context. However, serious data limitations make such comprehensive examinations difficult, if not impossible, in many applications.

Three major limitations on the availability of political conflict data for measurement purposes are the result of administrative, technological, and statist constraints. Administrative constraints refer mainly to incomplete and imperfect record keeping. The measurement of social phenomena depends on the collection and recording of vital information; this, in turn, depends on the quality of knowledge and direction regarding what information to acquire and keep, the capacity to collect and record that information thoroughly, and the ability to maintain those records. Such extensive administrative capacity is a relatively recent and resource intensive development that has been largely monopolized by the state, especially in regard to the scope, consistency, and density of the historical record. Most historical records are highly selective and stylized accounts. Various authors, scholars, and historians have augmented shortcomings in the public record with private efforts but even these efforts have been subject to critical influence by state authorities.

Private contributions to the public record are particularly important in regard to information concerning political conflict, as will be discussed further below.

The administration of the public and historical record is heavily dependent on prevailing information and communication technologies. As our general technological capacity to collect, store, and retrieve information has expanded, so, too, has the scope and density (and accuracy and reliability) of the records available. It is really only with the expansion of the independent news media and the advent of the computer in the latter half of the 20th century (and, particularly, the personal computer in the late 1980s) that the data collection and record keeping enterprise has come of age. Even so, in most cases, the state retained a dominant and controlling grip on information resources through the end of the 20th century. The enormous costs of data administration remain daunting, if not prohibitive, for most private enterprises. Further complicating the scope and quality of the public record are the obvious security and strategic implications associated with information regarding political conflict. The interests of state security and sovereignty, particularly prior to the end of the Cold War in the early 1990s, very often led state authorities to actively suppress, or even distort, rather than effect and facilitate the collection of political conflict information. This suppression was especially acute in regard to internal conflict situations. A general recognition of the power inherent in the monopoly control and manipulation of information and the powerful role of information in the political conflict process, has contributed to greater emphasis on private, independent media; free and open communication; and expanded role of international organizations in the collection and dissemination of information in the contemporary period.

The state's preeminent role in data collection is even more apparent in the structure of the information that is available. The state, or country, is the primary focal group and unit of analysis for the vast majority of all political data and measures. The global state system strongly conditions the historical record; nearly all data is collected with the state as the measured unit. Problems for analysis arise from the fact that states themselves vary quite dramatically across many important attributes and, so, these politically "sovereign" units may not be statistically "comparable" units. Historical data focusing on other substate, or subnational, social units is extremely rare and severely restricted for most states for most years. Analysis of the internal conditions and dynamics of political conflict within states has been hampered by the lack of disaggregated information concerning variations across substate units.

The historical record is, of course, unevenly distributed. The more powerful and affluent countries are much more likely to keep and maintain detailed records on a much broader array of conditions and situations. Poorer countries may not keep even the most rudimentary records. Adequate record keeping is even less likely to occur during periods of general social turmoil and political conflict. Interestingly, Nazi Germany with its obsessive militancy, and cruelty, kept extensive and meticulous records of its crimes and those records have provided a rare insight into the intensity and totality of the war enterprise, both internal and external, for both perpetrators and targets.

In short, social phenomena are complicated situations nested within complex conditions and circumstances involving multiple, independent actors and, so, are difficult to measure directly or definitively. As such, social phenomena present serious difficulties for both measurement and analysis. The measurement, and quantitative analysis, of political conflict is a relatively recent development in the social sciences and we must keep this in mind when examining and evaluating progress in this endeavor.

General Principles of Measurement and Analysis

The inherent complexities of social phenomena present challenges for measurement and analysis. In fact, the use of the term "measurement" in regard to social phenomena is misleading and imparts a false sense of precision. "Estimation," "classification," and "coding" are the preferred terms for the most common procedures of collecting, organizing, and recording information on social phenomena. Measurement and analysis are intricately intertwined in the social sciences: measures derive, at least partially, from theory and analysis and analysis is critically conditioned by the measures available, as well as the measures actually used in the analysis. Measures are further refined as a result of advances made in analysis. In social group behavior even the most straightforward attributes are rarely, accurately known mainly because social actions involve varying degrees of spontaneity and the actions of interest cannot be isolated from surrounding or related activity. For example, the simple number of individual participants in a social action very often cannot be accurately or reliably measured because no direct procedure is employed to count and track such involvement, people move in and out of the situation over the course of time, the participatory actions of people who are not physically present may have been crucial in creating the situation, the action may attract partial or indirect involvement from bystanders or others who are stimulated by and/or react to the social action, etc.

Because of the complexities of social actions, these phenomena are usually subject to multiple levels and types of measurement. The general levels, or modes, of measurement are four: typology (classification), attributes, qualities, and dynamics. The initial step in

measurement is classification: identifying which of the myriad social phenomena constitutes a distinct type of "case" or "event" of analytic interest and can be consistently distinguished from other cases and events; that is, deciding what needs to be measured. A "case" typically references a particular social group or "actor" over a specific time frame; an "event" refers to a specific type of action. The most simple and ideal classification scheme would dichotomize, or separate, all social actions into events and nonevents with few "borderline" cases (i.e., cases that are difficult to classify). A more complex scheme would parse cases or events into multiple, related categories or typologies. The issue of borderline cases is crucial in any classification scheme; having too many such (indistinct) cases undermines the validity and utility of the scheme. The most frequent and fundamental measures, then, are the event count and event frequency, simply the number of occurrences of a particular type of event and events per unit time.

Having identified the object of analytic interest as a particular event or case, the researcher can proceed to measure and code its various parameters. The parameters of social phenomena are of three general types: attributes, qualities, and dynamics. Attributes are one-dimensional, defining properties, such as beginning and ending dates, duration, number of participants, number of fatalities, actor, target, and location. Qualities refer to general, multidimensional characteristics of social units that represent (reasonably) stable and comparable patterns of complex normative, behavioral, instrumental, or institutional factors. Each social phenomenon is in large part a unique mixture of complex traits that combine to produce a distinctive and recognizable quality. For example, personal dictatorships, oligarchies, military juntas, hereditary monarchs, and one-party states are different forms of government but share a certain quality of autocratic rule. An important facet of the complexities of social phenomena are its dynamic qualities, that is, changes over the course of the phenomena in the magnitude, intensity, scope, etc. of its several definitive attributes and qualities. For example, violent political conflict situations often fluctuate in levels of commitment among three main interactive strategies: conventional (protest and negotiation), unconventional (militancy and armed conflict), and withdrawal (noninteraction). Of particular importance in violent conflict events are escalatory and deescalatory dynamics. Each of the three general types of parametric measures: attributes, qualities, and dynamics, provide important information for the quantitative analysis of social phenomena typologies.

Earlier, it was noted that measurement and analysis are best considered an interactive and reiterative sequence in regard to the study of social phenomena, rather than separate phases of the research process. Measurement and analysis inform one another and advances in either one may lead to refinements in both aspects. Ted Robert Gurr provides a succinct procedural "map" of the circular research process in his 1972 book, *Politimetrics: An Introduction to Quantitative Macropolitics*. In that scheme, an initial "hunch or observation" stimulates the articulation of (1) a "theory" from which "testable hypotheses" may be derived. In order to test, or validate, the theory, the researcher engages in (2) "problemation," that is, identification of "the most fundamental problem requiring solution if a progressive development of theory about a subject is to occur." Having decided upon the focus of inquiry, the researcher proceeds to (3) "variable or unit specification." Specification is the essential act of measuring social phenomena; specification must be definitive, explicit, and sufficiently detailed that any researcher will understand the variable or unit in the same way. Directly related to specification is (4) "operationalization," wherein coding rules and procedures are designed whereby "reliable and valid" measures and indicators can be obtained that are in accordance with the variable or unit specification; these rules and procedures must be so precise that any independent researcher will obtain the same values when applying the coding rules to specific cases. Data collection proceeds by (5) applying the coding rules systematically over the entire "universe of analysis." Subsequent (6) analysis and (7) interpretation of results, then, further informs, and refines, the research process. The quality of measurement itself is judged according to the standards of validity (is it an adequate measure of what it purports to represent?); accuracy (is it precise enough to allow detailed or subtle distinctions among cases?); and reliability (is it comparable among cases and does it yield consistent results in successive measurements of the same case?).

In summary, the process of classification, estimation, and coding of information regarding inherently complex social phenomena such as political conflict is a theory-driven procedure that produces numerical codes and indices. This codified data is recorded and stored as "representational" numbers; representational as the relationships among the numbers used are not necessarily numeric. Social science data, unlike data in the physical sciences, cannot be precisely measured under controlled conditions. As such, they rarely approximate ratio, or even ordinal, numeric scales and, so, great care must be taken when using such "soft" and "fuzzy" data in quantitative (especially statistical) analyses. The strength of analysis and interpretation of social science data depends on the quality of the measures used and this quality is established through proper coding procedures and verified through a painstaking confidence-building process. Confidence is established and increased through extensive analysis, that is, by establishing patterns of consistency in the application of techniques, the cross-examination of multiple parallel analyses, and the substitution of alternative indicators.

Measuring Political Conflict

The state is a crucial actor in all political conflict processes, both with actors in the external (interstate) environment and actors in the internal (intrastate) environment. Historically, the state system has dominated political processes and our capabilities to measure and analyze those processes since the Peace of Westphalia in 1648. The interests of state (or national) security are critically affected by political conflicts and, so, the state, as the primary collector and recorder of social science data has been in position to actively suppress or distort information regarding political conflict behavior, especially as political conflict often poses a direct challenge to the viability of the state itself. The suppression of information in the interests of state security is most effective in regard to internal conflicts, as the sovereign state has long held a virtual monopoly over information on its internal affairs. Of course, the complex motives and incentives to suppress or distort reports concerning serious political conflict episodes affect all parties that are directly involved or have important stakes in the outcome. Disinformation has strong security and strategic value for all parties involved, often leading to wide discrepancies in key information. Great caution is required when measuring political conflict; multiple reports should be acquired whenever possible and sources should be evaluated for reliability.

Nearly all political conflicts in international affairs are channeled through and controlled by the institutions of states; thus, the concentration of power in the state system at once greatly reduces the number of potential conflict actors and imposes discipline on the course of conflict interactions through international norms and law (despite the popular fiction of an anarchic world system). Political conflict between states enjoys greater visibility than that within states as (1) the concentration of power in states greatly increases the conflict potential between states, thus, commanding greater attention, and (2) there are in all cases at least two independent actors and, therefore, two independent seats of information gathering and dissemination. In addition, conflict between states is more institutionalized and public and, so, more likely to be observed and recorded by private parties and representatives from disinterested states. Political conflict within states, in addition to being overshadowed by the security interests of the state, is largely ad hoc, undisciplined, complex, and diffuse. For all these reasons, the systematic measurement and analysis of external political conflict has progressed more rapidly than has the measurement and analysis of internal political conflict.

Quantification in studies of the most extreme (violent) forms of political conflict is a relatively recent addition to the social sciences. The systematic quantification of the problem of interstate war was greatly aided by the high profile of the institution of war in state politics, human fascination with the spectacle and horror of war, and the preeminent place of war in the historical record. The pioneering work in the quantification of classical wars is Quincy Wright's 1942 work titled *A Study of War*. In that study, Wright codified "all hostilities involving members of the family of nations [independent states] ... which were recognized as states of war in the legal sense or which involved over 50,000 troops." Wright's very narrow, statist treatment of the problem of political conflict as a legal condition of war between sovereign states fit neatly within the conventional "state security" perspective. Wright's threshold for identifying only the highest profile cases of political conflict ensured that his collection of cases would be comprehensive given the limitations on information. An alternative perspective was offered in the work of Lewis Richardson in his 1960 study titled *Statistics of Deadly Quarrels*, which purported to include all political events that involved a "quarrel" (i.e., a hostile dispute) and at least one fatality. Lewis' very low threshold for identifying cases guaranteed failure as such detailed information was not generally available at that time. However, Lewis' broad approach to the measurement of violent political conflict in many ways foreshadows the emergence of the "human security" perspective in the late 20th century.

A major point of difference between the approaches taken by Wright and Richardson concerns their conceptions of the "most fundamental problem requiring solution," that is, problematic. Both accept that the transformation of political conflict to the systematic use of violence is the core of the problem. However, Wright takes the conventional approach in assuming that distinct forms of political violence events, or events occurring at different "levels of analysis" (i.e., individual, state, and system levels), can be identified and categorized and that the within category causal relationships are fundamentally similar, whereas the causal relationships across categories are fundamentally different. The Lewis approach, on the other hand, assumes that all forms of political violence share essentially similar causal relationships and that the different forms that political violence appears to take are largely a product of the circumstances within which a particular political conflict develops. A corollary to the problematic question, regardless of the approach taken, concerns the "point or points" at which nonproblematic political conflict transforms to problematic political conflict, that is, the "process."

Interstate (External) Political Conflict

Interstate War and Militarized Disputes

Whereas Wright stood as the pioneer in the identification and measurement of war, J. David Singer emerged as its

paragon. Building on Wright's work and earlier work by Pitirim Sorokin, Singer, and his partner Melvin Small, established the Correlates of War (COW) project in the mid-1960s. The Correlates of War created a comprehensive database recording the basic attributes (inclusive dates, duration, general location, state participants, general outcome, and estimated number of "battle deaths") of each interstate war case beginning with the end of the Napoleonic wars in 1816. In a separate effort, Jack S. Levy compiled and reported basic information on war cases beginning in 1495 in his 1983 book, *War in the Great Power System*.

The standard measure of war magnitude adopted by the COW project: "battle deaths," still stands as the industry standard, however, the COW project's high threshold of 1000 annual average battle related deaths has been relaxed somewhat by subsequent researchers. The Uppsala armed conflicts data set, directed by Peter Wallensteen, provides a compilation of war events (including both interstate and civil wars) over the contemporary period beginning in 1946 that uses a 25 battle-death per annum threshold. Some argue that this threshold is too low to distinguish war events from other forms of political violence. The Uppsala war data also categorizes war episodes according to general levels of magnitude: minor (>25 battle deaths), intermediate (>100), and major (>1000). A common criticism of the strict "battle death" measure is that it does not take into account the often substantial noncombatant or civilian casualties of war, that is, the "battle-related deaths." Ruth Leger Sivard was one of the first to estimate the numbers of battle-related deaths in war events (interstate and civil) over the post-World War II period.

Complicating the issue of measuring the magnitude of war is the fact that the actual numbers of deaths in many wars, whether battle or battle-related, remain unknown; most such tabulations are only crude estimates. This is particularly true in regard to civil wars. Information that could be used to create other, more detailed, measures of war magnitude, such as injuries, damage caused, area affected, or costs, are not available, or even estimable, for most wars. Monty Marshall argues that violent political conflict is the social equivalent to "storms" that ravage the societal ecosystem with an identifiable potential that produces chaotic effects. Borrowing from meteorology, he developed a 10-point scale of the comprehensive "societal impact of war" taking into account the full spectrum of war's destructive effects on complex social systems, including general damage to human resources, population dislocations, weakening of social networks, deterioration of environmental qualities, infrastructure damage and resource diversions, diminished qualities of life, and increased nonreciprocal resource transfers.

In addition to refining Wright's original list of war cases, the COW project expanded the universe of inquiry to include a new category of war termed "extra-systemic war." Whereas interstate war occurs between two or more independent states, in an extra-systemic war an independent state "engages in a war with a political entity that is not an interstate system member;" these wars are designated as one of two types: imperial or colonial. The COW project also expanded its data collection to include all interstate military conflict events, irrespective of the occurrence of fatalities, or battles. The category of "militarized interstate disputes" includes situations of political conflict "between sovereign states below the threshold of war and include explicit threats to use force, a display of force, a mobilization of force, or the use of force short of war." It has also collected data on the conflict-related attributes of states: "national material capabilities" (power) and cultural attributes (ethnic and religious subnational groups), and attributes of the state system: system membership, formal alliances, borders and contiguity, and territorial change.

Interventions

Whereas the conceptualization of "interstate war" depends in large part on what Wright has referred to as the "legal condition of war," whether that sense of legality is explicit or implicit, the use of military force by a third party to intervene in or interfere with the course of an interstate or civil war in which it is not directly involved generally lacks this legal sense and is usually regarded an "intervention." "Military interventions" are military operations intended to alter the course or outcome of an ongoing war in favor of the interests of the intervening party; multilateral military interventions are conducted to promote common values such as the enforcement of legal principles. "Humanitarian interventions" are military operations intended to alter the course or outcome of a war in the general interests of alleviating human suffering or limiting the war's brutality and destructiveness.

There are three independent data collection efforts that focus on interventions; each of the three collections cover the contemporary period (i.e., since the end of World War II). Herbert Tillema has compiled information on contemporary cases of "foreign overt interventions," which involve any use of military force by one country outside its own borders. Tillema's concept of intervention is similar to the COW concept of "militarized interstate dispute." Frederic Pearson and Robert Baumann have collected information on cases of "international military interventions" that involve a use of military force to intervene in a civil conflict in a foreign land. Patrick Regan has compiled cases of "third party interventions" based on a much broader definition of what constitutes an intervention; his data includes cases of military, economic, and diplomatic interventions. Related to this broader definition of intervention are data collections on issues such as "arms transfers,"

"peacekeeping operations," "conflict mediation," and "foreign support."

Foreign Policy Crises

Similar to the COW project conceptualization of "militarized interstate dispute" is the concept of "foreign policy crisis" formulated by the International Crisis Behavior (ICB) project established in the mid-1970s by Michael Brecher and Jonathan Wilkenfeld. For the ICB project, a "crisis" must involve at least one state and is initiated by "a specific act, event or situational change which leads decision-makers to perceive a threat to basic values, time pressure for response and heightened probability of involvement in military hostilities." The ICB data covers the period beginning in 1918 through the present.

International Interactions

Official political interactions between the governments of countries are very formal and stylized daily "events" that are used to communicate, or "signal," various levels of conflict and cooperation. Signaling events may range from acknowledgment of subordination (e.g., surrender or acquiesce); through denials, criticisms, complaints, or protests; to more hostile demands, warnings, or threats; and even ultimatums, boycotts, seizures, and attacks. Two early compilations of "events data" are Edward Azar's Conflict and Peace Data Bank (COPDAB) and Charles McClelland's World Event Interaction Survey (WEIS). COPDAB data coverage begins in 1948 and was originally developed to study the interactive dynamics associated with the problem of "protracted social conflict;" as such, it includes both state-to-state events and interactions between states and substate groups. The WEIS data coverage begins in 1964 and focuses on official political interaction events between states as reported in the *New York Times*. Both projects attempted to develop conflict scaling techniques by which they could use events data to measure short-term changes in levels of political conflict. These global events data records have been compiled from news reports and involve tens of thousands of events per year. Events data collection has been extremely labor-intensive and costly. Recent advances in the conversion of textual news reports to electronic files has contributed to the development of machine (computer) coding techniques that may greatly reduce the costs of recording, compiling, and archiving events data.

International Terrorism

International, or transnational, terrorism refers to a very special form of violent political conflict event. The conventional conceptualization of international terrorism refers to a violent act by a member or members of a substate political group subject to one state's jurisdiction against a target associated with or under the nominal protection of another established state or suprastate zauthority. There are two data bases that have collected information on international terrorism events. The RAND–St. Andrews Chronology of International Terrorist Incidents begins its coverage in 1968. The International Terrorism: Attributes of Terrorist Events (ITERATE) data base originally compiled by Edward Mickolus also begins coverage in 1968.

Domestic (Internal or Civil) Political Conflict

The human fascination with the external subjects of war and conquest has been chronicled since the beginnings of civilization. The capacity to deter attacks and to wage and win wars (power) was even heralded as the true measure of the "good" state through much of human history. The ability to ensure domestic social order was viewed as a requisite for state power and status. The Hobbesian notion of inherent social disorder and the Machiavellian principle of state preeminence in politics combined to forestall critical examination of the modes and methods of state authority and the dynamics of internal political affairs. Greater attention to the qualities of citizens' rights, that began with John Locke's reflections on governance and gained impetus with the movement for the abolition of slavery, became imperative with the socialist and fascist convulsions that marked the 20th century.

Whereas the evolved formalities of the state system created a fairly disciplined structure based on relative power capabilities and imposed a fair degree of order and conceptual simplicity on the conflict behavior of states, the intricacies of domestic political conflict appeared especially chaotic and remained poorly understood. Modernity facilitated and energized the mass mobilization of latent constituencies that increasingly challenged the political status quo. The ground-breaking studies of "social revolutions" conducted by Crane Brinton (*The Anatomy of Revolution*) and Barrington Moore Jr. (*Social Origins of Dictatorship and Democracy*) and Franz Fanon's expose of the perversions of colonialism (*The Wretched of the Earth*) in the mid 1960s were among the first to critically examine the complex relationships between governments and the governed.

Civil War and Revolution

While the radical, Marxist ideal of social revolution piqued general interest in the dynamics of political conflict, it was the noted similarity of civil warfare to interstate warfare and the rise of "wars of independence" during the decolonization period following the end of the Second World War that first informed the systematic collection and study of domestic political conflict episodes. The COW project expanded its treatment of major wars by adding a third category, "civil war," to complement its

collection of data on interstate and extra-systemic wars over the period since 1816. In the COW project, a major civil war is defined according to four criteria: "(a) military action was involved; (b) the national government at the time was actively involved; (c) effective resistance (as measured by the ratio of fatalities of the weaker to the stronger forces) occurred on both sides; and (d) at least 1000 battle deaths resulted during the civil war." Greater attention has been drawn to the subject of civil wars as they became increasingly prevalent in the latter half of the 20th century, in contrast to the outbreak of interstate wars, which remained quite rare. Most alternative compilations of war events accept the basic COW classification scheme.

More recently, a new compilation of domestic warfare events during the period beginning in 1955 was produced for the US Government's State Failure Task Force. The list of "state failure" events includes three categories of civil violence: (1) ethnic wars, (2) revolutionary wars, and (3) genocides (ethnic mass murder) and politicides (political mass murder). Each war event is scored annually according to three measures of magnitude: number of rebel combatants, number of battle-related deaths, and size of territory directly affected by the war (genocide and politicide events are scored annually only for number of deaths). The State Failure data set is unique also because it combines information on a fourth category, "adverse regime change" events (that is, major changes toward greater autocracy), with domestic violence events and combines time-related events into complex "state failure" events.

Genocide and Human Rights Violations

An alternative perspective on civil violence, made particularly salient by the Nazi Holocaust during World War II, focuses on the state's tremendous power advantages over its own citizens. This "human rights" perspective looks at the problem of "state terror" or the systematic use of terror and violence by state authorities to subdue actual and potential opposition to its authority by individuals and substate groups. The Purdue Political Terror Scale is a five-point scale designed by Michael Stohl that provides annual scores, beginning in 1980, for the general quality of each state's treatment of its own citizens. Other researchers have compiled information on cases of political mass murder, such as Rudolph Rummel's data on cases of "democide" (mass murder by governments) and Barbara Harff's cases of genocide and politicide (incorporated in the State Failure data set mentioned above).

Governance

The crucial link between the general qualities of governance and domestic political conflict is recognized and incorporated in the Polity data series, originally designed and compiled in the early 1970s by Ted Robert Gurr. Government is charged with the primary responsibility for managing political conflict and preventing outbreaks of violent contention between political groups. The Polity scheme examines the "patterns of authority" that characterize different regime types. It combines information on the institutional qualities of executive recruitment, constraints on executive power, and the general tenor of political competition to score both the autocratic and democratic qualities of regimes, beginning in 1800.

Conflict Events

There are two broad data collection and archiving research enterprises that have been engaged in compiling information regarding daily domestic political conflict events. The Cross-National Time-Series Data Archive (formerly named the Cross-Polity Survey) was established in 1968 by Arthur S. Banks. The Banks' data is derived from the *New York Times* daily news files, begins coverage in 1815, and records annual numbers of events in nine categories of domestic conflict, including guerrilla warfare, government crises, purges, riots, revolutions, antigovernment demonstrations, coups, assassinations, and general strikes. The World Handbook of Political and Social Indicators was begun under the direction of Charles Lewis Taylor in the late 1960s; its data begins coverage in 1948. The World Handbook also uses the *New York Times* as its primary source but supplements that general news coverage with six regional news sources; it has compiled event counts on the same general types of domestic political conflict events as the Banks' data but with finer distinctions such that information on up to 38 separate event types are recorded. Both conflict events data collection efforts suffer from problems associated with human processing of large volumes of information on daily occurrences. Neither of these data projects distinguishes among subnational groups or interests actively engaged in the conflict events or levels of magnitude; they simply report raw event counts. Current plans call for updating the World Handbook using machine-coding techniques which, if proven effective, will greatly reduce the time and cost of recording events data.

Minorities and Ethnic Conflict

Perhaps the most ambitious effort to date at "unpacking" the state and recording detailed information on substate social groups and their conflict behavior has been the Minorities at Risk (MAR) project established in the late 1980s by Ted Robert Gurr. The MAR project is the first conflict data base that uses the "ethno-political group," rather than the "nation-state," as its unit of analysis; data coverage begins in 1946. The 2002 version of the MAR data set provides a comprehensive listing of 285 current and 62 historical "minorities at risk" around the world. A "minority at risk" is defined as follows: "The group collectively suffers, or benefits from, systematic discriminatory

treatment vis-à-vis other groups in a society and the group is the basis for political mobilization and collective action in defense or promotion of its self-defined interests." The general categories of information coded for each group include group identity characteristics, discrimination, group organization, collective interests (grievances), sources of transnational support, and conflict behavior (intragroup factionalism, intergroup communal conflict, protest, rebellion, and government repression).

Public Opinion

Political conflict stems from the perspectives, interests, and values of individuals and, so, gauging differences, strengths, and changes in public opinion can provide crucial information on potential and actual responses to policy issues and initiatives. Public opinion surveys have become a routine instrument for gauging public response in open societies. Comparative, cross-national surveys of public opinion are more problematic as language and cultural differences among societies complicate the survey design and interpretation of results. The first regular, multinational opinion surveys were begun under the aegis of the European Community in 1970. The Eurobarometer surveys are conducted twice annually and were expanded in 1990 to cover central and eastern Europe. Drawing upon the experience of the Eurobarometer and recognizing the need to factor in information regarding differences in cultural values, the European Values Survey was begun in 1981. Building on that experiment, Ronald Inglehart expanded coverage of the values survey to include countries representative of all major cultural regions. The World Values Survey has collected information from 65 societies around the world. More recently and in response to the increasing globalization of political issues, the Pew Research Center initiated, in 2000, the Pew Global Attitudes Project, which samples public opinion in 49 countries around the world.

State and Human Security

The "behavioral revolution" that began in earnest in the 1960s sought to apply quantitative methodologies, adapted from the physical sciences, to augment the empirical quality of the social sciences and, in particular, the measurement and study of political conflict. While a significant rift remains between the more "purist" advocates of qualitative and quantitative methods, it is clear that substantial progress has been made in the development and application of macrocomparative and statistical techniques in political science and that these procedures have transformed the field of inquiry in very fundamental ways. The evolution of social science methodologies has benefited especially from major advances in information, computation, and communication technologies, and particularly from the global processes of greater openness and democratization that have encompassed the world at the beginning of the "next" millennium. Having emerged from a strict "state-centric" political culture, the new methodologies have contributed to a shift toward greater attention to individual human rights and the articulation of a "human security" perspective that both challenges and complements the conventional "state security" perspective.

At the beginning of the 21st century, it is clear that the simplicity and discipline of the classic Westphalian "state system" is giving way to a more complex and integrated "global system." Tantamount to this change is a virtual explosion in the numbers and types of international organizations (from about 500 in the 1950s to over 16,000 today) and the volume and densities of international transactions, and with these, also, the potential for political conflict. Our ability to effectively manage political conflicts depends critically on our abilities to measure and monitor the attributes, qualities, and dynamics of those conflicts.

See Also the Following Articles

Experiments, Political Science • Political Science • Political Violence

Further Reading

Azar, E. E., Jareidini, P., and McLaurin, R. (1978). Protracted social conflict: Theory as practice in the Middle East. *J. Palestinian Stud.* **8,** 41–60.

Bremer, S. A., and Cusack, T. R. (eds.) (1995). *The Process of War: Advancing the Scientific Study of War*. Gordon and Breach, Luxembourg.

Cioffi-Revilla, C. (1990). *The Scientific Measurement of International Conflict: Handbook of Datasets on Crises and Wars, 1495–1988.* Lynne Rienner, Boulder, CO.

Coser, L. A. (1956). *The Functions of Social Conflict.* Free Press, New York.

Gleditsch, N. P., Wallensteen, P., Eriksson, M., Sollenberg, M., and Strand, H. (2002). Armed conflict, 1946–2001: A new dataset. *J. Peace Res.* **39,** 615–637.

Gurr, T. R. (1972). *Politimetrics: An Introduction to Quantitative Macropolitics.* Prentice-Hall, Englewood Cliffs, NJ.

Gurr, T. R. (1993). *Minorities at Risk: A Global View of Ethnopolitical Conflicts.* United States Institute of Peace Press, Washington, DC.

Marshall, M. G. (1999). *Third World War: System, Process, and Conflict Dynamics.* Rowman & Littlefield, Boulder, CO.

Marshall, M. G. (2002). Measuring the societal impact of war. *From Reaction to Conflict Prevention: Opportunities for the UN System* (F. O. Hampson and D. M. Malone, eds.). Lynne Rienner, Boulder, CO.

Richardson, L. F. (1960). *Statistics of Deadly Quarrels.* Boxwood Press, Pittsburgh.

Sarkees, M. R.Wayman, F. W.Singer, J. D. (2003). Inter-state, intra-state, and extra-state wars: A comprehensive look at their distribution over time, 1816–1997. *Int. Stud. Quart.* **47,** 49–70.

Singer, J. D. (1968). The incompleat theorist: Insight without evidence. *Contending Approaches to International Politics* (K. Knorr and J. N. Rosenau, eds.). Princeton University Press, Princeton, NJ.

Small, M., and Singer, J. D. (1982). *Resort to Arms: International and Civil Wars, 1816–1980.* Sage, Thousand Oaks, CA.

Vasquez, J. A. (1993). *The War Puzzle.* Cambridge University Press, Cambridge, UK.

Wright, Q. (1942/1965). *A Study of War,* revised Ed. University of Chicago Press, Chicago.

Political Science

John L. Korey
California State Polytechnic University, Pomona, California, USA

Glossary

event data analysis The coding and content analysis of interactions between nations (or other actors) reported by newspapers, wire services, or other sources.
JudgeIt A program for analyzing the impact on election outcomes of existing or proposed legislative districting plans.
nominate A method of analyzing legislators' ideological preferences by locating them in *n*-dimensional space based on their roll call votes.
paradox of voting (Arrow's impossibility theorem) The rule that given more than two alternatives and absent the assumption that preferences are single peaked, no method of voting will ensure that there will be majority support for any one alternative over each of the others.
prisoners' dilemma A non-zero-sum game in which each player's pursuit of rational self-interest leads to nonoptimal overall results.
Samplemiser A Web-based program for filtering out noise variance in a series of cross-sectional surveys.

Almost since political science became recognized as a separate academic discipline in the late 19th century (the American Political Science Association was not founded until 1903), its practitioners have struggled over its identity. By the 1960s, its place as a field that is at once both a social science and humanity finally had become generally accepted, although sometimes grudgingly. Today, political science as a social science draws on methods shared with the other social sciences and has developed a few of its own. Following a brief history of empirical research in the discipline, this article provides some illustrations of methods that have been used in political science to gather data (through experiments and quasi-experiments, survey research, case studies, and content analysis), to develop formal models of political behavior, and to carry out statistical analysis of data.

History

In their history of American political science to 1980, Somit and Tanenhaus dated "the first rigorous application of statistics to political data for analytic purposes" to the 1898 work of A. Lawrence Lowell. By charting election results in the United States and Great Britain, Lowell showed that "oscillations" tended for a variety of reasons to work against the party in power, thereby producing equilibrium in two-party systems.

However, it was not until the 1920s, with the emergence of the Chicago school led by Charles E. Merriam, that a systematic effort was made to create a "new science" of politics focused on the development and testing of empirical hypotheses. In an influential 1921 article, Merriam called for organized efforts at systematic data collection, more and better use of statistical techniques, and more extensive borrowing from other social science disciplines, especially sociology and social psychology.

This approach flourished for a time, but it did not really begin to reach maturity until the end of World War II. In the years that followed, what would become known as the behavioral approach to the study of politics clearly became part of the mainstream of the discipline. In 1962, 21 universities formed the Inter-University Consortium for Political Research in order to archive data and provide advanced quantitative training. (In 1975, the organization became the Inter-University Consortium for Political and Social Research. Today, it is a worldwide entity with more than 500 member institutions.) In the mid-1960s, the National Science Foundation recognized political science as a social science. With this, as Somit and Tanenhaus

observed, "the last bastion of resistance to the legitimacy of the behavioral movement had fallen."

Success produced its own backlash. For a time, a pitched battle was fought between behavioralists and traditionalists over the definition and direction of the discipline. Increasingly, however, it came to be accepted that a fully adequate study of politics would have to embrace a variety of ways of knowing. As early as 1961, Dahl argued the need for greater dialogue between behavioralists and exponents of other approaches, including political philosophy. Before the end of the decade, Easton suggested that political science was already moving into a "postbehavioral" period in which social science methods would continue to be employed but with more concern for their relevance to the pressing issues of the day. Since then, social science methods have been a major, although not an exclusive, focus of scholarship in the field.

Table I presents evidence of this. The table classifies articles from the 2001 issues of arguably the three leading journals in the discipline, the *American Political Science Review* (APSR), the *American Journal of Political Science*, and the *Journal of Politics*. Overall, quantitative analysis of empirical data is found in three-fourths of all entries. Some of these articles developed and tested formal models, but most proceeded more inductively. Similar, earlier classifications distinguish between articles with only low-level quantitative analysis (such as percentages and means) and those with more statistically complex analysis. No such distinction is made here since, with a few exceptions, all of the articles in this category of Table I employed relatively complex methods. The second category in the table includes articles containing analytical (predominantly mathematical) models and frameworks and discussions of methodological issues but no quantitative analysis of empirical data, although some did include simulated data and two involved qualitative empirical analysis. Except for these last two, critics might dismiss articles in this category as "inference without evidence." However, they were clearly intended to develop empirically testable social science hypotheses. Finally, only approximately one in seven articles are classified as humanities-oriented analysis (mostly political philosophy, with one article each in public law and literary analysis), and even a couple of these included some qualitative empirical analysis. In 2001 at least, the discipline's flagship journal, the APSR, was somewhat more hospitable to humanities-oriented research than the other two ($p < 0.01$ for the overall table).

Obviously, debates about the direction of the discipline continue. Echoing Easton, critics (such as backers of the Perestroika movement of the last few years) fault behavioral research for elevating methodological rigor over relevance to pressing social and political problems. However, critics of behavioral research are more likely now to call themselves postmodernists than traditionalists.

Experiments and Quasi-experiments

Political science is, of course, a largely nonexperimental discipline. Exceptions to this rule, however, have a long pedigree. Experiments and quasi-experiments (involving manipulation of one or more independent variable by the researcher but lacking one or more of the other elements of a true experimental design, including random assignment of subjects to groups) have been carried out in both field and laboratory environments.

Field Research

Perhaps the earliest example of the use of experimental design in political science is described in Gosnell's 1927 report of his efforts to discover methods of improving levels of voter turnout. Gosnell selected 12 census enumeration districts in Chicago intended to reflect a cross section of the city's demographics. Within each district, households were divided (although not on a truly random basis) into control and experimental groups. Before the 1924 presidential election, those in the latter group were

Table I Articles in Leading Political Science Journals, 2001[a]

	American Political Science Review (%)	*American Journal of Political Science* (%)	*Journal of Politics* (%)	*All*
Quantitative, empirical data	55	82	85	74
Analytical theory; no quantitative, empirical data	20	9	6	12
Humanities-oriented analysis	25	9	8	14
Total (%)	100	100	100	100
N	51	57	48	156

[a] Replies to other articles were counted as separate entries; review articles, book reviews, and the address of the president of the American Political Science Association were excluded. Totals may not agree due to rounding.

sent mailers urging them to register. (Chicago did not employ a system of permanent voter registration.) Those in this group who did register then received material urging them to vote.

Gosnell found that there was a substantial difference in registration rates between the two groups. Mailers urging registered voters to turn out had little additional impact since turnout among registered voters was very high even without additional stimulus. A follow-up study of the much lower turnout local elections of 1925, however, did show an increase in turnout in the experimental group. Gosnell concluded that greater efforts at civic education would "undoubtedly have an immediate and continuous effect upon the interest shown in elections."

The line of research pioneered by Gosnell has been continued, albeit sporadically, to the present. A 2000 article by Gerber and Green is clearly in the Gosnell tradition. Gerber and Green randomly assigned registered voters in New Haven, Connecticut, either to a control group or to experimental groups in which subjects received one or more stimulus: face-to-face contact, direct mail, and telephone calls. They found that face-to-face contact substantially improved voter turnout, that direct mail produced a modest increase, and that telephone calls had no impact. They concluded that declines in recent decades in personal contact with voters by parties and other organizations help explain "the ongoing mystery of why turnout has declined even as the average age and education of the population has risen."

In the sophistication of its design and methods of analysis, the Gerber and Green study is, as would be expected, worlds apart from Gosnell's work. Some things, however, change very little. A postcard sent by Gosnell included a cartoon showing the ghost of a disheveled-looking man labeled "The slacker who won't help defend his country in time of war" extending his hand and saying "Hello, brother" to a much more reputable-looking man labeled "The slacker who doesn't vote when his state needs his help." A postcard sent by Gerber and Green showed a photo of the Iwo Jima Memorial with the caption, "They fought ... so that we could have something to vote for." Although obviously different in tone, both communications employed the same underlying appeal to voting as an act of patriotism analogous to defending one's country in time of war.

Although often contrasted with experimental research as a method, survey research commonly employs experiments as a means of improving survey design. In their pioneering study of voting in the 1940 election, researchers at Columbia University divided their sample into an experimental group and three control groups in part to test for reactivity that reinterviewing respondents in a panel study might produce. Survey researchers also very commonly conduct experiments to develop and test measurements. Sniderman and Carmines, for example, randomly assigned respondents to two or more groups to assess the effect of question wording and question order on attitudes toward a series of race-related issues. They concluded that such issues are often framed in ways that inadvertently underestimate the potential for building interracial political coalitions.

Laboratory Research

Barber's 1966 study of 12 local government finance boards is an early example of laboratory research in political science. Barber invited these boards to participate in mock budget meetings conducted in a social psychology lab at Yale University. He argued that use of actual boards would provide more realistic results than those usually produced by small group experiments. Barber acknowledged that his research was not a true experiment in that he used actual rather than randomly assigned groups and in that all 12 groups were assigned the same tasks. The one variable that was manipulated was the presence or absence of the board's presiding officer, who in each case was called out of the lab midway through the session. Barber found that the absence of an active chair led to an increase in negative communications, whereas the absence of a chair with a passive style led to a decrease in conflict.

In 2002, Green and Gerber noted that "recent years have witnessed a resurgence of interest in laboratory experiments" in political science. The precision of experimental designs often makes them especially well suited for the testing of formal mathematical models. Another area that has lent itself more than most to laboratory analysis has been the examination of the impact of campaign commercials. Indeed, Iyengar has remarked that such studies "now constitute a 'dominant' methodology for political communication researchers." An example of such research was described by Ansolabehere *et al.* During the course of several actual political campaigns, subjects viewed 30-second simulated ads similar to those used in the campaigns. The visual components of the ads were held constant, and subjects were randomly assigned to groups exposed to versions of the ads that differed in the negativity of the announcer's text. The research showed that attack ads tend to have a demobilizing effect—that is, reduce the subject's intent to vote.

In general, however, experimental and quasi-experimental designs continue to represent a relatively small portion of political science research. The journal *Experimental Study of Politics* was founded in 1971 but stopped publishing on a regular basis a decade later. Green and Gerber concluded that this relative neglect is unfortunate since "well-conducted experiments are typically more persuasive arbiters of causality than comparable nonexperimental research."

Survey Research

In their 1924 groundbreaking study of the 1923 mayoral election in Chicago, "Non-Voting: Causes and Methods of Control," Merriam and Gosnell employed a variety of methods of gathering relevant information, including a survey in 1923 of more than 5000 Chicago residents. Although not employing random sampling techniques, the study was intended to be representative in terms of "sex, age, nationality, economic status, occupation, [and] length of residence." For almost the next two decades, most of the advances in the study of mass political opinion were made by the Gallup Organization and other commercial polling firms.

In 1940, scholars at Columbia University's Bureau of Applied Social Research carried out a systematic random sample of residents of Erie County, Ohio, that became the basis for "The People's Choice" by Lazarsfeld *et al.* The survey consisted of four panels, three of whose members were interviewed on two occasions and a fourth whose members were interviewed seven times during a period of 6 months. Four years later, the Bureau of Applied Social Research joined forces with the National Opinion Research Center (located at that time at the University of Denver) to conduct a nationwide survey. This survey was a two-wave panel consisting of pre- and postelection interviews.

These works are important landmarks in the use of survey research to study political attitudes and behavior. The early study with the greatest long-term impact, however, was a small ($N = 662$) national survey in 1948 conducted by the University of Michigan's Survey Research Center. This would be the first of what would become known as the American National Election Studies (NES). NES surveys have been conducted in each presidential election year since 1948 and in each midterm congressional election year starting in 1954. The cumulative file for 1948–2000 contains data from 44,715 interviews. Since 1970, the University of Michigan's Center for Political Studies has been in charge of the project. In 1978, the NES became a "national research resource" supported by the National Science Foundation (NSF). Along with two other ongoing NSF-funded efforts, the General Social Survey and the Panel Study of Income Dynamics, the NES project is a leading example of what Sapiro called "big social science."

Some sense of the impact of the NES on scholarship can be obtained by an examination of the bibliography that the NES maintains on its Web site of works using its data. The list (of books, chapters, journal articles, conference papers, internal NES reports, and a few entries in the popular media) comprises more than 3600 entries. Figure 1 (which excludes undated entries, internal NES reports, and a few duplicate entries) shows 5-year moving averages for the number of entries in the bibliography by year of publication from 1952 through 2001. It is possible that the bibliography is less complete for the earlier years. Conversely, delay in adding entries to the list no doubt accounts for what appears to be some drop-off near the end of the period. At the very least, however, the chart shows that since the 1970s, the NES has had a large and sustained impact on the conduct and dissemination of scholarship.

Survey research is hardly confined to the study of American opinion, and there are many examples of applications in other countries, including the European Commission's "big science" Eurobarometer, which has been measuring attitudes since 1974 (with forerunners

Figure 1 National Election Studies bibliography: Citations by year. Compiled from National Election Studies (2002). The NES bibliography. Available at http://www.umich.edu (accessed August 7, 2002).

of the Eurobarometer dating from 1962). Other important multinational and interdisciplinary survey efforts include the several waves of the World Values Survey and, in political science specifically, the Almond and Verba five-nation study. An important recent initiative specifically focused on the study of politics is the Comparative Study of Electoral Systems (CSES). Like the NES, the CSES is funded by the NSF. It includes both postelection surveys and aggregate data and will eventually expand to approximately 50 countries. The survey research portion of the project consists of postelection interviews including 16 questions intended to be completed in approximately 10 minutes. Because of varying election calendars, the first module of the study was designed to cover the period 1996–2000. Data are available for 19 countries for elections through 1998. A total of 32,022 cases are included in the pooled sample. A second module is planned for the 2001–2004 cycle. References to the study have begun to appear in the literature: The CSES Web site listed (as of August 7, 2002) 46 such references (from passing mentions to analyses based on the data) in its bibliography.

Case Studies

Long a staple of political science research, case studies have the advantage of examining politics holistically. Sometimes this comes at the expense of the ability to generalize findings by placing them in a rigorous theoretical context. That such is not an inherent limitation of the case study method can be seen in two very different applications of the method, one in which individual actors provide the cases for analysis and the other in which the cases consist of entire political systems.

Monroe questioned why some people act in ways that are, to varying degrees, at odds with their self-interest. Through interviews, she examined 25 individuals whose behavior fell at different points on a continuum from self-interest to altruism (the entrepreneur, the philanthropist, the hero, and the rescuer of Jews during the Holocaust). Her work explicitly challenged rational choice theories explaining behavior wholly in terms of self-interest (although sometimes defining self-interest so broadly as to make the argument tautological).

Monroe concluded that various hypotheses from sociology, economics, biology, and psychology only very imperfectly serve to distinguish the altruist, especially the hero and the rescuer, from the entrepreneur. She found that what does seem to be the distinguishing characteristic is what she called "perspective," concluding that altruists perceive their relationship to others in more universal terms than do most people, with lower boundaries between self, kin, and group identity and identification with others generally.

A very different application of the case study approach is that of Weaver and Rockman and colleagues, who sought to provide a theoretical framework for a series of case studies examining whether and how institutional differences influence the capabilities of government. Their case studies compared the American experience in a number of issue arenas with those of other industrialized democracies. Beginning with the contrast between presidential and parliamentary systems, they analyzed the ways in which this distinction, although important, interacts with various other institutional and noninstitutional factors across time and in varying policy contexts.

An important subset of the case study method is participant observation, an approach that political scientists have borrowed from anthropology. The examples used here derive from the study of the U.S. Congress.

Richard Fenno has spent his career studying Congress, is a leading practitioner of participant observation, and has written trenchantly on this method in a series of essays published in 1990. Although most studies of congressional behavior have taken place within the confines of the nation's capital, Fenno followed members back to their states and districts in order to study their relationships with their constituents. In the 1970s, he explored the "home styles" of 18 members of the House of Representatives. Later, he did the same for several members of the Senate.

For his research to be successful, it was important that Fenno develop a high level of rapport with his subjects. This might mean anything from helping to change a member's flat tire to blending in with an election campaign by pitching in with stamping envelopes, working a phone bank, or engaging in other "materially trivial" forms of participation. Fenno was aware of, and wrote perceptively about, the dangers of too much rapport. Such overidentification runs the risk of interfering with the researcher's judgment. Fenno maintained that in their research, "political scientists ... should refrain from engaging in any behavior that has the intention of affecting political outcomes." To guard against these dangers, he developed a very strict set of ethical standards as a way of maintaining appropriate boundaries between researcher and subjects, declining even to register with a political party or join any interest groups. Subsequent experience led him to modify some of his self-imposed limitations. For example, following publication of his research on Dan Quayle after the former senator had become vice president, Fenno decided to abandon his former practice of refusing to give interviews to the media.

Despite his efforts to maintain a strictly professional relationship with his subjects, Fenno found that he could not maintain complete detachment: "I could not bring myself to be indifferent to their electoral success. *I wanted them all to win.*"

As Fenno acknowledged, his research has involved much more observation than participation. This is in

marked contrast to what happens when a political scientist becomes a political actor and then writes about the experience partly from the perspective of his or her scholarly discipline. Such actors, of course, do not take the same vow of political celibacy advanced by Fenno. North Carolina Democrat David Price first came to Congress in 1987. He has since written a memoir based primarily on his experiences but drawing on the literature of his discipline, including the work of Fenno. Somewhere in between Fenno and Price on the participation–observation spectrum is the work of scholars such as Redman and also Dwyre and Farrar-Myers, who have worked on legislation as congressional staffers and then written case studies attempting to shed light in a systematic way on the legislative process.

Content Analysis

A long-standing concern among some scholars of international relations has been the development of databases, derived from public records such as newspapers or wire services, concerning events involving interactions among nations. Such scholars have hoped that by uncovering patterns of behavior in these data, they may be able to develop a sort of "early warning system" for the prevention of international conflict.

Two pioneering efforts at developing international event databases were Azar's Conflict and Peace Data Bank (COPDAB) and McClelland's World Event/Interaction Survey (WEIS). The COPDAB database includes information from various sources for the years 1948–1978. The WEIS database includes the period from 1966 to 1978 and covers events reported in the *New York Times.* Both are available from the Inter-University Consortium for Political and Social Research.

The Kansas Event Data System (KEDS) is one effort to extend (e.g., by using electronic rather than manual coding) and update COPDAB, WEIS, and other earlier work. The KEDS project Web site provides freely downloadable software and a series of regional data sets. Philip Schrodt and colleagues at KEDS have used these data to study such questions as the prediction of conflict in the Balkans, crises phases in the Middle East, and media coverage of the Intifada.

The Intranational Political Interactions project applies the events data analysis framework to conflict within nations. Downloadable data files are available for 10 countries in Africa, Latin America, and Asia.

Formal Rational Choice Models

Employing what is variously referred to as rational choice or public choice models, a number of scholars have sought to apply ideas originally developed in economics to the study of political choices. Three examples that have each led to an extended dialog within political science are described here.

The Paradox of Voting (Arrow's Impossibility Theorem)

In 1951, economist Kenneth Arrow described what he called the "well-known 'paradox of voting.'" Although he did not claim to have originated it, he is credited with the systematic formulation of what has also come to be known as Arrow's impossibility theorem. The theorem states that given more than two voting alternatives, and absent the assumption that they are "single peaked" (i.e., that an individual's first preferred choice determines his or her second choice), there is "no method of voting ... neither plurality voting nor any scheme of proportional representation, no matter how complicated" that will guarantee an unambiguous aggregate preference. For example, suppose that in 1996 three voters were asked to choose between Clinton, Perot, and Dole in that year's presidential contest. Suppose further that their preferences were ordered as follows:

Voter 1: Dole, Perot, Clinton
Voter 2: Perot, Clinton, Dole
Voter 3: Clinton, Dole, Perot

Since there are various considerations that might govern an individual's choices among these options, all three voters may have been acting quite rationally as individuals. Taken collectively, however, any one of the three options would have been rejected by a majority when pitted against only one of the alternatives. A majority would have preferred Dole to Perot, Perot to Clinton, and Clinton to Dole.

Arrow's theorem has given rise to a substantial body of literature across a number of different disciplines, including political science. Jones *et al.*, for example, employed computer simulations to show that when voters are able to order their preferences across all options, the likelihood of producing a majority decision is smaller the larger the number of voters. However, when voters are indifferent among some choices, the problem is more serious for small groups, such as committees, than with a large electorate.

The Prisoners' Dilemma

Originated in the 1920s by mathematician and physicist Von Neumann and introduced into economics in the 1940s by Von Neumann and Morgenstern, game theory has become a common approach to the study of political cooperation and conflict. One of the best known games is the prisoners' dilemma. Lalman *et al.* (1993) noted that

"the prisoners' dilemma . . . together with the voters' paradox, is by far the most widely known and celebrated example of the use of formal theory in political science."

The prisoners' dilemma is a non-zero-sum game. If two prisoners are being separately interrogated about a crime that they may have committed, their optimal strategy would be to cooperate by remaining silent. Not knowing what the other will do, however, it is individually rational for each to defect—that is, to confess in exchange for a reduced sentence.

This might suggest a rather pessimistic view of human interaction, but this need not be the case. Axelrod showed that iterative prisoners' dilemma games can lead to cooperation among players over time. He demonstrated that a strategy he called "tit for tat" ("don't be envious, don't be the first to defect, reciprocate both cooperation and defection, and don't be too clever") encourages cooperation and tends to win out against other strategies in a wide variety of contexts. Among the examples he discussed is the "live and let live" strategy adopted, under certain conditions, by soldiers on both sides during the trench warfare of World War I.

As Axelrod noted, cooperation is not always desirable (e.g., for the police who are interrogating the prisoners). Geer and Shere used this same insight to critique the common assumption of advocates of the "responsible party" approach that competition within parties (through devices such as the direct primary) weakens accountability by reducing party cohesion. They pointed out that since interparty competition is an iterative process, parties become aware of each other's strategies over time and learn to cooperate. They may reach an agreement to share patronage and other political benefits and avoid addressing voter concerns. In a more or less closed party system, such as the two-party dominant system in the United States, interparty competition by itself may not provide a mechanism for challenging such an arrangement. Intraparty competition through an open system of nominations introduces challengers who, like the prisoners in the prisoners' dilemma, have an incentive to not cooperate with the other players.

Downs's Party Competition Model

In 1929, the economist Hotelling proposed a model explaining why retailers in direct competition often choose to locate their stores in close proximity to one another. He argued that all else being equal, either of two competing companies could improve its market share by moving closer to the center of town. Near the end of his essay, he devoted a paragraph to arguing briefly that this same logic explains why in America's two-party system, "each party strives to make its platform as much like the other's as possible." Twelve years later, another economist, Smithies, suggested that if demand is elastic, a company will at some point lose more business from customers at the outskirts of town than it will gain from its competitor by moving closer to the geographic center. Although Smithies did not make the connection explicit, others have noted that nonvoting can be seen as the political equivalent of elastic demand. Some people will choose not to vote if they perceive that both parties are too distant from their preferences, and this might help explain why the Republican and Democratic parties are only somewhat similar ideologically.

In 1957, another economist, Anthony Downs, developed the political implications of these ideas systematically, considering the model under varying conditions (regarding the distribution of voters' ideological preferences, the number of parties, the entry of new parties, and the type of electoral system). His ideas have had a major influence on the way in which political scientists have thought about party competition, and his essay on this topic is a centerpiece of his book, *An Economic Theory of Democracy*, which Grofman called "one of the most influential and frequently cited works in social science of the post-World War II period."

Statistical Analysis

This section overlaps the previous ones to varying degrees. By their nature, experimental and survey research almost always require statistical analysis of the data collected. This is often true for content analysis, but sometimes such analysis is purely qualitative. As noted previously, formal models may or may not be tested against real or simulated data. Fenno notes that because of its reliance on small samples, participant observation does not lend itself well to quantitative analysis. Because of the unrepresentative as well as small size of the samples, not to mention the necessarily unscripted and unstructured nature of the observations, it is better suited to generating than to testing hypotheses. The same is true of most, although not all, case studies generally.

Overall, as shown in Table I, political research usually does involve statistical analysis of one sort or another. Contemporary political science research employs the panoply of statistical methods found in the social sciences generally, from the simplest to the most elaborate. Over time, these methods have become increasingly complex. Sometimes, however, a very basic technique can be employed with substantial elegance, as in the following examples.

The standard deviation is normally used either for simple descriptive purposes or as an intermediate step in a more elaborate calculation. Seldom is it used directly to create a variable. One exception is its use by Beck to demonstrate the decline of regionalism in American politics. Using election returns for the presidential contests

from 1896 through 1992, Beck charted the standard deviations among the states in the winning candidates' share of the popular vote, demonstrating a generally steady decline in interstate variation. A similar example is the use by Smith and Gamm of the standard deviation in roll call voting scores as a measure of lack of cohesion within political parties in Congress. Change over time in this measure was used by the authors to assess differences in the importance of party leadership.

Another example having to do with the creation of measures is the use of "feeling thermometers." A common limitation of most survey data is that although surveys are often the most direct way to measure political attitudes and behavior, the resulting measures are mathematically weak, usually only nominal or ordinal. Feeling thermometers, which have been used by the National Election Study since 1964, are usually treated as interval measures. To construct these thermometers, respondents are asked to indicate how warmly or coolly they feel about a political party, a group or category of persons, or an individual candidate or other public figure. A score of 50 indicates neutrality, with scores in the 50–100 range indicating warm or favorable feelings and lower scores reflecting cool or negative feelings. An example of the use of these thermometers in research is the work of Bolce and de Maio, who showed that negative feelings toward Christian fundamentalists have become an important predictor of feelings toward the Republican and Democratic parties, and of voting choice, even when other factors are held constant.

The treatment of feeling thermometers as interval measures has not gone unchallenged. Using data from the 1992 NES, Jacoby examined a measure created by subtracting the Clinton thermometer from the Bush thermometer. He found that intervals near the middle of the scale reflect larger differences in candidate preference than those near the extremes. Jacoby concluded that the level of measurement of these and other "pseudo-interval" measures should be tested rather than merely assumed.

More complex statistical techniques used in political science run the gamut from common forms of analysis, such as analysis of variance and correlation and regression, to more specialized tools, such as factor analysis, cluster analysis, canonical correlation, discriminant analysis, LISREL models, and Monte Carlo simulation. All of these techniques are ones that political science has borrowed from other disciplines. In addition, political scientists have developed new methods designed to address (more or less) discipline-specific problems.

NOMINATE

Students of the legislative process have long been concerned with the development of measures of roll call behavior. A common practice has been to adopt the ratings provided by interest groups such as the American Conservative Union and the (liberal) Americans for Democratic Action. These ratings consist simply of the percentage of times that legislators have voted in agreement with the group's position on "key" votes the group has selected. Other researchers have sought to develop less subjective inductively derived scales. Thurstone used factor analysis for this purpose as early as 1932, and Belknap did the same with Guttman scaling in 1958.

Poole and Rosenthal have developed a method called NOMINATE (for "NOMINAl Three-step Estimation"). This procedure is used to locate legislators in n-dimensional space based on the ideological preferences revealed by their voting patterns. The same procedures can be used to locate other actors (including interest groups and, in studies of the American Congress, the president) who take positions on roll calls. Applying NOMINATE (and more recent refinements to their original procedure) to the American Congress from 1789 through 1998, Poole and Rosenthal found that through most of America's history, most voting on roll calls can be accounted for by a single dimension closely associated with political party affiliation. This pattern was established very early and has been maintained except during periods of party breakdown (of the Federalists and, later, the Whigs). A second, much less powerful dimension accounted for intraparty divisions, most notably the north–south split within the Democratic Party during the civil rights era in the decades following World War II.

Poole and Rosenthal extended their work to the comparative study of legislative bodies. They found that across a variety of different party systems, roll call behavior in several European legislatures and in the European Parliament could be explained well in only one or two dimensions, with the most important being a Left–Right division. Generally similar results were found for the nonparty United Nations General Assembly (with the first dimension pitting NATO countries against their opponents and the second pitting north against south).

JudgeIt

In winner-take-all electoral systems (such as are used in Great Britain, France, the United States, and much of Asia and Africa) there is a complex relationship between a party's share of votes and its share of seats in legislative bodies, and an extensive body of literature has developed in efforts to unravel this relationship. First, district lines may be gerrymandered—that is, drawn to deliberately favor some parties at the expense of others. Second, there is a tendency, even in the absence of gerrymandering, for minority parties to waste votes (since only the party finishing first in a district wins representation), thus providing bonus seats to the leading party.

A specific form of this effect is the so-called "cube law," which states that in a two-party, single-member plurality system the ratio between the parties' proportion of seats will be equal to the cube of the ratio between their proportion of votes. Originally noticed at the turn of the 20th century, the cube law was first set forth formally by Kendall and Stuart. Others have since criticized the specific form of this "law" on both theoretical and empirical grounds, but there seems to be little or no dispute that some such effect usually, although not invariably, does occur.

Gelman and King developed a program called JudgeIt to sort out these effects. Among other things, JudgeIt provides estimates of the effect of "partisan bias" (favoring one party over another for a given division of votes) and "electoral responsiveness" (the effect of changes in the division of votes on the number of seats won). JudgeIt can be used to analyze an existing set of districts or to predict the impact of a proposed districting plan. It has been used both in scholarly research and in the redistricting of a number of states.

Samplemiser

Practitioners managing campaigns, journalists reporting on campaigns, and scholars doing trend analysis (on a variety of subjects) sometimes make use of series of cross-sectional surveys conducted over a period of days or longer, with each cross section often consisting of a fairly small sample. A difficulty exacerbated by these small sample sizes is that of distinguishing between noise variance and genuine change over time. Samplemiser is a program designed to filter out noise variance and "smooth" trend lines. In addition to its obvious application to day-to-day tracking polls conducted in the closing days of a campaign (in which a total sample on any given day might consist of only a couple of hundred interviews), Samplemiser is also useful in studying much larger samples, sometimes conducted over periods of years rather than days, when there is interest in studying relatively small subsets of the total.

The creators of NOMINATE and JudgeIt have made source codes for their programs freely available on the Internet. Samplemiser can be run interactively on the authors' Web site.

See Also the Following Articles

Case Study • Content Analysis • Experiments, Overview • Experiments, Political Science • Field Experimentation • Laboratory Experiments in Social Science • Politics, Use of Polls in • Qualitative Analysis, Political Science • Quasi-Experiment • Surveys

Further Reading

Almond, G. (1996). Political science: The history of the discipline. In *A New Handbook of Political Science* (R. Goodin and H. Klingemann, eds.), pp. 50–96. Oxford University Press, New York.

Fenno, R. (1990). *Watching Politicians: Essays in Participant Observation.* Institute of Governmental Studies, Berkeley, CA.

Gelman, A., and King, G. (2001). JudgeIt: A program for evaluating electoral systems and redistricting plans. Available at http://gking.harvard.edu/judgeit

Green, D., and Gerber, A. (2002). Reclaiming the experimental tradition in political science. In *State of the Discipline* (H. Milner and I. Katznelson, eds.), Vol. 3. American Political Science Association, Washington, DC.

Grofman, B. (1996). Political economy: Downsian perspectives. In *A New Handbook of Political Science* (R. Goodin and H. Klingemann, eds.), pp. 691–701. Oxford University Press, New York.

Iyengar, S. (2002). Experimental designs for political communication research: From shopping malls to the Internet. Available at http://pcl.stanford.edu/common/docs/research/IYENGAR/2002/Expdes2002.pdf

Lalman, D., Oppenheimer, J., and Swistak, P. (1993). Formal rational choice theory: A cumulative science of politics. In *State of the Discipline* (A. Finifter, ed.), Vol. 2, pp. 77–104. American Political Science Association, Washington, DC.

Sapiro, V. (1999). Fifty years of the National Election Studies: A case study in the history of "big social science." Paper presented at the annual meeting of the American Political Science Association, Atlanta. Available at http://polisci.wisc.edu/users/sapiro/papers/bignes.pdf

Schrodt, P. (1995). Event data and foreign policy analysis. In *Foreign Policy Analysis: Continuity and Change in Its Second Generation* (L. Neack, J. Hey, and P. Haney, eds.). Prentice Hall, Englewood Cliffs, NJ.

Somit, A., and Tanenhaus, J. (1982). *The Development of American Political Science*, 2nd Ed. Irvington, New York.

Political Violence

Avram Bornstein
John Jay College of Criminal Justice in the City University of New York, New York, New York, USA

Glossary

genocide Mass murder targeted to annihilate a particular group, including religious, ethnic, racial, and national groups.
militarism The ideology or practice of military domination in nonmilitary spheres of society such as the political, economic, or cultural.
post-traumatic stress disorders Persistent manifestations of anxiety caused by a traumatic event including insomnia, loss of concentration, guilt, loss of self-worth, loss of trust, hopelessness, violent outbursts, panic, and dissociative disorders.
reactive aggression Destructive behavior motivated by experiences of injury or loss, or as a reaction to a perceived threat.
structural violence Systematic deprivation resulting in chronic poverty and hunger, usually reinforced by militarized violence.
symbolic violence Attacks on human dignity and the denial of humanity including degrading representations of a group or forcing people to perform humiliating acts.
truth and reconciliation commissions Official bodies created to investigate human rights abuses of the previous regime and its opponents that aim to bring violent conflict to an end.

Politics is the struggle to maintain or transform governance over economic, bureaucratic, intellectual, cultural, or other spheres of collective activity. Violence is an action that does harm to an object. Political violence, therefore, is harmful action intended to influence or shape collective action. Political violence can target the material environment, the body, the mind, or the social order, and can include armed, structural, or symbolic forms. The motivation for choosing violence may be coordinated with clear material objectives, or it may follow complicated psychological and cultural imperatives. Acts of violence can influence victims, perpetrators, and witnesses, and it can inspire resistance or terrorize into submission. Representations of violence in writing and public ritual may work to mitigate emotional trauma and inhibit retaliation, but they can also facilitate the reproduction of violence.

Introduction

Prussian army officer and military theorist Carl von Clausewitz observed, in his famous treatise *On War* (1833), that war is merely a continuation of politics. Derived from the Greek for city or citizen, politics is a word that points to participation in governance. Political processes are those in which groups struggle and compete over a variety of goals. The essence of politics is often said to be the art of persuasion. Sociologist Max Weber, for example, wrote that charisma, tradition, and rational planning were all forms of authority that can effectively shape social action. Public ritual, literature, landscaping, architecture, education, social welfare, and the law can all shape orders of governance. Philosopher Michel Foucault emphasized the spatial, temporal, and operational organization of social institutions that discipline the body and mold self-knowledge. But when persuasion is incomplete or ineffective, coercion may become the next option. Somewhat like Clausewitz, Weber saw that the decisive means of politics is violence.

Forms

Armed Violence

Armed violence is the exertion of force meant to injure or destroy an object. Many species use their bodies to attack or defend against other bodies, but humans magnify their arms with armaments and armor. Weaponry has made

possible mass annihilations. Romans, Slavs, Crusaders, Mongols, and others wiped out entire cities over the centuries. Tens of millions died in rebellion in China in the late 19th century. Millions of indigenous Americans were murdered, and millions of Africans killed and enslaved by Europeans since the late 15th century.

In the 20th century, the increased growth of weapons manufacturing sped the pace and scope of corporeal destruction. It was the century that coined the term *genocide*: *genos* from the Greek for group and *cide* from the Latin for killing. Over 100 million people, mostly civilians, died in politically driven wars and violence including: over a million Armenians killed by Turkish assaults; 12 million Jews, Romani, leftist activists, and homosexuals killed by Nazis; over 30 million died under Stalin's rule in the Soviet Union; and over 45 million killed in China under Chaing Kai-shek and Mao Tse-tung. Millions more were slaughtered in Indonesia, Yugoslavia, Cambodia, Pakistan, the Congo, and Rwanda. Death squads tortured and murdered thousands in Guatemala, El Salvador, and Chile. Atomic bombs were dropped on two cities in Japan and a nuclear arms race ensued. The carnage continued to the end of the century. According to the United Nations, during the 1990s, about three and a half million people were killed by political violence, about 15 million lived as refugees, at least 22 million lived as internally displaced persons, and there were an estimated 300,000 child soldiers.

Although down from the 1990s, at the beginning of the 21st century weapons expenditures amounted to around $800 billion. The United States is by far the largest manufacturer and consumer of weapons, although several other countries have significant industries including Bulgaria, China, France, Germany, Israel, Romania, the Russian Federation, South Africa, and the United Kingdom. Major weapons manufactures, Lockheed–Martin, Boeing, Raytheon, General Dynamics, Northrop Grumman, and the Carlyle Group, produce missile systems, aircraft, ships, and other items costing millions or billions of dollars. The United States' military budget for 2003 was around $360 billion, and the President requested approximately $400 billion for the 2004 defense budget. By comparison, NATO, Australia, Japan, and South Korea were spending a combined total of about $225 billion, Russia about $65 billion, and China about $47 billion. This extraordinary spending is the product of what the former United States president and heroic World War II general Dwight Eisenhower called "the acquisition of unwarranted influence, whether sought or unsought, by the military–industrial complex. The potential for the disastrous rise of misplaced power exists and will persist."

Perhaps more devastating than large offensive systems are the proliferations of small weapons, like pistols, assault rifles, machine guns, grenade launchers, small mortars, and shoulder antitank and antiaircraft missiles. The United Nations estimates that there are around 500 million of such weapons in the world, between 40% and 60% of which are illicit. New nonlethal technologies, such as electro-shock stun weapons and chemical crowd-control devices, are also being marketed and sold globally. While these can reduce fatalities in some cases, they are also used to torture civilians.

There has been some movement toward arms control. Land mines, which claim 15 to 20 thousand victims in 90 countries each year, have been banned by 141 states, but China, Russia, the United States, and others refuse to sign the Land Mine Treaty. While the Strategic Offensive Reductions Treaty has led the United States and Russia to agree to make modest reductions of deployed nuclear weapons by 2012, the United States has only selectively supported the Nuclear Nonproliferation Treaty and the Comprehensive Test Ban Treaty, and may resume banned nuclear testing.

Structural Violence

Not all physical violence is directly caused by armed force. People also suffer from administrative or managerial actions that promote chronic and severe inequality. Usually reinforced by armed violence, structural violence inflicts pain slowly by systematically keeping people in poverty and material vulnerability. The concept of structural violence is at odds with economic models that frame poverty only as a product of mechanistic market forces. To identify some poverty as structural violence indicates a perpetrator and a victim, rather than a winner and loser.

In the United States, groups that are victims of recurring structural violence include indigenous Americans and African Americans. After waves of military and paramilitary violence from the 16th to the 19th centuries, in the 20th and 21st centuries indigenous Americans continue to have land, minerals, timber, fish, and other natural resources taken from their territories, while others are being poisoned by the dumping of toxic or radioactive industrial wastes. In the United States, African enslavement for plantation work formally ended in the 19th century, but segregation, enforced by Jim Crow laws and antiblack terrorists like the Ku Klux Klan, kept many African Americans in poverty as tenant sharecroppers into the 20th century. Industrial employers drew African Americans to cities, but systematic denial of rights through neglect of education, health care, employment, and other public services, combined with heavy policing and incarceration, have kept opportunity limited for many.

Across the globe, systematic disparities have grown with a handful of rich nations, including the United States, the United Kingdom, Japan, Germany, France, Canada, Italy, and Saudi Arabia, managing trade and industry through the International Monetary Fund, the World Bank, and the World Trade Organization. They have used financial and military assistance to encourage governing elite to impose

economic austerity conditions. These "structural adjustment" programs favor the deregulation of production and trade and the elimination of public welfare, making conditions attractive to investors seeking high returns. In theory, these measures assist in the economic development of "emerging" markets, but few debtor nations have eliminated their debts, many have suffered environmental and social devastation, and wealth has generally gone to small groups of local elite and international investors. At the start of the 21st century, the systematic gap globally between those with material privilege and those without sufficient resources is wider than ever before in human history. According to the UN, the richest 5% of people have incomes 114 times greater than the poorest 5%; the United States' wealthiest 25 million earn about the same as the world's poorest two billion people; in 1999, 2.8 billion people lived on less than $2 a day, 1.2 billion of whom were surviving on less than $1 a day; more than 30,000 children die each day due to curable diseases; and while extreme poverty decreased in South Asia, East Asia, and the Pacific, it rose dramatically in Africa.

Symbolic Violence

Political domination not only strikes at the material integrity of a society, it also attacks the dignity of victims. Emotional well-being and a sense of self-worth can be hurt by armed and structural violence, as well as by witnessing degrading representations of one's self or being forced to perform humiliating acts. Such symbolic violence is meant to injure or destroy the recognition of mutual personhood. The Nazi Holocaust began with acts of humiliation like forcing Jews to wear yellow stars in public, which meant to indicate that they were inferior. European colonial violence and governance were also accompanied by white supremacist images of the colonized. Africans were represented as stupid, ugly savages. So-called "Orientals," a term that refers to all "eastern" peoples from Moroccan Arabs to the Japanese, have been portrayed as sadistic, despotic, and misogynist barbarians. These forms of symbolic or representational violence define the victim as something not quite human. Dehumanization places the victim outside the community, beyond the circle of moral behavior, and allows the withdrawal of empathy. By helping ordinary people to distance themselves from the pain of those suffering, symbolic violence allows them to commit or condone horrible acts of armed and structural violence.

Actors

States

The major military and policing forces of the world, and the major wars and genocides, have been organized by modern states. Modern states vary greatly, but whether weak or strong, they often include overlapping networks of coercive and administrative agencies. Max Weber described a state as a sovereign, territorial, and compulsory organization that claims a monopoly on the legitimate use of violence to enforce its order. Enlightenment philosophers argued that such state organizations are good because they could eliminate violence among citizens, in favor of specialized agents that defend against other states and criminals. State agents are directed, but not necessarily controlled, by a particular government. In many states, militarized forces and business elites have noticeable power over official governing bodies and those whom they are to protect.

Privateers

Contrary to the idea that states monopolize legitimate violence, the business of war, in all its aspects, has flourished in the private sector, and products may be purchased by states or nonstate actors. While mercenary soldiers are often from vulnerable populations and receive minimal compensation, others are on the payroll of large private companies led by the former personnel of state-sponsored militaries. There are massive contracts for noncombat services like information technology, food, and cleaning and maintaining vehicles, buildings, and grounds. Many companies also provide lethal equipment, training, planning, protection, reconnaissance, targeting, and even combat. Among the largest are Airscan, Northrop Grumman, DynCorp, Military Professional Resources, Armor Holdings, Vinnell Corporation, Sandline International, and Executive Outcomes. Private mercenary forces have entered battles in several African countries, particularly where profitable resources like mining operations ensure their payment. They have also worked in the former Yugoslavia, Afghanistan, and with the infamous Colombian Cali cocaine cartel led by Carlos Castaño. To avoid accountability, some states use such services to create proxy militia.

Partisans

The proliferation of weapons has made it possible for many groups with antistate or antigovernment agenda to mount campaigns of violence, sometimes with the backing of other states. The victims often label their attackers terrorists, which is a pejorative term usually referring to nonstate groups who use violence to intimidate for political changes. Some of these antigovernment and antistate groups have sizable guerrilla forces and control significant territory within the state from which they operate. These would include the Fuerzas Armadas Revolucionarias de Colombia (FARC), the Zapatistas in Mexico, Tamil Tigers in Sri Lanka, the Shining Path in Peru, Hizbollah in southern Lebanon, UNITA in Angola, Kashmiri forces in India, and Mujahadeen in Afghanistan.

Largely due to the superior power of their opposition, some groups hold no significant territory and must operate in smaller units. In the United States, a militia movement organized training camps and networks implicated in the bombing of the Federal Building in Oklahoma City in 1995 that killed 168 people. Several Islamic groups have operated within and against a number of countries including Hamas in the Israeli occupied Palestinian territories, al-Gama'a al-Islamiyya in Egypt, and the Islamic Salvation Front in Algeria. According to the United States government, Al-Qaeda launched attacks in Afghanistan, Yemen, Saudi Arabia, East Africa, Indonesia, and the United States, where on September 11, 2001, close to 3000 people were killed in New York and Washington, DC. Like most other militaries, some of these groups organize financing, purchasing, training, and planning through networks that spread across several nations.

Multistate Alliances

In the 20th century, multistate organizations became directly involved in political violence, in the name of preventing continuing bloodshed. In 1948 the United Nations emerged and served as a significant global multistate organization. For its first decades, the United Nations mainly focused on the Arab–Israeli conflict, but in the 1990s its scope expanded into over a dozen countries including Cambodia, Kuwait, El Salvador, Guatemala, Haiti, East Timor, eight African nations, two former Soviet states, and much of the former Yugoslavia. In 2003, there were 14 U.N. peace keeping missions, with about 37,000 security personnel from 89 countries, and 10,000 more supporting personnel. In the 1990s, the United Nations also helped organize two ad hoc international criminal tribunals, one for the former Yugoslavia and the other for Rwanda. In 1998, an International Criminal Court, to complement national judicial systems when they are unwilling or unable to investigate or prosecute war crimes, was approved by 120 U.N. members, but China and the United States, among others, refused to participate.

Motivations

Economic

Violence can often be instrumental to obtaining desired objects. For several thousand years, the power of violence allowed victors to demand tribute from the vanquished. With trans-Atlantic colonialism European monarchs centralized their control at home and went from merely taxing peasant and mercantile subjects to sponsoring extensive monopolies on trade and even direct organization of production. The Spanish, Dutch, English, and French struggled over resources from the Americas, Africa, South Asia, and East Asia. The slave trade and the seizure of land for plantations and mines defined much of the colonial process. In the 20th century, military forces continued to be deployed around the world to secure important markets, especially for primary resource extraction in the developing world such as mines, narcotics, timber, and oil.

Psychological

Clausewitz's famous phrase, that war is a continuation of politics, is often understood incompletely. Clausewitz was comparing this proposition with its antithesis, that war is nothing but a brutish act. He explained that war is a synthesis of both brute aggression and political calculation with a great deal of chance. While rational economic interest may shape conflict, there is often too much violence to be explained merely by instrumental motivation or material want. Sometimes violence aims only to make the victim suffer. Sigmund Freud argued that such brutish aggression is innate, an instinct, like the drive for pleasure. Ethologists have described aggression as common to most animals and humans. Some sociobiologists have argued that aggression is a selected trait in the brain involving the production of neurotransmitters like noradrenalin, dopamine, and serotonin in the limbic system.

Though there are biological components of violence, humans are not always violent. Contextual factors often cause violence to manifest. Psychological theories have emphasized the role of frustration in the creation of aggression projected either toward the source of frustration or a substitute object. Erich Fromm wrote that while humans can develop the love of life, if frustrated humans develop destructive behaviors such as the desire to control others, sadism, or the desire to destroy things, necrophilia. Such a personality prefers violence to solve a problem rather than sympathetic efforts. Human destructiveness often comes from experiences of injury or loss that produce reactive aggression. From this perspective, collective forms of violence may be a reaction to a perceived threat, a "counter" violence. Revolutionaries, counter revolutionaries, terrorists, and counter terrorists, all portray their movements as the offended victims in order to motivate and justify violent retaliation. Even if such reactionary aggression can achieve no clear political goal, such violence appeals to a psychology of recognition. Despite the futility of armed resistance against overwhelming militaries and police apparatuses, those feeling overwhelming frustration may pursue suicidal revenge rather than submit to their oppressors.

Cultural

Violence takes very specific cultural forms. The choice of targets is often made according to culturally based, often

relatively local notions of affiliation or identity that combine notions like racial, ethnic, gender, or religious difference. The colonial torture and murder of indigenous Americans and Africans was often predicated on European fantasies of taming wild, heathen, and satanic races. The Nazi Holocaust took place within a context of long-standing anti-Jewish and anti-Semitic myths of Jews as a race of conspirators and despicable materialists. Not only the target, but also the act may be structured by specific cultural distinctions. The castration of African American men when lynched expressed the pervasive white supremacist stereotypes of African men as rapists. In the Hutu genocide of Tutsi in Rwanda, the cutting off of limbs and breasts was patterned by a local logic of healthy flow in the body and body politic. In the Balkans, rape became a weapon of war largely because of the powerful cultural shame involved with such violence. The target and the method show distinct cultural patterns.

There are also international patterns of identification that transcend local ideas of culture. Modern religious fundamentalism, characterized by textual literalism, moral certainty, and outstanding organizational capacity, spread across state borders in the 20th century and included Christian, Islamic, Hindu, Buddhist, Jewish, Sikh, and other expressions. Many of these groups provide education, social services, and electoral support, and they have come to influence some states and their militaries or even form independent paramilitary groups. A second more pervasive global cultural influence on violence is militarism. In addition to the central role of military groups in politics and industry, somewhat like religious fundamentalism, the military has strong symbolic significance. Those in the military are directly exposed to messages that associate manhood, honor, and violence. The general public, too, receives similar messages. The military parade has become a standard ritual of state display. Audiences across the globe watch the movies, many from Hollywood, that make military violence seem glorious and titillating. These cultural images, like ancient warrior myths, may shape widespread patterns of political violence.

Impacts

Trauma

Dead bodies, amputated limbs, torched farms, slaughtered livestock, bombed out buildings, these are the most visible traumas of war. Violence's powers to kill or maim people, and destroy natural resources and community infrastructures, are the most obvious costs to the victims of battle. But there are less visible traumas. Bruno Bettelheim's studies of Holocaust survivors revealed surprising patterns of emotional burden. Some victims experienced guilt for surviving when so many died. Others described a sense of identification with their jailers and executioners. Many psychologists have observed post-traumatic stress disorders in which the anxiety of past trauma continues to manifest in behaviors and experiences such as insomnia, depression, violent outbursts, and dissociative disorders. Such experiences can tear apart family and community and destroy images of self, trust of others, and a sense of hope.

The trajectory of trauma has several intersubjective aspects. Psychotherapists have long argued that for healing to occur, victimized people must develop a way to talk about, or even commemorate and mythologize what occurred. Through these processes the trauma comes to define part of their identity as members of a group. Public rituals, monuments, testimonial narratives, historical studies, and bodily practices are familiar forms through which people collectivize emotional trauma. Such memorial practices can transmit trauma from one individual to another, across a family, to a community or even transgenerationally, and while they may help people mourn, they may also keep feelings for retribution alive and frame future reprisal.

Domination

One consequence of violence is that fear can inhibit the organization of resistance to domination or exploitation. Public torture and execution, and some military attacks, are meant to display the overwhelming power of the state and terrify subjects into obedience. The threat of future violence can exert great control. Political actors sometimes use old, or even create new enemies that heighten popular anxiety for political advantage. As Randolph Bourne said, "War is the health of the state."

Resistance

Violence is often permitted or encouraged to secure the passivity, compliance, or elimination of particular populations; but, sometimes it can create solidarity and communities of resistance. For example, the mothers of those who disappeared at the hands of the Argentine military were united in their grief and demonstrated in the Plaza de Mayo in Buenos Aires to demand knowledge of their missing loved ones and, eventually, to call for criminal prosecution of the perpetrators. Cycles of violence between groups, Pakistanis and Indians, Tamils and Sinhalese, Palestinians and Israelis, and many others, while intended to intimidate, may only harden defiance on each side and perpetuate reprisal.

Representations

Writing

Several international organizations, like Human Rights Watch, Amnesty International, Physicians for Social

Responsibility, and many local organizations and individuals witness and record violence to work against suffering in at least three ways. Some victims need to talk about their trauma in order to render it speakable, to mourn the dead, or to reconstruct a self image. Giving voice and face to suffering can reduce the social distance between victims and those who may be sympathetic allies. Writing about violence can expose abuses of human rights and incriminate those responsible for crimes.

However, in other circumstances, writing about violence and recording suffering can further the terror it purports to explain. Talk of violence can increase terror, fear and panic. Reading about or watching repeated television images of suffering may also desensitize viewers to violence rather than create empathy. Rather than being a step toward ending suffering, when watching violence stirs panic or becomes a voyeuristic form of entertainment, it encourages more violence.

Rituals

Healing people and communities from trauma and breaking cycles of violence often requires public ceremonies to communicate the restored dignity of victims. Sometimes ritual feasting or exchange reestablishes recognition between people. Modern courts have ritualized establishing truth and extracting retribution to restore a sense of justice. A new and significant ritual to emerge in the 20th century was the truth and reconciliation commissions (TRC). TRCs are official bodies that investigate human rights abuses committed by governmental or rebel forces in order to ease the transition to a new political order by creating a renewed sense of justice and legitimacy. In South Africa, the TRC was charged with the tasks of creating a complete record of past abuses, making recommendations to prevent future abuses, restoring the dignity of victims, and granting amnesty to perpetrators who confess fully. Anglican Archbishop Despond Tutu recognized that to heal the nation, the restoration of justice and dignity must happen without retribution. Retribution would have been an obstacle to establishing truth and could have perpetuated the violence of the conflict. In addition to South Africa, in the last quarter of the 20th century, about two dozen TRCs were created across the world including Guatemala and East Timor.

See Also the Following Articles

International Relations • Political Conflict, Measurement of

Further Reading

Bauman, Z. (1989). *Modernity and the Holocaust.* Cornell University Press, Ithaca, NY.

Farmer, P. (2003). *Pathologies of Power: Health, Human Rights, and the New War on the Poor.* University of California Press, Berkeley, CA.

Giddens, A. (1987). *Power, Violence and the State.* University of California Press, Berkeley, CA.

Keane, J. (1996). *Reflections on Violence.* Verso, New York.

Lifton, R. J. (1979). *The Broken Connection: On Death and the Continuity of Life.* American Psychiatric Press, Washington, DC.

Mahmood, C. K. (1996). *Fighting for Faith and Nation: Dialogues with Sikh Militants.* University of Pennsylvania Press, Philadelphia.

Schirmer, J. (2000). *The Guatemalan Military Project: A Violence Called Democracy.* University of Pennsylvania Press, Philadelphia.

Tambiah, S. (1986). *Sri Lanka: Ethnic Fratricide and the Dismantling of Democracy.* University of Chicago Press, Chicago.

Taylor, C. (1999). *Sacrifice as Terror: The Rwandan Genocide of 1994.* Berg, New York.

Tilly, C. (1990). *Coercion, Capital, and European States, AD 990–1992.* Blackwell, Cambridge, MA.

Politics, Use of Polls in

Jeffrey A. Fine
University of Kentucky, Lexington, Kentucky, USA

D. Stephen Voss
University of Kentucky, Lexington, Kentucky, USA

Glossary

benchmark poll Large public opinion survey, typically taken at the beginning of a campaign to gauge the initial party, candidate, and policy preferences of constituents; helps shape campaign strategy by providing vital information about what kinds of messages will motivate the public.

dial groups Small groups of individuals who are given handheld electronic devices that allow them to register their ongoing feelings about a speech, debate, or political advertisement while they are being exposed to the message; may not be representative of a constituency or audience, but can provide detailed information about the types of people in the sample.

direct mail Surveys mailed (or emailed) to potential voters; recipients participate voluntarily, and so may not be representative of the electorate. A relatively inexpensive way to poll, thus commonly used by public officials to measure statewide or district-level opinions of constituents.

exit poll Survey conducted by media outlets during primary and general elections to project the outcomes before the ballots are tallied. Questionnaires ask respondents about their vote choices, as well as various demographic questions; the accuracy of the poll depends on the representativeness of both the selected precincts and the participating respondents at each precinct.

focus group A group consisting of about 7 to 10 individuals who are talked through political issues by a trained moderator. The modest size of the group does not permit a representative sample, but focus groups do provide more detailed information than can be obtained from traditional survey research. Because focus groups can indicate how individuals respond to alternate ways of phrasing the same political messages, for example, campaigns use them to plan advertisement strategies.

media ratings Surveys that provide information about television viewership in various markets, as well as demographic information about the audience. These ratings are used to determine the costs and the target audiences for television political advertisements.

push poll An unscientific poll conducted by political campaigns under the pretense of objectivity; although evidence about how constituents intend to vote in an upcoming election may be obtained, the main purpose of a push poll is to disseminate political propaganda that either benefits those conducting the poll or damages the campaign of an opponent.

recruitment survey A poll conducted by political parties very early in the campaign cycle to provide potential candidates with information about the viability of their candidacies.

straw poll Although used to refer generally to a poll with a large but unscientific sample, the main usage in politics applies to surveys conducted among the membership of a political party. Due to the sampling methods of a straw poll, results cannot be generalized to the entire population, but are useful for determining which issues, messages, or candidates can motivate the party faithful.

tracking poll A survey taken frequently during the latter stages of a campaign to monitor public reactions to various political events, such as speeches and political advertisements.

trend poll A survey taken intermittently during a campaign following the benchmark polls; results indicate changes in mood and issue and voter preferences that occur throughout the campaign process.

Public opinion polls are a common feature of democratic politics. Numerous political actors, regardless of whether they are governing or are involved in conducting elections, take advantage of the usefulness of polls to characterize citizen attitudes and policy preferences.

Introduction

Literature and history are replete with stories of despots who nonetheless worried about the sentiments of their subjects. All four gospels of the *Bible*, for example, portray Roman governor Pontius Pilate as reluctant to crucify Jesus; he does so to pacify the people he rules. Shakespeare's "Julius Caesar" opens with the tribune Marullus berating citizens for their political fickleness, then climaxes as Marc Antony's funeral oration whips his audience into a mob against those who have taken power at Caesar's expense. Western history is littered with the corpses of real rulers who failed to recognize the danger of their declining popularity. Notorious examples include English King Charles I in the 17th century, French King Louis XVI in the 18th century, and Russian Czar Nicholas II in the 20th century. Even autocrats have—to paraphrase the late political scientist V. O. Key, Jr.—found it prudent to heed public opinion.

Compared to autocrats, politicians in democratic systems may have to place more emphasis on satisfying the citizenry, because elections institutionalize mass participation. But prior to World War II, the tools available for gauging popular impulses in democracies were not much more effective than were those possessed by other regimes. Elected leaders usually relied on trusted advisors, well-connected party bosses, or savvy legislators to gather information about voter sentiment. Leaders also have paid close attention to voices in the media, treating them as indicative of regional attitudes. For example, after Teddy Roosevelt invited Booker T. Washington to dinner (the first instance of a U.S. president entertaining an African-American man at the White House), the critical response of Southern journalists convinced Roosevelt to shy away from future contact with Black leaders. In short, elected officials of the past customarily depended on limited, and usually biased, avenues of popular expression.

The birth of scientific survey methods, however, opened vast new resources for democratic leaders seeking to measure popular sentiment. Technological and intellectual developments in the aftermath of World War II, such as scientific sampling, advanced statistical analysis, and behavioralism, combined to create workable polling techniques that politicians could use to promote their own interests. They have exploited the new techniques to strategize election campaigns, to shape political platforms, to select among public policies, and to guide legal rulings. As a result, public opinion polls have contributed significantly toward increasing the responsiveness of democratic governments, and perhaps also toward increasing the ability of a nation's leaders to manipulate the citizenry.

Because the scientific aura surrounding surveys appeals to populist, pragmatic, and empiricist strains in the national culture, political actors in the United States have been especially quick to adopt polling innovations. It is no accident that the first commercial pollsters emerged in the United States or that, until recently, at least, pioneering polling techniques and electioneering strategies have typically originated in the United States. American media organizations, meanwhile, have learned that audiences appreciate seeing themselves reflected back in newsprint or in broadcasts. Opinion polls serve as the centerpiece of political journalism in the United States. The remainder of this discussion therefore focuses, out of convenience, on the use of polls in American politics.

History of the Use of Polls

The desire to survey public attitudes is as old as the American political system. Although the birth of modern polling changed the face of public opinion research, allowing unprecedented access to the pulse of the public mood, experimentation with surveying popular opinion preceded the development of sound methods for doing so. One of the earliest techniques was the use of the "straw poll," an unofficial vote, to gauge public opinion and support for various candidates for office. The practice traces back to a time before scientific polling methods, perhaps as far back as the elections of 1824. Sometimes voters attended meetings called explicitly to assess their political preferences; sometimes voters answered straw polls taken during militia or party meetings called for other purposes. Political activists also sometimes conducted straw polls by leaving "poll books" in public places for citizens to provide information about their voting preferences. All of these methods rely on unrepresentative samples; they cannot be generalized to all voters.

Another early sampling technique involved mail-in ballots. A periodical called *Literary Digest* successfully predicted the winner of every presidential election between 1916 and 1932 by tabulating millions of mail-in ballots. However, the magazine distributed ballots to an unrepresentative sample culled from automobile registration lists and telephone directories, and relied on respondents to complete the questionnaire voluntarily. Because respondents with telephones and automobiles during this depression-era election were much wealthier than the typical voter, the *Digest* incorrectly predicted an overwhelming Republican victory in 1936, rather than a Franklin Delano Roosevelt (FDR) reelection. The 1936 election, however, was the first time that scientific polls were taken by George Gallup, Elmo Roper, and Archibald Crossley. Gallup not only predicted that FDR would defeat Alf Landon, he also predicted that *Literary Digest* would get the prediction wrong because of bias in their sampling methods. That Gallup's organization could outperform *Literary Digest*, despite a sample of thousands rather than millions, underscored the importance of scientific sampling

methods. This success led to widespread development of sampling techniques around midcentury—not only in politics, but also in marketing.

Some of the innovations in survey sampling techniques resulted from generous government investment. The U.S. Department of Agriculture established a Division of Program Surveys in 1939 to study farm opinion. The division, directed by Rensis Likert, was the first governmental body to conduct public opinion research regularly. It developed numerous polling techniques that have become standard. These included using open-ended rather than yes-or-no questions, asking multiple questions on the same topic to develop a scale of opinion, and following methods of probability sampling that limited the discretion of interviewers when they obtained respondents. The federal government continued to commission surveys during World War II. In 1942, for example, the Office of War Information established a Surveys Division that conducted more than 100 studies of civilian attitudes about wartime problems. During the same period, a U.S. Army research branch studied troop morale. The wartime research programs took advantage of the specialized insights of multiple social sciences and so contributed enormously to an understanding of how opinions are formed and changed. They also helped sustain the polling profession and cultivate its expertise during a time of limited resources.

Trial-and-error experimentation contributed significantly to technical polling innovations. For example, Gallup's sampling methods came up short in the 1948 presidential election; his organization's prediction that Thomas Dewey would oust President Truman led some newspapers to print embarrassingly wrong articles headlined "Dewey Defeats Truman." Some blamed Gallup's faulty prediction on a decision to stop polling nearly 2 weeks before the general election, even though approximately 14% of respondents remained undecided at that time. Others blamed Gallup's use of quota sampling, a method that seeks samples perfectly representative of population demographics but allows individual interviewers more discretion to pick cooperative and easy-to-reach subjects. Gallup's high-profile failure led to several adjustments in survey techniques. For example, polls taken in the days before a presidential election became more common and firms increasingly opted for random samples rather than those collected under demographic quotas.

Use of Polls by Politicians/ Officeholders

Almost every U.S. president has tried to gather intelligence on public sentiments in one way or another, starting with George Washington, who reportedly rode around the countryside to gauge citizen opinion toward the federal government. But America's wartime president, Franklin Roosevelt, was the first who could exploit sound polling data, because scientific survey techniques were only just developing during his administration. For example, FDR used polls to tap public sentiment toward some New Deal programs, especially Social Security. He was also interested in popular views on World War II and the role that the United States should play in the war. Roosevelt often checked public opinion data before making key decisions relating to the war, such as whether Catholics approved of bombing Rome in 1944.

Harry Truman, FDR's successor, was much more uncertain of the validity and accuracy of polls. Like many politicians during his time, Truman did not believe that polls could represent the opinions of the public at large. Gallup's faulty prediction that Truman would lose in 1948 only encouraged such skepticism, and may have slowed the widespread adoption of public opinion polling for more than a decade. Not until the Kennedy presidency did the White House start using polls extensively. The use of polls by John F. Kennedy (JFK) began well before he won the 1960 presidential election. Kennedy commissioned Louis Harris to conduct polls during the campaign. Before the Democratic National Convention, a Harris survey showed that Kennedy's leading Democratic opponent, Hubert Humphrey, might be politically vulnerable in West Virginia and Wisconsin. Kennedy increased his attention to these states, and although he fell short in Wisconsin, he was victorious in West Virginia. He eventually won the nomination and the presidency. Once in office, Kennedy retained Harris' services so that he could gauge his own approval ratings as well as probe specific policy preferences of the citizenry.

Following JFK's assassination, Lyndon B. Johnson (LBJ) used polling data to measure public support for his domestic agenda. LBJ especially concentrated on public opinion late in his presidency, because he was extremely concerned with how the American people perceived their country's involvement in Vietnam. In 1966, Johnson's nightly reading included summarized results of a series of questions relating to public support for the war. President Richard Nixon took the use of polls to new heights. In his first year in office, Nixon commissioned more private polls than Johnson commissioned during his entire presidency. Nixon was obsessed with public opinion, particularly his own approval rating. During his reelection campaign in 1972, Nixon had his pollsters working frantically to find the best strategy to run against both contenders for the Democratic nomination, Hubert Humphrey and George McGovern.

Public opinion polls have never lost their critical importance since the Nixon administration. Following Nixon's resignation, President Ford examined strategies

for maneuvering out of the political hole left by the Watergate scandal. President Carter felt that public opinion was so important that he gave his pollster, Patrick Caddell, an office in the White House. Reagan met with his pollster, Richard Wirthlin, almost monthly to monitor public support for the administration and its policies. George H. W. Bush kept close tallies on public opinion and reportedly relied on poll results to shape his posture with respect to Iraq. President Bill Clinton employed surveys to a great extent, not only to conduct his election efforts but also to shape his policy stances, resulting in some observers calling his administration a "horserace presidency." Clinton made no secret about the role of pollsters in his White House. He commissioned regular polls about every aspect of American political life, at first from Stanley Greenburg and later from Dick Morris. Morris, in a kiss-and-tell book written after he fell from grace, denies that Clinton used polls to select his policies, but verifies that Clinton was constantly aware of the political significance of his policy decisions. Clinton's White House used polls to determine which actions were winning the most support and to shape public messages accordingly, resulting in a highly politicized and highly responsive form of governance that scholars call "the permanent campaign," because of its reliance on tactics once reserved for electioneering. During the 2000 presidential election, Texas Governor George W. Bush criticized the Clinton White House's reliance on surveys, promising that he would discontinue the "permanent campaign" and instead govern based on firmly held principles. Although this posture may have appealed to voters weary of Clinton-era scandals, it ignored the reality of contemporary political life, which is that polls are an invaluable and regular tool used by politicians because they are commonly accepted as a legitimate and accurate representation of mass opinion. Indeed, once in office Bush apparently recognized the necessity of watching public attitudes closely. His White House cut back on polling relative to Clinton's, but it still surveyed public opinion quite frequently by historical standards, and Bush confidant Karl Rove moved smoothly from campaign consultant to domestic advisor.

Although conventional wisdom states that presidents have used public opinion polling to gauge popular sentiment and cater to it, recent work challenges this claim. These scholars assert that politicians monitor public opinion to determine how to present their messages to win public support for their own policy preferences. Rather than following public opinion, politicians strategically craft their political messages to manipulate public opinion, in order to feign responsiveness.

Members of Congress usually lack the resources for extensive polling, although they can sustain some polling activity through the party organizations or through well-funded campaign organizations. Officeholders therefore supplement scientific polling methods with direct-mail (or constituent) surveys, which provide a cheap means for checking the mood of attentive constituents in their home districts. These polls are questionnaires sent to constituents' homes, often at public expense. Recipients determine whether to participate in the survey, and respondents complete the surveys outside the presence of interviewers. Direct-mail surveys are widely used because they provide information about voter preferences at a low cost to politicians at all levels of government.

Use of Polls by Campaigns

The use of public opinion polls becomes most intense during election years; they are invaluable to politicians seeking to attain or retain office. Political campaigns use an extensive battery of polling methods, each providing candidates with different sorts of information. Well-financed campaigns customarily use a series of benchmark, trend, and tracking polls during the election season. Early in the campaign cycle, candidates conduct benchmark polls to measure the initial issue preferences and likely vote choices of a large sample of individuals within their districts or states. These polls also can provide vital information about the relative name recognition of candidates and their opponents and can test different campaign messages. They may contain both open-ended and highly structured questions. Candidates for the U.S. House of Representatives often conduct these polls as early as a year before the general election, and Senate candidates may commission these polls 3 years prior to the election.

As campaigning commences following the initial benchmark poll, shorter trend polls, then tracking polls, are taken intermittently, up to the date of the general election. Trend polls gauge shifts in opinions and preferences that occur after the initial benchmark poll. Trend polls allow a campaign to gauge how well its messages are resonating with various portions of the constituency. As the election season moves into its final days, though, many campaigns move to tracking polls. Tracking polls generally use small samples and few questions, but they may occur daily, which allows a candidate to measure public opinion continuously. Tracking polls can help the campaign staff determine how respondents feel about various political advertisements, speeches, and policy positions. Given the expensive nature of conducting public opinion polling, most House candidates conduct tracking polls only during the last few weeks of the campaign.

A more problematic type of survey, appearing with increasing frequency in modern campaigns, is the push poll. Unlike objective public opinion polls, which seek to determine where respondents stand on candidates

or issues, push-poll questions intentionally provide participants with negative information about opposition candidates (or, on occasion, positive information about candidates on whose behalf they are being conducted). Motives for push polling vary. These polls are a useful way to test criticism of the opposition before investing in negative advertisements. They also tend to produce skewed survey responses, which may be important to a campaign that wants to claim publicly that it enjoys significant voter support. But probably the biggest reason why push polls are growing in popularity is that they provide a means to criticize opponents at a time when voters think they are hearing from an impartial survey organization.

Another important campaign resource is the focus group. These small-group interviews typically consist of 8–12 voters. A paid moderator leads free-flowing discussions with participants about a variety of topics important to the candidate. The role of the moderator is to steer the discussion without imposing a formal setting or skewing the answers, providing the campaign with a sense of how voters perceive the campaign. The moderator's task might be to test particular issue positions, to try out different ways of framing an issue, to evaluate how voters perceive the candidates, or to test out criticism of opponents to figure out which attacks raise the most ire. Focus groups are usually not representative of an entire constituency, given their small size. Indeed, sometimes campaigns intentionally select a skewed membership, perhaps to concentrate on swing voters or on a particular advertising demographic. Nevertheless, focus groups can provide very detailed and specific information unavailable from more scientific forms of survey research. Perhaps the most important trait of focus groups, from the point of view of political campaigns, is the rich detail they provide on voter reactions in an environment simulated to resemble the "real world" of political discussion. Dick Morris claimed that the Clinton reelection campaign often conducted focus groups to test the effectiveness of various political advertisements. They not only used focus groups to test the independent impact of their ads, they also tested which ads were the most effective responses to opposition critiques.

A new technique, used by both political campaigns and politicians, is the use of dial groups. Much like focus groups, dial groups are not representative of the entire population, but they can still provide information unavailable by other means. Participants are given "dials," which are small electronic devices with a wireless connection to a computer. Participants may be asked questions, or they may view political advertisements, speeches, or debates. These respondents turn the dial on the device, indicating their gut-level reaction to whatever they are witnessing, giving high numbers when they like what they are hearing and low numbers otherwise. Because respondents repeatedly adjust their responses over the course of the survey, dial groups offer real-time qualitative and quantitative information rather than just blanket judgments after the fact. Campaigns therefore receive detailed, instant feedback about the way audiences perceive different portions of a campaign event.

Use of Polls by Political Parties

Political parties once played a central role in campaigns, because they were the main mechanism for organizing a ticket and persuading voters to show up on election day to support that ticket. These critical mobilizing agents faded in importance over the course of the 20th century, however, and especially after World War II. Candidates increasingly constructed their own campaign organizations from funds provided by powerful political-action committees (PACs), and approached voters through high-technology advertisements rather than via door-to-door appearances by party workers.

To a certain extent, the parties restored their political importance in the 1980s by becoming brokers of cheap campaign consulting, including public opinion polling. Today, parties conduct and fund polls on behalf of the party candidates. In addition to benchmark, trend, and tracking polls, they also fund recruitment surveys. These surveys allow those considering a run for political office to examine the viability of their candidacies. Although political parties may split a poll's costs with the campaign for which it is commissioned, the parties have found ways of deflecting the financial burden away from the candidates. For example, the parties sometimes release poll results to the press, which means that the polling expenditures do not count as campaign contributions or expenditures.

In the 1990s, the parties began using polls and focus groups to shape appealing, unified political messages as a means of countering the localized nature of candidate-driven elections. Newt Gingrich, along with fellow Republican candidates during the 1994 congressional elections, used focus groups to assemble their "Contract with America" ideological package. Through various national surveys, the Republicans were able to construct a package of policies that were supported by popular majorities. The 1996 Democratic response, their "Families First" agenda, used polling data to support their initiatives in an attempt to regain some of the seats lost in the 1994 election.

Political parties use the same tools for measuring opinion as do candidates and government officials, i.e., large public opinion polls, focus groups, and direct-mail surveys. However, parties also traditionally conduct straw polls (often held among the party faithful) to determine the popularity of potential nominees for office. Political

parties and their candidates also use poll information to determine the most effective audiences for advertisements. A firm called Nielsen Media Research regularly reports on the viewership of various television programs; the report reveals not only how many people watch a certain program, but also certain demographic information about each audience. This is critical information to campaign managers who want to buy television time—that is, set up "media buys"—that will spread the campaign's message widely and target the right voters.

Use of Polls by the Media

Media organizations use several types of polls. Some of these polls serve little more than an entertainment function, whereas others genuinely assist with the news coverage of campaigns and elections. Call-in polls and Internet polls are examples of unscientific polls that have become common in recent years. The two sorts of polls are similar. Call-in polls prompt radio or television audiences to answer questions by calling a telephone number, whereas Internet polls provide the same opportunity to people who visit a Web site. Both call-in polls and Internet polls differ from legitimate public opinion surveys because respondents choose whether to participate. This lack of randomness means that the results will not be representative of the entire population. The resulting bias is likely to be especially large when these polls are taken for entertainment purposes, because individuals who participate tend to be those with the strongest opinions on an issue. Yet, despite the near uselessness of these polls as a source of political information, audiences apparently find the results interesting.

Media outlets conduct exit polls on election night, hoping to use that information to determine the winner of a particular election before the official tally arrives. The organization generates a representative sample of precincts, and then stations pollsters at each precinct during an election. To get a representative sample of those who turned out to vote, the pollsters select voters as they leave the precinct station. Usually participants receive a questionnaire to fill out and drop in a ballot box set up at the polling site. These questionnaires ask how the respondent voted in various contests, as well as a variety of demographic questions. The results of these polls allow making projections on election night. They also allow analysts later to probe how different social groups voted and to speculate on why particular candidates won or lost.

Exit polls are expensive. They take advance preparation and a large, well-trained staff. As a result, the major television networks and the Associated Press banded together in the 1990s to create a single group, Voter News Service (VNS), to conduct the exit polls. VNS has collected exit poll data since 1993, saving the partner television network and news organizations (ABC, CBS, NBC, FOX, CNN, and Associated Press) millions of dollars every election year. Although all of the sponsors receive access to the same data, they analyze the results individually. Until the 2000 election, exit polls primarily faced criticism from those who were upset with how the media used the exit data. News organizations sometimes broadcast their projections while voting was still underway, prompting critics to accuse them of swaying voter turnout. In particular, party activists worried that knowing presidential election results from eastern states might discourage voters in the western states from participating at the same rates as they otherwise would have.

By contrast, the methodology of exit polls received little scrutiny before the 2000 election debacle—when every media outlet, using exit poll data provided by VNS, originally called the state of Florida for Al Gore, then called the state for George Bush, before stating that the election was "too close to call." Flaws in the VNS procedure apparently caused the embarrassment, ranging from the method for estimating absentee ballots, to the method for selecting precincts, and on to the system for transmitting poll results to journalists. The number of actual absentee ballots cast, for example, was nearly twice the quantity predicted by VNS. The sampling method, meanwhile, apparently pumped up the number of votes for Gore. VNS initiated numerous reforms in the wake of the 2000 election. These changes will include updating VNS's computer systems, improving the statistical models used by the organization, and creating better measures of absentee ballots. However, VNS could not get the system working in time for the 2002 congressional contests, leaving the long-term fate of the enterprise in doubt.

Another media practice to receive criticism in recent years is that media outlets increasingly engage in "horserace journalism." This sort of coverage pays little attention to campaign issues, the sort of information that might aid undecided voters attempting to choose among candidates. Instead, it focuses on the likely outcome of an election, the sort of information sought by the large number of partisans who already know who they support and want to know if their horse is going to win the race. Preelection polls play a central role in this sort of coverage, because they provide the best information about who is leading a contest and why they are ahead. Critics note that projecting an election winner before the fact provides little useful social or political function, but might distort election results—for example, by preventing dark-horse candidates from attracting public support by getting their message out to voters through the news. Polls also tie up media resources that otherwise might support

investigative journalism, creating news rather than finding it. However, media polls show little sign of abating.

Conclusion

Politicians possess a much better resource for measuring public opinion than they have ever enjoyed before: the scientific public opinion poll. This powerful new tool allows modern governments to be much more responsive to voters than those in the past were. Polls have become such a common device in politics that, for some social scientists, they raise the specter of a hyperactive democracy, one that slavishly accommodates the whims of public opinion. Social critics bemoan the replacement of public leaders with public followers and worry that political discourse has become little more than an echo chamber of ill-considered popular attitudes. Nonetheless, polls have been shown to be invaluable for politicians seeking to obtain and retain public office, for political parties seeking to exercise influence over their members without the inducements of the past, and for political journalists seeking a source of news unmediated by the "spin doctors" who mastermind election campaigns. It is unlikely that a practice so useful to so many political actors will be allowed to fade in significance.

See Also the Following Articles

Census Undercount and Adjustment • Census, Varieties and Uses of Data • Election Polls • Election Polls, Margin for Error in • Polling Industry • Polling Organizations • Survey Questionnaire Construction

Further Reading

Eisinger, R. M. (2003). *The Evolution of Presidential Polling.* Cambridge University Press, New York.

Fiorina, M. P., Peterson, P. E., and Voss, D. S. (2004). *America's New Democracy,* 2nd Ed. Longman Publ., New York.

Geer, J. G. (1996). *From Tea Leaves to Opinion Polls: A Theory of Democratic Leadership.* Columbia University Press, New York.

Jacobs, L. R., and Shapiro, R. Y. (2000). *Politicians Don't Pander: Political Manipulation and the Loss of Democratic Responsiveness.* University of Chicago Press, Chicago, IL.

Morris, D. (1999). *Behind the Oval Office: Getting Reelected Against All Odds.* Renaissance Books, Los Angeles, CA.

Traugott, M., and Lavrakas, P. (1996). *The Voter's Guide to Election Polls.* Chatham House Publ., Chatham, NJ.

Wayne, S. J. (2000). *The Road to the White House 2000: The Politics of Presidential Elections.* Bedford/St. Martin's, Boston, MA.

Polling Industry

Frank Newport
Editor in Chief of the Gallup Poll, Princeton, New Jersey, USA

Glossary

census The calculation of the characteristics of a population based on measuring every member of that population.
direct-to-the-media polls Polls that have not undergone scientific peer review of prior publication in journals or outlets that provide quality control and that are published or released directly in media outlets as received from the polling organization that conducted the poll.
full-service polling companies Businesses that provide complete polling services, including development of the questionnaire, drawing the sample, interviewing, data analysis, and report writing and summary. These are usually differentiated from niche firms that focus on only one aspect of the polling process.
polling A process that uses scientific principles of probability sampling to obtain a representative sample from a population and then uses measure of attitudes from this sample to estimate population values on specific variables of interest. Polling usually refers to a process in which the results from the sample are released publicly.
probability sampling The process of selecting samples from a population in such a way that every member of the population has an equal or known probability of falling into the sample.

Any discussion of polling companies must start with a precise definition of the word "polling." Polling is often used in casual conversation to describe a process by which the views of a group of people are measured and tabulated. Most people have heard the term polling used in common conversation, for example, "poll the jury." We also use the word "poll" to describe the place where we vote (i.e., "go to the polls on election day").

Introduction

In this article, the term polling is used in reference to a process that is more specific than the typical generic use of the term. The definition of polls focuses on a process that has three key parts: (i) the attempt to measure human attitudes or opinions, (ii) the effort to use sampling techniques to get those measures from a small group of people and generalize them to a large population, and (iii) the public release of the results so that everyone in the population is exposed to them.

The first of these components is straightforward. One can in theory conduct a poll of breeds of dogs represented in a community or the types of minerals represented in a geologic area. Here, discussion is confined to research involving human attitudes and opinions.

The second point in this definition is particularly important because it differentiates polling from more general processes of survey research. It is possible to measure the opinions of every member of a population under consideration. This is commonly called a census, and it is an easy process when the population of interest is small (e.g., a school or a business) and an extremely complex process when the population of interest is an entire country (the best example of which is the biennial census in the United States). However, censuses are not polls under the definition used in this article because they do not include the process of sampling and generalizing from that sample to a population.

The third point is key in terms of the discussion of companies that engage in polling. There is a very wide group of procedures that have as their object the measurement of human attitudes and that use scientific random sampling techniques to efficiently generalize to a large population. However, many of the results of these procedures are not released to the public but are instead analyzed and used only in confidential privacy

by the organization commissioning them. These come under the rubric of "market research" and are outside of the purview of this article. We are focusing here only on the process by which survey research results are released publicly after they are gathered and analyzed.

History

The widespread use of polls in the United States began in the 1930s when several men widely recognized as pioneers of the polling process began to apply the principles of random sampling to human attitudes. These men included George Gallup, Elmo Roper, and Archibald Crossley. Later, two other pioneers in this field, Louis Harris and Daniel Yankolovich, joined this roster of prominent pollsters. Most of these pollsters founded full-service polling companies. The Gallup Organization, for example, was a very successful and profitable company focusing primarily on providing polls to newspapers from the mid-1930s to World War II. Firms carrying the names of Gallup, Roper, Harris, and Yanklovich still exist today. However, it is important to note that none of these firms focus exclusively on polling today. All conduct publicly released polls as part of a much broader set of research services.

This transformation is symptomatic of the general trends that have affected the polling industry during the past half-century. There is little question that the promise of polling and its value to society have been realized successfully—as evidenced by the ever increasing number of polls that are reported in the news media and thus are available to the citizens of the countries in which they are conducted. However, the value of the polling enterprise to provide a foundation for a commercial, for-profit business has been much less successfully realized.

A significant challenge facing the polling industry has been the search for ways in which polls can be conducted, analyzed, and reported in a manner that provides incentives for profitable, creative, and entrepreneurial companies. This challenge has not been successfully addressed. As a result, the polling industry today is not the vibrant collection of fast-moving competitors that characterize other business sectors in which there has been rapid innovation and advancement. Rather, the polling industry (at least in the United States) has evolved to the point where most polling is conducted in one of the following ways: (i) as a ancillary service provided by firms mainly focused on providing market research for business and industry; (ii) by nonprofit government and educational institutions; (iii) by niche providers that focus only on one small aspect of the polling process; and (iv) by companies attempting to use methodologically slipshod techniques involving the Internet or robotic, impersonal telephone calling to provide high-volume, low-cost polling. The economics of polling has to a large degree dictated that traditional, for-profit polling companies have a very difficult time surviving.

As noted previously, this is not to say that there is a diminished demand for polls. Most observers agree that there are more polls released into the flow of media news coverage today in the United States and throughout the world than ever before. This is particularly true, for example, before a presidential election in the United States, or when there is a major news story about which the public's opinion is germane. Part of the explanation for the increased number of polls has to do with the increased number of news outlets reporting news, particularly cable television news channels and Internet news sites. However, there also seems to be a genuine increase in the thinking of media gatekeepers (i.e., newspaper editors and television producers) that polls are a legitimate and interesting component of news coverage.

When we look at polling from a business perspective today, we find a paradox that results from the difference between this demand for polls, on the one hand, and the inability of this demand to support a vibrant commercial polling industry, on the other hand. We can examine the reasons behind this paradox by discussing in more depth at the types of organizations willing to commission and pay for polls.

One type of group or organization interested in sponsoring polls consists of nonprofit foundations and educational and government entities. These organizations sponsor and pay for polls because they believe polls are in the public interest and help fulfill the purpose for which the organization was founded. (In this sense, polling follows the same model as nonprofit public radio, sponsored by government and charitable contributions for the public good.) The Philadelphia-based Pew Foundation, for example, invests each year in sponsoring a significant polling program through the Pew Research Center. Its polls are released in their entirety, free of charge, to anybody who wants them. The Harvard School of Public Health, along with the nonprofit Kaiser Foundation, sponsors a series of polls asking Americans about their health and health-related issues. Several small colleges, most notably Quinnipiac in Connecticut and Marist College in New York, sponsor polls of their region in an effort to attract recognition and brand awareness. The University of New Hampshire conducts polls of its home state, as do a number of other state universities. Nonprofit advocacy or charitable groups will also sponsor polls that focus on their particular issue or concern.

Relatively few of these nonprofit entities pay full-service prices to outside firms to conduct the polls. Many conduct the polls themselves. Others work out

an arrangement by which the cost of the polls is shared with research organizations interested in providing them.

A second and very important group of organizations that sponsor polls consists of news and media outlets that support polling programs primarily because they believe polls add an important element to their news coverage and/or because they believe that having their name attached to polls will increase brand awareness and legitimacy. This use of polls by news and media outlets is a time-honored tradition. George Gallup, who founded the Gallup Organization, initially released his polls through a syndication service that provided poll results, analysis, and interpretation to newspapers in a ready-to-use format. Fortune Magazine was one of the original sponsors of Roper Polls. In today's news environment, almost all major news organizations sponsor polls, including CNN/USA Today (with Gallup), ABC/Washington Post, CBS/New York Times, Fox News, NBC/Wall Street Journal, the Los Angeles Times, Time Magazine, Newsweek Magazine, and the Associated Press.

However, these media polls are not numerous or frequent enough to sustain an industry of polling companies focused just on providing them to news and media outlets. Additionally, the media organizations that sponsor these types of polls have developed a series of ways of keeping the costs low (discussed later).

A third type of organization that sponsor polls consists of "normal" for-profit companies that commission polls primarily because they believe that having their name attached to the results when they are publicly released is good business and makes sense from a "branding" viewpoint. (This latter motivation can be called the "Goodyear" motivation because it reflects the same rationale used by the Goodyear Corporation in sponsoring its famous blimps.) These types of polls are often associated with the companies' areas of interest, based on the hope that either the results are favorable to their business objectives or the association of their name with topics related to their business will generate name identification and legitimacy.

Why Polling is not Profitable

There are a number of specific reasons why the demand for polls as represented by these three types of organizations has not led to the development of a large and thriving industry of polling companies. Many polls are conducted in-house by the sponsoring organizations, much as would be the case for organizations that use in-house legal or accounting services. This situation is particularly true of government and educational organizations, as well as for some news and media outlets that have ongoing polling programs. These entities have developed self-contained polling units and therefore the ability to subsume polling into their internal business structures. The New York Times/CBS polls, for example, are conducted using in-house polling operations supervised by employees of these two sponsoring media outlets. A number of colleges and universities in the United States conduct polls within institutes or academic departments, in some instances as part of an educational or learning experience. This is true for the Quinipiac Poll, the University of New Hampshire polls, and the Marist College polls. The implications of these procedures for the polling industry are quite obvious. In-house polling lessens the need for private, independent polling organizations.

Even when outside polling companies are used to conduct polls, however, there are several factors that mitigate against the development of a robust and profitable polling industry. First, conducting publicly released polls is an attractive proposition to many research companies because it helps their business to be associated with news stories potentially read and seen by hundreds or thousands of people. In other words, many research organizations are quite eager to get involved in the business of providing polls because of the valuable increase in name identification that results. The supply is higher than the demand, and this makes a business focused just on conducting polls less attractive for all but companies that can provide polls in a high-volume, low-cost way or to conduct polls as a sideline to other businesses.

A second factor results from the specific nature of the typical news poll or other polls conducted for brand identification purposes. These types of polls do not require a great deal of consulting or other value-added services from the organization conducting the poll. The sponsoring organizations may need only the bare essentials of the poll process, eliminating the demand for the less tangible "intelligence" services that are more profitable for full-service polling companies. This in turn diminishes the ability of polling companies to charge for the typical value-added consulting, analysis, and interpretation that provide high margins in research.

These factors have created a situation in which companies providing polling services today often do so as a sideline to other, more profitable businesses, much of the time as a loss-leader to generate attention and increase name awareness. On the other hand, companies interested in polling have focused on the attempt to harness the Internet and other technologies to provide polls on a high-volume, low-cost basis. Some of these companies have attempted to conduct polls entirely without human intervention using telemarketer-type calling technology coupled with recorded voices and the capability to allow respondents to answer questions via the touch-tone pad on their phone. Other companies attempt to gather large numbers of individuals who have e-mail addresses and then to persuade these individuals to follow through and complete online polls.

There is another factor that limits the potential profitability of polling companies that attempt to provide polls as their primary business. There is usually not much of a direct-to-the-bottom-line payoff to organizations or entities that sponsor polls. Much market research (i.e., survey research whose results are not released to the public but used in a proprietary fashion by the organization or entity that commissions the research) is of great value to the commissioning organization because it can be used by it to achieve a business goal. This includes research relating to brand names, product characteristics, and the positioning of political candidates. Companies and political campaigns are willing to pay high prices for the provision of this type of information in the same way that they pay for strategic consulting and positioning studies because these organizations can ultimately find a relationship between the recommendations developed as a result of the research and revenues and profits. This unique value provided by marketing research gives companies an incentive to seek out and pay for the best possible research services, including research that is of the highest quality and that involves higher level consulting and expertise.

There is a key differentiating factor that prevents polling companies from justifying these same high costs. Poll results by definition are released to everyone and therefore have no unique or proprietary value to the commissioning entity. The commissioning organization cannot gain a unique advantage or a significant business edge based on poll results that are essentially available to anyone who wants them, including competitors, regardless of who pays for them. This lack of a direct connection between polls and an organization's bottom line puts even more pressure on an organization interested in commissioning polls to control costs and obtain polling services at the lowest possible price. This in turn compresses the return on investment for polling companies and restricts the viability of polling as a stand-alone business.

There are business benefits other than revenues and profits that accrue to organizations that sponsor polls. Many of these benefits are based on the value of legitimacy and name identification that comes from being associated with polls, the use of polls as a component of news coverage, or polls as part of an educational or governmental function. However, the fact that poll results are by definition shared with everyone severely limits the number of business organizations willing to pay for polls on a systematic or continuing basis and, more important, limits the amount of money organizations are willing to pay for polls. Thus, the situation today is one in which there is a considerable demand for and interest in polls but a limited willingness or capability on the part of those who have this interest in paying for them.

In summary, several major factors that affect polling as an industry today have been reviewed. First, the fact that many organizations conduct polling in-house restricts the need for outside polling companies. Second, organizations commissioning polls do so as either a public service or as a means of building name awareness and burnishing the company's image, neither of which has the direct-to-the-bottom-line value of proprietary market research. This creates a situation in which organizations often have more limited budgets for polling and seek out lower cost polling providers. Third, conducting publicly released polls has intrinsic value to many research companies, which means that there are usually more research companies interested in providing polls than there is demand. All of this results in a situation in which the business of polling is not one that has, to date, developed into an industry with strong revenues and high profitability. This has kept the number of companies interested in doing nothing but providing polling services at a quite low number.

Differentiation of Polling Services

As has been the case in many industries over the years, one of the major changes in the field of polling has been the differentiation of the polling process into discrete phases, each of which can be executed by specialized companies focused just on one aspect of the process. These companies operate profitably in a particular niche by providing their service to a wide variety of research and business clients. The fact that firms can provide these component services allows those interested in conducting polls to in essence "assemble" a poll by obtaining each part of the poll from a different vendor. A random sample of the adult population can be purchased from a sampling company, field interviewing services can be commissioned from a high-volume company that does nothing but telephone interviewing, and basic analysis of the polling data can be obtained from a company that does nothing but such data tabulations. Each of these niche players provides these services to market research clients as well as polling clients and sustains a profitable business in that manner. The ability of organizations interested in commissioning a poll to contract with these individual vendors provides a further disincentive for companies to enter the polling arena as pure, full-service polling providers.

Thus, there are a number of forces at work that limit the ability of companies to sustain a full-time business doing nothing but providing poll services. Although there are more publicly released polls today than at any other point in the polling industry's relatively brief history, the total number of polls is still small enough to limit the number of businesses needed to provide them. Some organizations that sponsor polls end up doing the polling work in-house, further limiting the need for outside polling companies. There are name identification and

branding benefits that accrue from conducting publicly released polls that create a situation in which companies compete to provide polls at very low margin rates. Most organizations that sponsor polls do not need the value-added consulting services that are often the most profitable components of the survey research process. The polling industry today has many niche companies that can provide specific elements of the poll process, thus allowing sponsoring organizations to organize a poll by contracting separately for each of its components.

All these factors have opened the polling industry to low-cost providers that attempt to compete by using the Internet or other high-technology techniques to obviate the cost of live human interviewers. Additionally, many of the companies that conduct polls are full-service market research companies whose main business is providing more lucrative and remunerative services in addition to polling—primarily market research services for business and industry. These companies often provide polling services as an adjunct to these businesses, sometimes as a way of obtaining brand identification. A market research company can become widely known as a result of providing a highly visible publicly released poll, helping the company establish legitimacy in the eyes of potential commercial clients.

In summary, the nature of the polling business today has created a situation in which there are few for-profit polling companies that do nothing but provide polling services. The landscape of companies involved in polling consists of the following: (i) nonprofit educational and government entities that conduct polling in-house; (ii) news and media organizations that conduct polling in-house; (iii) full-service research companies that conduct polling as a low-margin side business; (iv) companies that provide specific components of the polling process, such as samples, field interviewing, or data processing; and (v) companies focused on provided commodity such as polling services using the Internet and other high-technology methodologies.

The Business of Polling and Quality Concerns

The structure and dynamics of the way in which the polling industry is configured today have repercussions on the overall level of consistency and quality of the polling that is released to the public. In general, the "stepchild" nature of polling results in a situation that leads, in some cases, to less than optimal quality.

As reviewed in the previous section, there are few full-service firms that provide polls as the major function of their business. This means that the polling industry misses out on specific benefits provided by larger firms concentrating just on polling. Full-service polling companies that conduct polls from beginning to end provide the positive benefit of quality control and a coherent monitoring of how the pieces of the polling puzzle fit together. In some ways, this is similar to the process of building a new house. A homeowner can contract separately with an architect, cement foundation specialists, carpenters, electricians, plumbers, roofers, and so on, but the odds that the results will be a well-built house are lower than if an experienced contractor is in charge of the entire process and supervises it from beginning to end. The fact that there are few full-service companies that provide polling services has led to a general diminishment in the standards of control of polling that would exist if full-service companies were more involved in the process. This is particularly true in the case of polls that are purchased from low-cost providers or assembled from the contributions of a series of different companies. This is little different from the situation that develops in any industry in which there are intense pressures to provide products or services in the cheapest manner possible.

There is another aspect of the way in which the polling industry is set up today that gives rise to concerns about poll data quality. This has little to do with the cost structure of the polling industry but, rather, is a result of the way in which poll results are released to the public. The vast majority of polls to which the public is exposed in print or broadcast are not screened, reviewed, or published in refereed journals before they are used by the news media. For the most part, this is a result of the fact that a good deal of polling is conducted primarily for immediate release to a lay audience through print and electronic media means, without moving through any sort of the more traditional scientific process.

This direct-to-the-media nature of the polling business includes survey research that is funded and conducted by media organizations (such as the CNN/USA Today/Gallup poll and the New York Times/CBS poll) and polling that is conducted by others with the primary intention of gaining release through the media (rather than via scholarly publication). The purpose of these polls—that is, polling for direct dissemination to a lay audience—may include understanding and accumulation of theoretical knowledge, but the parallel, latent, or, in some instances, overt motivation often becomes one of generating and sustaining reader and viewer interest. The major purpose of scientific polling intended primarily for a scientific audience, in contrast, is the accumulation of knowledge.

The nature of direct-to-the-public polling means that it is often released within days or even hours of its completion. There is no peer review and journalists have little opportunity to review the research before it moves into the news stream. Journalists have a higher probability of being guilty of inaccurate or incomplete reporting when they summarize polling that is handed to them

by nonscientists intent on getting it published, when they summarize polling that they commission themselves, and/or when they report polling that is conducted without having gone through a scientific or scholarly screening process.

It could be argued that direct-to-the-media polls actually simplify matters because the needs of the media are directly taken into account when the polling is done. In other words, with the encumbrances of the need for scientific reporting not an issue, the execution and reporting of polls in "media-ready" form can be done much more quickly and in a fashion that is more targeted to reporters' needs.

However, this simplification comes at a cost. In the drive to meet media requirements and to gain access to media channels, pollsters may run the risk of creating more superficial, selective, or incomplete research and reporting than they might otherwise.

Scientists too have desires for public recognition of their work, and competition in science often results in rushed research and can distort truth. Scientists also can conduct bad, sloppy, or poorly designed research and have problems when their conclusions are not dramatic or easily summarized. However, a scientific orientation at a minimum encourages a desirable motivation for the research: adding to the accumulation of knowledge. The criteria for media use are different. Media polls are focused more on adaptability to the media's format and the interest value of the research report to the print or broadcast audience. The need to get results to the public quickly and easily can supersede the need to present results that are accurate and complete. The probability of less than optimal reporting of polling results can increase significantly in these situations in which research is conducted without the limiting context provided by scientific context and motivation. The balance, in short, gets out of whack.

Of course, direct-to-the-media polling does not always avoid or ignore scientific procedures. A good deal of this type of polling is built on a foundation of science and the scholarly approach to the accumulation of data. Many of the researchers who work for the media and report polling data directly to the public are highly trained and experienced polling professionals. The issue, however, is that there are no universal requirements or standards of quality in these situations, and thus what the public receives in these direct-to-the-media polls can suffer.

Summary

There are few "pure" polling companies left in existence today, at least in the United States. The structure of the polling industry as it has evolved during the past 70 years makes it extremely difficult for companies to focus just on conducting polls and at the same time maintain a profitable business. As a result, a good deal of polling is done as a sideline to other business and/or is commissioned by organizations searching only for the cheapest, bare-bones way to conduct the research. Some polling is assembled by purchasing the various components of the polling process from niche providers. The nature of the polling industry has also opened the door to a variety of low-cost, high-volume providers that conduct polls using the Internet or computer-based telephone-calling equipment. Most polling is conducted and reported through the news media without the peer-review process that is associated with other scientific work, sometimes within hours or days of being completed.

These facts, coupled with the increased demand for polling today, have resulted in a situation in which the quality and validity of poll results released to the public are often of a lower level than would be desired.

See Also the Following Articles

Census Undercount and Adjustment • Census, Varieties and Uses of Data • Election Polls • Election Polls, Margin for Error in • Politics, Use of Polls in • Polling Organizations • Survey Questionnaire Construction

Further Reading

Asher, H. B. (2001). *Polling and the Public: What Every Citizen Should Know.* CQ Press, Washington, DC.

Bogart, L. (2000). *Commercial Culture: The Media System and the Public Interest.* Transaction Publishers, Somerset, NJ.

Eisinger, R. M. (2003). *The Evolution of Presidential Polling.* Cambridge University Press, Cambridge, UK.

Lavrakas, P. J., Traugott, M. W., and Miller, P. V. (eds.) (1995). *Presidential Polls and the News Media.* Westview, Boulder, CO.

Mann, T. E., and Orren, G. R. (1992). *Media Polls in American Politics.* Brookings Institution, Washington, DC.

Moore, D. W. (1995). *Superpollsters.* Four Walls Eight Windows, New York.

Polling Organizations

Jeffrey A. Fine
University of Kentucky, Lexington, Kentucky, USA

D. Stephen Voss
University of Kentucky, Lexington, Kentucky, USA

Glossary

completely automated telephone surveying (CATS) A polling technique in which an automated voice interviews participants and a computer directly processes the answers, by interpreting responses punched into a touch-tone telephone, for example.

computer-assisted telephone interviewing (CATI) A polling technique that prevents human error in the sampling and interviewing because phone numbers identified by the sampling frame are directly computer dialed and interviewers are walked through the questions in the survey. CATI systems also add flexibility by randomly assigning questions to a subset of respondents or by randomly ordering possible responses to a question.

dial groups Small groups of individuals who are given handheld electronic devices that allow them to register their ongoing feelings about a speech, debate, or political advertisement while they are being exposed to the message; may not be representative of a constituency or audience, but can provide detailed information about the types of people in the sample.

Literary Digest An influential periodical that successfully predicted presidential election outcomes from 1916 through 1932, using a survey that sampled millions of telephone users and car owners; collapsed after a terribly wrong prediction for the 1936 election caused by the class bias in its sampling frame.

Nielsen ratings Statistics generated by Nielsen Media Research that report the overall size of the audience for broadcast programs. Not only do these ratings help determine advertising costs for television programs, the company's data on media audiences also allow both commercial and political advertisers to target their messages efficiently.

polling alliance A cooperative effort among disparate news organizations to conduct polls; allows control over survey methods yet keeps costs down compared to individually organized polling efforts. Generally unites news organizations from different media formats (for example, newspapers with either magazines or television networks).

quota sampling Selecting poll respondents systematically to ensure that a sample matches the demographic characteristics of the larger population.

sampling bias Polling data slant that results from using improper or unscientific methods to select the people whose opinions will be sampled. Not to be confused with "sampling error," which is not a form of bias but instead results naturally when using small numbers of people to estimate the opinions of a larger group.

straw poll Although used to refer generally to a poll with a large but unscientific sample, the main usage in politics applies to surveys conducted among the membership of a political party. Due to the sampling methods of a straw poll, results cannot be generalized to the entire population, but are useful for determining which issues, messages, or candidates can motivate the party faithful.

Public opinion polls are a complex and costly enterprise. Although they have become ubiquitous in contemporary society, polls did not always dominate social discourse in the way they do now. Rather, early polling companies got off to a rocky start, including embarrassing failures when they tried to predict presidential election winners in 1936 and 1948. However, survey practices have been institutionalized in the form of professional polling organizations. Most of these organizations focus on marketing research, and some of the highest profile organizations

generate large quantities of political and social data. Regardless of their prime function, though, polling organizations have helped establish stable survey methods and have also been leaders in the innovation of better methods. The wealth of data produced using modern, scientific polling techniques has become an invaluable resource for both journalists and social scientists.

Market Research Organizations

Market research companies regularly studied public preferences early in the 20th century, decades before academic polling became prevalent. Such marketing organizations exist in large numbers today. They conduct surveys on behalf of clients interested in the kinds of products and advertisements that appeal to consumers. Although product development and consumption patterns sometimes interest social scientists, this is not really the focus here. Multitudes of such organizations do exist around the world, and many of these businesses consist of little more than small operations squared away in tiny shopping-mall offices, grabbing respondents from among the pedestrians passing before their storefront.

Marketing research, however, has played an important role in the development of polling, beyond its function of probing consumption patterns. The early marketing firms that survived by helping clients increase sales held a significant stake in getting the answers right. The profit motive encouraged marketers to develop sophisticated and successful polling techniques. Eventually, their methods trickled down to other forms of social research, often because individual pollsters carried their expertise from the commercial world to academia, to research institutes, or to specialized polling firms. George Gallup, who would later become perhaps the most widely recognized pollster, got his start performing market research for the *St. Louis Post-Dispatch*, conducting interviews with subscribers to inquire about the columns they liked to read.

Before Gallup began conducting survey research on public issues, political polling consisted almost entirely of election forecasting conducted by marketers. Marketing organizations used political polls as a form of publicity. They would project election results as a high-profile way of showing off their techniques. For example, market research analyst Archibald Crossley established Crossley Incorporated in 1926. This organization, though predominantly involved in commercial survey research for private corporations, nonetheless regularly issued election projections. Crossley's ability to predict the correct outcomes, at times more accurately than Gallup or Elmo Roper, demonstrated that market-research firms could conduct political polls validly. The effects of Crossley's work are still seen today, in that many political actors and organizations still use marketing firms for much of their consulting needs.

Early Techniques for Measuring Public Opinion

Before the birth of the modern poll, public opinion research was labor intensive, time consuming, costly, and often inaccurate. The most common techniques for measuring public opinion were various types of "straw" polls. The purpose of these polls was to determine how citizens would vote in upcoming elections. The polls (votes) were taken in various gatherings, from meetings called specifically to gauge voter intentions to meetings of militia or political parties. To get a sense of the overall public opinion in any election, those interested this information would need to conduct straw polls regularly, during a variety of different public meetings. The effort required a great deal of time and manpower. Due to the extreme time requirements and labor constraints, straw polls seldom produced meaningful results; their samples were not representative of the electorate and the methodology for conducting this research was inconsistent.

Literary Digest, an influential periodical, developed the most famous political poll of the early 20th century. The magazine's methods initially seemed very accurate, primarily because it collected a massive sample that dwarfed the numbers found in straw polls. *Literary Digest* sent out tens of millions of questionnaires to people all over the United States. The surveys were mailed to people whose names appeared on automobile and telephone registration lists. As long as political behavior did not correlate strongly with access to cars or telephones, the method worked. Indeed, using these techniques, the *Literary Digest* poll correctly predicted the winner of every presidential election from 1916 to 1932. The New Deal political realignment after 1932 changed party politics, however. By 1936, Republicans were notably more likely than Democrats to own luxuries such as telephones or cars, so the magazine's days as a preeminent election prognosticator were numbered.

Rise of the Representative Sample

Survey specialists had criticized the *Literary Digest* poll before 1936. Informed social scientists and market research analysts recognized the methodological problems with the sampling methods used in the poll. Among these critics were George Gallup, Elmo Roper, and Archibald Crossley, all of whom instigated scientific opinion polls in 1936 to demonstrate the flaws of the famous *Literary Digest* effort. Gallup in particular strongly believed that

the use of sophisticated, scientific polling techniques would lead to a more-accurate portrayal of electoral preferences. During his time at the University of Iowa, Gallup conducted careful research on the attributes that helped department store sales representatives achieve success. His doctoral dissertation looked at the reading habits of newspaper subscribers, using a sample of approximately 1000 consumers. In this dissertation, the approach followed, eventually known as the "Gallup method," was in essence the same set of techniques that Gallup would employ when he began political polling.

In 1935, Gallup and Harry Anderson founded the American Institute of Public Opinion (AIPO) with the purpose of accurately determining public opinion by the use of representative sampling. Gallup incorporated the technique of quota sampling, common to market researchers, into his methods for political polling. Gallup and Anderson selected participants proportionally based on demographic information about the population. Each interviewer would receive particular quotas when identifying respondents: a certain balance of men and women, a certain balance of employed and unemployed, etc. This technique ensured demographic representativeness, but it did not ensure accurate social measurement, in part because it allowed interviewers wide discretion in identifying the respondents who appeared in the sample. Nevertheless, the 1936 election showed both the capability of these methods—even when used to collect modest samples—and the inadequacy of large but unrepresentative samples. *The Literary Digest* returned with its massive polling apparatus that year, using the same automobile and telephone registration lists that it had used in previous elections. The poll predicted that Republican candidate Alfred Landon would defeat Democratic candidate Franklin D. Roosevelt (FDR) by almost 20% of the vote. Of course, FDR went on to win more than 60% of the popular vote, which translated into 523 Electoral College votes. His victory dealt an immediate deathblow to the *Literary Digest* poll.

Scientific opinion polls proved more accurate in 1936, despite their much smaller samples. Gallup not only forecast FDR's victory over Landon, for example, he predicted that *The Literary Digest* would incorrectly predict the outcome. Gallup knew that *The Literary Digest* would make the wrong prediction due to sampling bias. The Gallup group found that higher income voters were more likely to support Landon, whereas lower income voters favored Roosevelt. The names on the automobile registration and telephone owner lists were those of wealthier and more conservative people than represented the population as a whole. This meant that the magazine's sample would not be representative even though it included over 2 million respondents. Although Gallup correctly forecasted the winner of the election, his prediction still missed the vote totals by approximately 7%. Some have attributed this error to the imperfections of the quota sampling techniques employed by Gallup at the time, requiring the sample of respondents to mirror the demographic distribution of the entire population. This method led to selection biases on the part of those conducting the surveys.

Gallup and Roper became famous for their ability to create polls that were representative of the population. Gallup also became an important voice in the debate over public opinion polls. In a 1940 book, *The Pulse of Democracy*, Gallup and Saul Rae asserted that the desire to measure public opinion on political issues was motivated by a drive to improve American democracy. In particular, they argued that representative samples would empower politicians, making them "better able to represent" than they had previously. Polls would replace "the kind of distorted picture sent to them by telegram enthusiasts and overzealous pressure groups who claim to speak for all the people, but actually speak for themselves." Public opinion research was promoted as vital to the health of a democratic political system, reforming a political system dominated by special interests by providing a stronger voice for "the people." Political polling would send public officials information directly from the voters so that the data would not have to be mediated through interest groups. Elected officials hearing the results of these polls would know that catering to special interests could cost them their political careers. However, just as the election of 1936 gave Gallup and his methods instant credibility, the close election of 1948 constituted a major setback to public perceptions of polling companies. Gallup and other pollsters, using the same quota sampling methods they had used in previous elections, predicted that Republican Thomas Dewey would defeat incumbent President Harry Truman. Though Gallup's predictions for 1948 actually were more accurate than the 1936 poll predictions in terms of each candidate's percentage of the vote, he forecast the wrong winner. The public image of polling companies fell as a result of this embarrassment. Gallup, Louis Harris, and Burns Roper would continue to conduct polls after 1948, hoping to regain public faith in their research methods, but their image did not recover for several years.

Modern Polling Organizations: Faith in Public Opinion Rebounds

During the 1950s and 1960s, Louis Harris conducted private polls for over 240 political campaigns. Harris served as a pollster for President John F. Kennedy, conducting numerous surveys for the administration about policy preferences and presidential approval ratings. The widespread use of polls by high-profile and

successful clients helped gain enough media attention to restore some of the public confidence in the industry. It became clear that pollsters could accurately gauge public opinion using scientific polling methods; some observers therefore consider the 1950s and early 1960s the modern era of political polling.

The polling companies that were founded by Gallup, Roper, and Harris began a period of enormous growth. These pollsters became common staples in political campaigns and in the daily routines of elected officials. Their survey agencies sold increasing numbers of reports to governmental actors, private corporations, and media organizations. They stepped up their polling on the hot political issues of the day, as well as on the public approval of the president. They also continued their tradition of accurate election forecasting. The overall enterprises are now massive indeed. Today, for example, the Gallup Organization conducts countless surveys every week, serving both the market research and political polling segments of the industry. Gallup gauges public opinion on various issues of the day, from controversial Supreme Court cases, to public sentiments on international wars and crises, to how Americans feel about the job performance of the president.

Eventually, numerous rivals joined Gallup, Harris, and Roper in the industry; prestigious examples include the Chilton Research Associates, the Princeton Survey Research Associates, the Pew Research Center, Yankelovich Partners Inc., and Rasmussen Research Group (now called Portrait of America). These companies generally receive less attention than the big three receive, but they play an equally important role in social measurement because their lower profile allows flexibility in techniques or question wording. For example, Rasmussen has experimented with ideological measurement that is more complex than the simple liberal/conservative scales customarily used in political polls. Such innovative poll results can receive significant news coverage.

In addition to the independent polling organizations, there are also numerous archives of public opinion data. These archives have made public opinion data widely available and are important sources for scholars and researchers alike. Among the largest and most frequented archives are the Roper Center for Public Opinion Research (located at the University of Connecticut), the Interuniversity Consortium for Political and Social Research (located at the University of Michigan), and the Odum Institute for Research in Social Science and the Louis Harris Archive (both located at the University of North Carolina at Chapel Hill). These smaller level polling companies also play a critical social role because they produce or develop services that otherwise would be unavailable to politicians and private industry. For example, government bodies such as the U.S. Department of Education hire lower level polling companies to conduct accurate surveys on behalf of the public. Other companies sell software and equipment that allow politicians, political parties, and campaigns to conduct "dial groups." This polling method allows political actors to gauge responses and opinions toward different aspects of speeches, debates, or advertisements while the event is still going on. In coming years, presidents may choose to use this technology to monitor responses of voters to portions of the State of the Union address, or candidates for political office may use dial groups to see what parts of a political speech or political debate particularly resonated with their audiences.

Not all of the competitors to Gallup, Roper, and Harris developed as for-profit polling companies. Government agencies, nonprofit organizations, and public-interest media outlets also started conducting important social research using opinion polls. For example, groups of news organizations and media outlets have teamed up to form major polling alliances. Some of the more well-known polling alliances are between television networks, newspapers or newsmagazines, and pollsters: ABC/*Washington Post*, CBS/*New York Times*, NBC/*Wall Street Journal*, *Time*/CNN, *USA Today*/CNN/Gallup, and Gallup/*Newsweek*. These alliances provide journalists greater direct input on the sampling methods used in polls, as well as allowing them relatively quick access to the data, but still allow individual organizations to share costs with others. Cheaper polls permit the news organizations to conduct large surveys more frequently. Their efforts have greatly expanded the overall body of social statistics available to analysts.

Large polling organizations are associated with academic institutions. Among these types of organizations are the National Opinion Research Center (NORC), associated with the University of Chicago, and the Survey Research Center (SRC), associated with the University of Michigan. NORC's clients include governmental departments and agencies, nonprofit organizations, and private corporations. *U.S. News and World Reports* commissions NORC to conduct some of the polls necessary for their various national rankings, such as their ranking of "America's Best Hospitals." The Survey Research Center has been conducting social science research for over 50 years and has been commissioned to conduct countless studies of public attitudes, beliefs, and values, including some of the seminal works on voting behavior.

Organizations do not necessarily need to gauge public opinion to be considered a polling company. The Nielsen Media Research organization provides valuable information about television audiences and viewers based on research conducted in every major television market in the United States. A "box" attached to a television set records each program that a household watches, as well as any changing of channels that occurs. The box also monitors which individuals in a household are viewing each program by allowing them to "log in" on a special remote

control. The resulting data, detailing which kinds of people are watching which television programs, allow both private companies and political actors to target specific audiences with advertisements. For example, corporations may wish to aim their ads at consumers who are most likely to purchase a given product, so they will select programs with a large audience among the target group. Political campaigns similarly target their ads at particular constituencies, such as swing voters or particular voting blocs likely to respond to a political message. The "Nielsen ratings," which report the overall size of each program's viewership, influence how much stations charge to air either corporate or campaign advertising.

Conclusion: The Future of Polling Organizations

A vast majority of polling resources are spent on consumer research. In fact, approximately 95% of all polling is conducted on market research. Nevertheless, the most widely publicized polls are still those conducted about political issues and figures. Polling information is almost seamlessly infused into daily news coverage and the national discourse on heated issues.

As technology has improved over time, the methods employed by pollsters and polling companies have changed. Early mail-in surveys gave way to face-to-face polling, which has yielded to telephone surveys. Within the modern era of random-digit-dialing (RDD) techniques, telephone polls dominate among the polling organizations. The methods have improved with time. Current polling companies have access to enormous telephone databases and statistical packages that allow those conducting the surveys to generate a random sample for the entire country (or any segment of the entire population), and analysis of the resulting data is easily done. Some polling companies use a new wave of computer techniques in collecting their telephone surveys. Computer-assisted telephone interviewing (CATI) and completely automated telephone surveying (CATS), relatively new innovations in telephone polling, facilitate the process of collecting public opinion data. With CATI, telephone numbers are generated randomly and are dialed by a computer that also displays the questionnaire on a computer screen for the interviewer. Questions may be randomly assigned to some respondents but not to others, and answers may appear in a random order so that respondents are not systematically influenced by where any one answer appears in a list. The CATI system then allows an interviewer to enter responses directly into a database, reducing the risk of human error during the data manipulation stage. With CATS, respondents interact with an automated voice that administers the entire survey. Neither technique is used universally, but both have become more common. In the coming years, technological advances will continue to shape the way that pollsters gauge public opinion on political issues and consumer interests.

Though polling methods have evolved as technology has improved, the pursuits to obtain a representative sample and an accurate reflection of the population have remained the same. Some popular polling methods, however, may not be so scientific. Internet polls are inexpensive and convenient to conduct, but respondents decide whether to participate (or, for that matter, whether to use the Internet). This selection process increases the risk of biased data; responses gathered in such a fashion cannot be generalized to the entire population. However, some scholars believe that the next technological innovation in the polling industry will be the advent of representative sampling methods for Internet surveys. For example, George Terhanian, the director of Internet Research at Harris Block, argues that Web-based polling eventually will be able to predict election results accurately. One company has already found ways to overcome some of the problems associated with Internet-based polling. Knowledge Networks is a market research company conducts surveys by using the Internet. Knowledge Networks has provided Internet access to a random sample of Americans who agree to participate in surveys. This Internet polling has been used in traditional market research as well as in public opinion polling. Knowledge Networks was commissioned to conduct polling during George W. Bush's 2003 State of the Union address, as well as polling to predict the outcome of the Gray Davis/Arnold Schwarzennegger 2003 California recall election. This kind of polling avoids the pitfalls typically associated with Internet polls and may represent a new direction in future polling endeavors.

See Also the Following Articles

Election Polls • Election Polls, Margin for Error in • Marketing Industry • Politics, Use of Polls in • Polling Industry • Telephone Surveys

Further Reading

Albig, W. (1956). *Modern Public Opinion.* McGraw Hill, New York.

Asher, H. (1998). *Polling and the Public: What Every Citizen Should Know.* CQ Press, Washington, D.C.

Gallup, G., and Rae, S. (1940). *The Pulse of Democracy.* Simon and Schuster, New York.

Voss, D. S., Gelman, A., and King, G. (1995). Preelection survey methodology: Details from eight polling organizations, 1988 and 1992. *Public Opin. Q.* **59,** 98–132.

Warren, K. F. (2001). *In Defense of Public Opinion Polling.* Westview Press, Boulder, CO.

Wheeler, M. (1976). *Lies, Damn Lies, and Statistics.* Liveright, New York.

Internet Literature

The Howard W. Odum Institute for Research in Social Science. (http://www2.irss.unc.edu).
Inter-university Consortium for Political and Social Research (ICPSR) (www.icpsr.umich.edu).
Knowledge Networks (www.knowledgenetworks.com).
The Louis Harris Data Archive (www2.irss.unc.edu).
National Opinion Research Center (http://www.norc.uchicago.edu).
Nielsen Media Research (http://www.nielsenmedia. com).
The Pew Research Center (http://people-press.org).
The Roper Center for Public Opinion Research (www.ropercenter.uconn.edu).
Survey Research Center (www.isr.umich.edu).

Population vs. Sample

John A. Grummel
West Virginia State University, Institute, West Virginia, USA

Glossary

accessible population The population that is actually accessible to the researcher. It may not be possible for a researcher to draw a sample from everyone (or everything) in the theoretical population because there may not be a complete listing of all the cases. The list of the population from which a sample is actually drawn is the accessible population. The accessible population does not have to be different from the theoretical population.

population The entire group of individuals, places, or other objects such as voters, organizations, political parties, and cities that conform to a set of specifications that a researcher wishes to study. Populations can be distinguished between the accessible population and the theoretical population.

population parameter A value, usually a numerical value that describes a population. Examples of population parameters include the mean, range, and standard deviation.

sample A set of individuals, geographical areas, or other objects selected from a population intended to represent the population in the study.

sampling error The discrepancy, or amount of error that exists between a sample statistic and the corresponding population parameter. If a sample is the same as a population there is no sampling error. The larger the sample the more closely the sample reflects the population of interest and as such the smaller the amount of error. Smaller samples are less like to mirror the larger population and the sampling error will be larger.

sampling frame The population from which the sample is actually drawn is known as the sampling frame. The sampling frame is a list of all known cases in a population. For example, the sampling frame for all residents in Los Angeles could be the Los Angeles phone book.

sample statistic A value, usually numerical, that describes characteristics of a sample such as the mean and standard deviation.

theoretical population The population about which one wants to make generalizations. The theoretical population includes all cases of the population of interest. The theoretical population may differ from the accessible population because all the possible cases might not be known or accessible to the researcher.

Determining whether to study a population or to study a sample from a particular population is usually not a difficult decision due to practical concerns such as costs in time and money. If it is impractical to study a population, there are certain problems that should be taken into account when studying a sample. The following are descriptions of populations and samples and their use in the social sciences. Also included is a discussion of why samples are studied more often than populations, as well as a general description of sampling, and concerns researchers should take into account when studying samples of larger populations.

One of the main points of research is to make inferences about events, people, nations, organizations, etc., in the population. By analyzing the results from a sample, for example, researchers hope to make general statements about the population, whatever the population under consideration. It would not be research, for example, if you tested three particular motorcycles to see which gets better gas mileage—unless you hoped to say something about the gas mileage of those models of motorcycles in general. In other words, a researcher might do an experiment on the effect of a particular method of teaching sociology using 50 students as participants in the experiment.

The purpose of the experiment would not be to find out how those particular 50 students respond to the experimental condition but rather to discover something about what works best in general when teaching sociology. The strategy and underlying logic of inference in almost all social and behavioral science research is to study a sample

that is believed to be representative of the general population (or of some particular population of interest). The sample is what is studied, and the population is an unknown that researchers draw conclusions about on the basis of the sample.

Definition of Terms

Broadly understood, a *population* is the entire group of individuals, objects, or cases, such as states, school districts, and parties) that a researcher wishes to study. By entire group, it is literally meant every single case. More specifically, a population is defined by Chein as the "aggregate of all cases that conform to some set of specifications." For example, by the specifications "pupils" and "attending public high school in the United States," the population would be defined as all pupils attending public high school in the United States. It is the population about which researchers wish to generalize. A researcher might desire to investigate the attitudes of public high school pupils concerning the quality of their educational environment.

One can distinguish between two types of populations: *theoretical* and *accessible*. The population that one desires to make generalizations about is known as the theoretical population. The theoretical population may also be referred to as the target or study population. Given the above example, all pupils attending public high schools in the United States would be the theoretical population. The accessible population is the population that is in fact accessible to the researcher. The accessible and theoretical populations are not always the same. Using the above theoretical population example of all pupils attending public high school in the United States there might not be, for whatever reason (i.e., reporting problems), a complete listing of all the pupils attending public high school in the United states. The available list would be the accessible population.

The size of any given population being considered depends on how the investigator defines the theoretical population. If the population is defined as all pupils attending public high schools in the United States this would be a large population, whereas defining the population as all pupils attending high school in a sparsely populated county would be a much smaller population. Examples of other large populations could include studying all the women on the planet, all the registered Republicans in the United States, or just married men in major cities, such as St. Louis, Chicago, New York, or Los Angeles. One might desire to be more specific, limiting the size of the population. For example, only married men who are registered voters in the United States (a smaller population) or conservative heads of state (an even smaller population) might be investigated. Furthermore, the specific nature of the population of interest depends on the research question and phenomenon being studied. For example, if the researcher decided to study only female pupils and their attitudes toward their educational experience, the population would be defined as all female pupils attending public high schools in the United States. These examples indicate that populations can vary in size from extremely large to very small.

It is quite important that the researcher identify the theoretical population that is being studied. If the theoretical population is not adequately defined or identified, any discussion of the misidentified population might be inaccurate. As noted earlier, populations are the aggregation of all cases conforming to set of specifications or specified elements. For example, specifications could refer to "people" but it could also refer to things such as businesses, interest groups, community organizations, minority businesses, corporations, political parties, parts produced in a factory, or anything else an investigator desires to study. Specifications could also refer to people or things in a particular geographic location, such as Canada or New York.

Once the accessible population is clearly defined, the researcher lists the members of the accessible population. For example, if the researcher is studying voting behavior, the members of the accessible population might be eligible voters within a particular geographic area as listed with the voter registrar's office. The accessible population becomes the sampling frame, the population from which the sample is drawn. All the elements—sets of specifications—that are part of the population of interest to the research question should be part of the sampling frame. The specifications could refer to individuals but it could also refer to objects or organizations such as businesses, interest groups, community organizations, minority businesses, corporations, political parties, parts produced in a factory, or anything else an investigator desires to study. If the elements are not included, any data collected for the sample might not be representative of the population being considered.

It is important to note that sampling is not always necessary, for example, if all the units in the population are identical. In some cases, for example, for smaller theoretical populations, it may even be possible to conduct a census as the United States does every 10 years. For a discussion of the debate regarding sampling and the census see Fienberg, Anderson and Fienberg, and; Brunell, as well as the additional readings listed in Further Reading. In social science research, however, this is not normally the case because of the immense cost and time that is often involved.

Since it is often impractical to study populations given limits of such things as time and money, typically researchers study samples of individuals in order to make inferences about populations. A sample differs

from a population in that a sample is a set of individuals (or things) selected from a population that is intended to represent the theoretical population being studied. It is important to note that a sample is the group of people or things that are selected to be in the study, not the group of people or things that are actually in the study. For example, all the individuals chosen to be in a study, for one reason or another, might not be able to participate. Researchers should always identify the sample in terms of the accessible population from which it was selected. Ideally, the sample is selected from a population using a strictly random procedure. A description of the random sampling procedure and issues of sampling are addressed below.

Like populations, samples can vary greatly in size. A sample might consist of 50 Mexican women who participate in a particular experiment, whereas the population might be intended to be all Mexican women. In an opinion survey, for example, 1000 individuals might be selected from the voting-age population through random digit dialing and asked for whom they plan to vote. The opinions of these 1000 people are the sample. Depending on the resources, such as money and time, the researcher could use a larger sample—for example, 2000 people. As illustrated below, researchers desire to study the largest sample possible because the larger the sample size the less error associated with the generalizations made concerning the population of interest. A researcher then hopes to apply the findings from the sample of 1000—that is, generalizes—to the larger population.

Statistical Terminology for Populations and Samples

Before discussing why samples are studied rather than populations, it is necessary to distinguish whether the data collected by a researcher is from a population or a sample. Any characteristic of a population is called a population parameter. The mean (arithmetic average of a distribution of scores), range (the difference between the highest and lowest score in a distribution), and standard deviation (reflects how well the mean of a variable represents the central tendency of a population or sample) of a population are examples of what are called population parameters. A population parameter usually is unknown and can only be estimated from what you know about a sample from that population.

A parameter may be obtained from a single measurement, or it may be derived from a set of measurements from the population. A characteristic of a sample is called a sample statistic. Like a population parameter, a sample statistic may be obtained from a single measurement, or it may be derived from a set of measurements from the sample. For example, the average of the scores (the mean) for a sample is a statistic. Another type of sample statistic would be the range of scores for the sample and standard deviation.

Table I Symbolic Representation of Several Population Parameters and Sample Statistics

	Population	*Sample*
Mean	μ	X
Standard deviation	σ	S
Variance	σ^2	s^2

As displayed in Table I, statisticians frequently use different symbols for a population parameter and a sample statistic. By using different symbols one can readily distinguish whether a characteristic, such as an average, is describing a population or a sample.

Why Study Samples Instead of Populations?

The decision whether to collect data for a population or from a sample is usually made on practical grounds. If conclusions are to be drawn about a population, the results would be most accurate if one could study the entire population, rather than a subgroup—a sample—from that population. However, in most research situations this is simply not practical or feasible. Studying an entire population is often too expensive and time consuming. Furthermore, some studies do not lend themselves to sampling, such as case studies. More importantly, one of the main points of research is to be able to make generalizations or predictions about events beyond our reach.

Sampling Concerns

There are a number of problems associated with using samples. Before proceeding with the problems, however, it is important to discuss the idea of random sampling. The objective of random sampling is to select a particular number of units from the sampling frame such that each case has an equal chance of being selected. Examples of procedures to achieve this objective include the use of a table of random numbers, a computer random-number generator, flipping a coin, or a mechanical device to select the sample. One problem with using samples, as discussed below, is that a sample provides only limited information about the population. There is always some degree of error associated with the information acquired from samples. Although samples are often representative of their populations, sometimes they are not. A sample is not

expected to give a perfectly accurate picture of the whole population.

A sample statistic (for example the mean) obtained from a sample generally will not be identical to the corresponding population parameter (population mean). There usually is some discrepancy between a sample statistic and the corresponding population parameter. This discrepancy is called *sampling error*. Sampling error, or margin of error, is the discrepancy, or amount of error that exists between a sample statistic and the corresponding population parameter. It gives the researcher some idea of the precision of the statistical estimates. There are two factors that reduce sampling error: sample size and variability. In other words, the less variability in the population or the larger the sample size, the lower the amount of error or discrepancy between the sample statistics and the actual population parameters.

It is essential that any sample is as representative as possible. A representative sample is a sample that looks like the population from which it was selected in all respects that are potentially relevant to the study. The distribution of characteristics among the elements of the representative sample is the same as the distribution of those elements among the total population. In an unrepresentative sample, some characteristics are over- or underrepresented. The researcher needs to be certain whether the sample is representative of the population being studied. If not, the study may not be externally valid. External validity refers to the degree a researcher, based on a sample, can make generalizations about the population of interest as well as other cases and at different times.

There are other problems associated with using samples. For example, to estimate unknown population parameters accurately from known sample statistics, three major problems must be dealt with effectively: the definition of the population, the sample design, and the size of the sample. The importance of defining the population was discussed above. To reiterate, if the population is poorly defined, errors in sampling can occur. If a researcher does not take care to properly define the population, the sampling frame will be inaccurate, and consequently the sample might be unrepresentative of the population the researcher actually desires to study. Thus, the accuracy of a sample depends on the sampling frame, and if the sampling frame is incomplete or inappropriate, sample bias will occur. In such cases, the sample will be unrepresentative of the population and inaccurate conclusions about the population may be drawn. Potential problems in sampling frames include incomplete frames, cluster of elements, and blank foreign elements. The problem of incomplete frames occurs when sampling units (a single member of a sampling population) included in the population are missing from the sampling frame. A detailed example is provided below. The problem of cluster of elements occurs when sampling units are listed in clusters rather than individually. For example, the sample frame of a particular study might include census blocks, whereas the study focuses on individuals and heads of household. A possible solution to such a problem would be to take a sample of blocks and then to list all the individual households in the selected blocks. The problem of blank foreign elements occurs when sampling units of the sampling frame are not part of the target population. The problem of blank foreign elements is illustrated by the following example. The population is defined as eligible voters but the sampling frame contains individuals who are too young or ineligible to vote.

An example of a well-known error in sampling frames, in this case incomplete frames that led to an extremely erroneous conclusion was the 1936 *Literary Digest* election poll. In 1936, Franklin Delano Roosevelt was running against Republican candidate Alfred Landon. In the poll, the largest in history (2.4 million individuals), *Literary Digest* predicted a victory for Landon by 57% to 43%. Despite this prediction, Roosevelt won by 62% to 38%. Despite the large sample size, the error was the largest ever made by a polling organization. The reason for the error was in the sampling frame. Questionnaires were mailed to 10 million people based on lists compiled from such things as telephone directories and club memberships. In 1936, few people had telephones and even fewer belonged to clubs. The sampling frame was incomplete and systematically excluded the poor and as such the sampling frame did not accurately reflect the voting population.

The second problem concerns sample designs and arises in connection with securing a representative sample. As noted above, an essential requirement of any sample is that it be as representative as possible of the population from which it was drawn. In sampling theory, a distinction is made between probability and nonprobability sampling. For a more in-depth discussion of these topics, see Carroll, Galloway, Glasgow, Handwerker, Harter, and Wright (this volume).

Probability sampling is one that can specify for each sampling unit of the population the probability that it will be included in the sample. In nonprobability sampling there is no way of specifying the probability of each unit's inclusion in the sample and includes such designs as convenience sampling, purposive sampling, "snowball" sampling, and quota samples. Although accurate estimates can only be made with probability samples, social scientists do employ nonprobability samples. Examples of research using nonprobability samples include the use of college students in psychological and public opinion studies by professors because of the convenience.

Probability sample involves some form of random sampling and, in most cases, is the ideal method of picking

a sample to study. Random selection, or random sampling, is a process for obtaining a sample from a population that requires that every individual in the population have the same chance of being selected for the sample. A sample obtained by random selection is called a random sample. Examples of random sampling include stratified random sampling, systemic random sampling, cluster random sampling, and multistage random sampling.

It is important not to confuse truly random selection from what might be called haphazard selection, such as just choosing whoever is available or happens to be first on the list—convenience sampling. It is quite easy to accidentally pick a group of people to study that is really quite different from the population as a whole. Unfortunately, it is often impossible to study a truly random sample. There are a number of possible problems. One problem is that there is usually not a complete list of the full population (although telephone directories come close). A second possible problem is that not everyone a researcher approaches agrees to participate in the study (telephone surveys, for example, are often biased against the wealthy and educated, who are away from home more and use answering machines to screen calls). Another potential problem is the cost and difficulty of tracking down people who live in remote areas to participate in the study. For these and other reasons as well, social and behavioral scientists often use various approximations to truly random samples.

The third problem concerns sample size. (For an in-depth discussion on sample size see Suchindram, this volume.) As noted above, there is usually some discrepancy between a sample statistic and the corresponding population parameter and that this discrepancy is known as sampling error. The following example is intended to illustrate how sample size may reduce or increase the amount of error, which has an impact on the level of confidence a researcher has in the conclusions he or she makes regarding a particular population. The equation for sampling error is

$$s = \sqrt{\frac{P \times Q}{n}}$$

The symbols P and Q equal the population parameters for the binomial, n equals the number of cases in the sample, and s is the standard error. To illustrate, assume that 70% of a college student body is in favor of ending affirmative action and 30% disapprove. Also, assume that the researcher is working with a sample of 100 students with 60% of the sample in favor (P) and 40% of the sample against (Q) ending affirmative action. Before proceeding, it is important to understand that the standard error is a valuable piece of information concerning how tightly estimates will be distributed around the actual population estimates (in this case support for ending affirmative action). According to probability theory, certain proportions of the sample respondents will fall within 66% of the mean, which are equal to plus or minus one standard error, from the true population mean. Furthermore, probability theory dictates that 95% of a sample will fall within plus or minus two standard errors of the true population mean and 99% will fall within plus or minus three standard errors of the actual population mean. When these numbers are entered into the equation, the standard error is 0.05, or about 5%. Based on the above discussion, with a sample of one hundred, if the researcher wants a 95% confidence level (plus or minus two standard error) the estimates will be plus or minus 10% of the actual value. If the sample size is increased to 200, the standard error is .035 or 3.5%, and the sample estimates will fall within plus or minus 7% of the true value. If the sample size is increased even more, say 500 the amount of error is even lower: 0.2, or 2%, indicating that the estimates are within plus or minus 4% of the actual population parameter. This example demonstrates that the greater the sample size, the lower the amount error and the greater the confidence a researcher will have in the conclusions he or she can make based on a given sample.

Conclusion

If possible, it is always better to study the entire population, but this is often simply impractical due to excessive cost, both in terms of time and money. It is for this reason that samples are often studied, and as outlined above, researchers need to take into consideration a number of potential problems when studying a sample from a given population. When researchers take into account the possible problems that can occur when studying samples, they can be almost as confident of their findings as if they had studied the entire population.

See Also the Following Articles

Measurement Error, Issues and Solutions • Sample Size

Further Reading

Anderson, M. J., and Fienberg, S. (2000). History, myth making, and statistics. *Politics Soc.* **33**(4), 783–792.

Anderson, M. J., and Fienberg, S. E. (1999). *Who Counts: The Politics of Census-Taking in Contemporary America.* Russell Sage Foundation, New York.

Aron, A., and Aron, E. N. (1997). *Statistics for the Behavior and Social Sciences.* Prentice Hall, Upper Saddle River, NJ.

Babbie, E. (2001). *The Practice of Social Research,* 9th Ed. Wadworth, Belmont, CA.

Brown, L. D., Eaton, M. L., Freedman, D. A., Klein, S. P., Olshen, R. A., Wachter, K. W., Wells, M. T., and Ylvisaker, D. (1999). Statistical controversies in Census 2000. *Jurimetrics* **39,** 347–375.

Brunell, T. (2000). Using statistical sampling to estimate the U.S. population. *Politics Soc.* **33**(4), 775–782.

Chein, I. (1981). An introduction to sampling. In *Research Methods in Social Relations* (C. Selltiz, *et al.*, ed.), 4th Ed. Holt, Rinehart, & Winston, New York.

Cohen, M. L., White, A. A., and Rust, K. F. (eds.) (1999). *Measuring a Changing Nation: Modern Methods for the 2000 Census*. National Academies Press, Washington, DC.

Edmonston, B., and Schultze, C. L. (eds.) (1995). *Modernizing the U.S. Census*. National Academies Press, Washington, DC.

Frankfort-Nachmias, C., and Nachmias, D. (1992). *Research Methods in the Social Sciences*, 4th Ed. St. Martin's Press, New York.

Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics*. Norton, New York.

Johnson, J. B., Joslyn, R. A., and Reynolds, H. T. (2001). *Political Science Research Methods*, 4th Ed. CQ Press, Washington, DC.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry*. Princeton University Press, Princeton, NJ.

Mulry, M. H., and Spencer, B. (1993). Accuracy of the 1990 census undercount adjustments. *J. Am. Statist. Assoc.* **88,** 1080–1091.

Robinson, J., Bashir Ahmed, G., Das Gupta, P., and Woodrow, K. A. (1993). Estimation of population coverage in the 1990 United States Census based on demographic analysis. *J. Am. Statist. Assoc.* **81,** 338–346.

Sanocki, T. (2001). *Student Friendly Statistics*. Prentice Hall, Upper Saddle River, NJ.

Schutt, R. K. (1999). *Investigating the Social World*, 2nd Ed. Pine Forge Press, Thousand Oaks, CA.

Trochim, W. M. K. (2001). *The Research Methods Knowledge Base*. Atomic Dog, Cincinnati.

Walsh, A., and Ollenburger, J. C. (2001). *Essential Statistics for the Social and Behavior Sciences*. Prentice Hall, Upper Saddle River, NJ.

Prevalence and Incidence

Stephen W. Marshall
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Glossary

case fatality rate The proportion of outcome-positive individuals that die.
closed population A population in which there is no migration or loss to follow-up.
competing risk (competing causes of death) The risk of death from causes other than the outcome of interest.
current case A person who is outcome-positive.
follow-up Monitoring a population (or subset of a population) to determine incidence.
incidence The occurrence of new outcome-positive cases in a population, quantified either in units of person-time at risk (incidence rate) or people at risk (incidence proportion).
incidence proportion (cumulative incidence) The proportion of the population that becomes outcome-positive over a defined period of time.
incidence rate (incidence density) The number of new cases divided by person-time at risk.
new case A person who has just become outcome-positive.
observation period (monitoring period) The period of time over which a population (or subset of a population) is followed to determine incidence.
outcome The disease or other ill-health condition of interest.
person-time at risk The sum of individual at-risk periods in a group of individuals.
prevalence The proportion of a population that is outcome-positive at a specific point in time.
prevalence pool The subset of a population that is outcome-positive at a specific point in time.
steady state When the flow of subjects into the population is exactly equal to the flow of subjects out of the population, and determinants of prevalence and incidence are static.

Prevalence and incidence are fundamental concepts that underlie the measurement of health outcomes in human populations. These concepts, although simple, are frequently misunderstood, even to the point where the terms prevalence and incidence are sometimes used interchangeably. Formally, prevalence is the proportion of a population that is outcome-positive (has the outcome of interest) at a specific point in time. Simply put, prevalence is the proportion of current cases in a population; a "current case" is a person who is outcome-positive at that point in time. Prevalence quantifies the status of the population with respect to the outcome of interest at a given point in time. It does not, however, express anything about how frequently new cases occur in the population. Incidence, on the other hand, measures the occurrence of new outcome-positive cases in the population. Two measures of incidence are of interest: incidence rate (or incidence density) is the number of new cases per unit of person-time at risk, whereas the incidence proportion (cumulative incidence) refers to the proportion of the population that becomes outcome-positive over a defined time interval. Prevalence and incidence proportion are proportions, but incidence rate is measured in units of inverse time.

Overview of Main Concepts

Introduction to Prevalence

Prevalence of disease quantifies the disease status of a population at a given point in time. Prevalence measurement addresses the question concerning what proportion of the population currently has the outcome of interest. People who are currently outcome-positive are often referred to as the pool of prevalent cases, or the prevalence pool. In Fig. 1, the pool of prevalent cases within a population at a given point in time is represented by

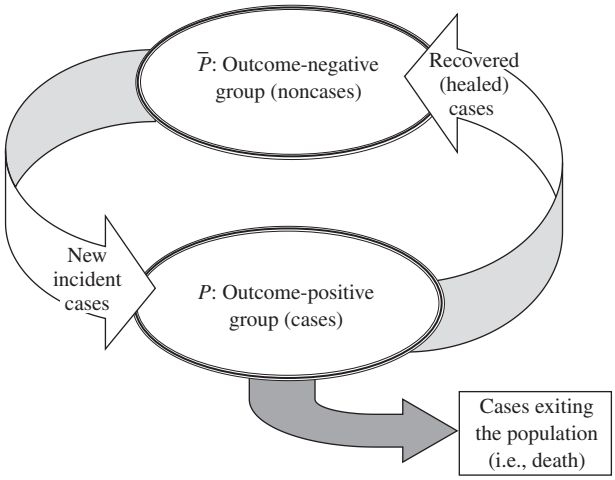


Figure 1 Relationship between prevalence and incidence in a closed population, $N = P + \bar{P}$. The group of people who are outcome-positive (lower circle) are prevalent cases. This "prevalence pool" is fed by incident (new) cases (left arrow). Cases exit the prevalence pool when they recover (right arrow) or die (lower arrow).

the lower circle (outcome-positive group, or cases). This is the prevalence pool. The upper circle represents the remainder of the population, cases that are outcome-negative at the same point in time. The two circles are a mutually exhaustive and exclusive partition of the population. The total population, N, equals $P + \bar{P}$. To quantify prevalence, it is necessary to sample from the whole population, and determine what proportion is outcome-positive. In essence, this amounts to determining the relative size of the two circles in Fig. 1. As people recover from illness, they return to the outcome-negative pool, at which time they are presumably again eligible to develop the outcome condition (e.g., reinfection); however, for some outcomes, reinfection is unlikely or impossible. People who do not recover die and exit the population.

Introduction to Incidence

Incidence quantifies the occurrence of new outcome-positive cases, i.e., how rapidly new cases are added to the prevalence pool. Measurement of the incidence of health outcomes in a population is an important aspect of documenting the burden due to various diseases and other sources of ill health. To measure incidence requires first identifying a group of outcome-negative people, and then prospectively following them over time to identify how many of them develop the outcome. In Fig. 1, incidence is the rate at which people flow along the left-hand arrow from the outcome-negative pool (top circle) to the prevalence pool (bottom circle). Whereas prevalence is measured at a single point in time, measurement of incidence involves prospectively following people over time to quantify the rate of conversion from the outcome-negative state to the outcome-positive state.

Can Prevalence Be Low When Incidence Is High (and Vice Versa)?

The answer is yes. The relationship between prevalence and incidence involves the duration of the outcome condition (the disease). If the disease duration is short, say, due to an extremely lethal virus, the prevalence pool will remain small, even if the incidence is high. If the disease is due to a virus such as Epstein–Barr virus, which causes infections from which people apparently never recover, but which is rarely directly lethal, then the prevalence pool can be large, even if the incidence is quite low.

Initial versus Recurrent Events

Recurrent events occur whenever an individual experiences multiple distinct episodes of the same outcome. For some conditions and/or individuals, the fact that they were outcome-positive once makes them unable to return to their original state, e.g., reinfection following the primary infection is very rare for some diseases. One approach to the problem of reinfections is to focus on determining only the incidence and/or prevalence of initial infections (i.e., first infections). Thus, episodes of reinfection would be excluded. In practice, it may be difficult or impossible to eliminate reinfection cases.

Measurement Concepts

Measurement of prevalence and incidence should utilize two types of statistical estimator: a point estimate (e.g., 35%, or 25 per 100,000 person-years) and an interval estimate, or confidence interval (e.g., 33–37%, or 22–28 per 100,000 person-years). Use of the 95% confidence interval (95% CI) is traditional. The purpose of the interval estimator is to quantify the precision of the point estimate, and it is a function of the sample size and the variability of the outcome in the population. In general, it is considered desirable to have a point estimator that is unbiased in expectation and an interval estimator with optimal coverage. Ratio, or difference, effect-measures can be computed by dividing, or subtracting, prevalence or incidence measures for subgroups within the population and, under certain conditions, these effect-measures permit inference concerning causal effects.

What Is Prevalence?

Definition of Prevalence

Prevalence is the proportion of the population that is outcome-positive at a given point in time. It measures

the current overall outcome-positive status of the population, whereas incidence reflects the pace of arrival of new cases. Prevalence is determined by the size of the population and the number of outcome-positive cases. Prevalence is sometimes referred as the prevalence rate, but prevalence is actually a proportion. It can never be less than zero or greater than one. It is often expressed as a percentage, e.g., "68% of the population tested positive for Epstein–Barr virus." Sometimes, the count of the number of prevalent cases is defined to be the prevalence (rather than the proportion).

Measurement of Prevalence

A common method of measuring prevalence is to conduct a one-time cross-sectional survey. Such a survey might simply involve a self-administered or interview-administered questionnaire, or a clinical test for the outcome. In some settings, it is practical to survey the whole population (to conduct a census). In most situations, however, some form of sampling is required for logistical reasons. A rich class of statistical designs for such surveys is available. Designs include simple random sampling, systematic sampling, stratified sampling, and cluster sampling. The common features of all these designs is complete (or nearly complete) enumeration of the population of interest, followed by a probabilistic sampling so that each member of the population has a known and defined probability of selection into the sample. These defined probabilities of selection can be used to make inferences about the whole population, based on the sample. In the simplest situation, a simple random sample in which all members of the population have equal probability of selection, the population prevalence is directly estimated by the sample prevalence P, defined as

$$P = \frac{y}{n}, \tag{1}$$

where y is the number of outcome-positive cases in the sample and n is the sample size. The standard error (SE) of the sample prevalence in a simple random sample is estimated by

$$\mathrm{SE}[P] = \sqrt{\frac{P(1-P)}{n}}, \tag{2}$$

and a 95% confidence interval can be estimated as

$$P \pm 1.96 \cdot SE[P]. \tag{3}$$

Prevalence Example

Joel M. Clingenpeel and S. W. Marshall, physicians concerned about head injuries, sought to quantify the prevalence of helmet rentals in U.S. ski areas. The study population (the sample frame) was enumerated on the basis of a listing of all U.S. ski areas. Hypothesizing strong regional variations in prevalence, a stratified sample was drawn to ensure that adequate numbers of ski areas would be sampled in each geographic region. In analysis, each respondent ski area was weighted by the inverse probability of selection, accounting for the stratification. The overall prevalence of helmet rentals at U.S. ski resorts in 2003 was 50.5% (95% CI: 48.2 : 52.7).

What Is Incidence?

Definition and Measurement of Incidence

Incidence quantifies the occurrence of new outcome-positive cases in the population. Whereas prevalence defines the proportion of the population that comprises current cases, incidence quantifies how quickly new cases arise. To measure incidence requires defining a subset of the population that is outcome-negative, and then prospectively monitoring this subgroup to determine how many of them become outcome-positive. This monitoring is commonly referred as to "follow-up." Incidence is defined by the length of time of monitoring, the number of people being monitored, and the number of new cases.

Two measures of incidence are of interest: incidence rate (the number of new cases per unit of person-time at risk) and incidence proportion (the proportion of the population that becomes outcome-positive over a defined time interval). These two measures both use the number of new cases divided by some measure of the population at risk. They differ, however, in how the population at risk is defined. Incidence rate is quantified in units of person-time at risk, whereas incidence proportion is quantified in terms of people at risk. Incidence rate is interpreted as an instantaneous measure, whereas the interpretation of incidence proportion is closer to an intuitive sense of individual risk.

Incidence Rate (Incidence Density)

The incidence rate is the number of new cases divided by person-time at risk. This epidemiologic concept of person-time at risk is a key element in computing incidence rate and it requires some explanation. Person-time at risk accounts for the fact that the length of follow-up may vary from one person to another. For example, once a subject has become outcome-positive, they are removed from the outcome-negative pool and do not contribute person-time at risk to the denominator of the rate calculation while they are outcome-positive. Once a person has recovered from the outcome and has returned to an outcome-negative state, they return to the pool of subjects eligible to sustain the outcome and can again contribute person-time at risk to the denominator of incidence rate

(this assumes, however, that the incidence for those who have a positive history of the outcome is identical to that for those who have no history, which is not true for certain infections). People who do not sustain the event of interest during the monitoring period contribute the full length of monitoring period to the person-time computation (in the language of lifetables and survival analysis, these individuals are referred to as "censored"). There are other reasons why people might cease to contribute person-time at risk during the monitoring period. Sometimes subjects are lost to follow-up for logistical reasons, or they may emigrate from the population. In addition, it is usually impossible to begin follow-up for all members of the outcome-negative pool at exactly the same moment. Finally, people may die during the monitoring period due to an outcome other than the outcome being studied (i.e., due to a competing cause of death). All of these complexities can be readily accommodated through the use of the concept of person-time at risk. The essence of person-time at risk is that each person has an individual amount of time during which they are at risk of becoming outcome-positive. The sum of these individual at-risk periods is the overall person-time at risk. Formally, incidence rate (IR) is estimated as

$$\mathrm{IR} = \frac{a}{\sum_i n_i}, \tag{4}$$

where the subscript i indexes the individual subjects, n_i is the amount of time that each individual was at risk of the outcome, and a is the count of new outcome-positive episodes during the observation period. The incidence rate is also sometimes referred to as the incidence density or the force of mortality (or morbidity, for nonfatal events). Because a is a count, the assumption is that it follows a Poisson distribution, and therefore its variance will equal its mean. Assuming that the person-time at risk is a fixed constant (i.e., does not have a probability distribution), then the standard error of the incidence rate can be estimated as

$$\mathrm{SE}[\mathrm{IR}] = \frac{\sqrt{a}}{\sum_i n_i}, \tag{5}$$

and a 95% confidence interval can be defined as

$$\mathrm{IR} \pm 1.96 \cdot \mathrm{SE}[\mathrm{IR}]. \tag{6}$$

The assumption that person-time at risk is a fixed constant might seem restrictive, but relaxing this assumption appears to have minimal effect in practice. However, Eqs. (5) and (6) invoke certain assumptions regarding study size. For small samples, it is preferable to use exact techniques to estimate the confidence interval. Exact techniques involve computing large numbers of permutations based on the observed data. StatXact is one of several software packages that implement the numerical algorithms required for exact computations.

Equation (5) also assumes that all outcome episodes are independent. If the outcome is transmitted (as in an infection), alternative techniques may need to be employed. If there are multiple outcome episodes per person during the observation period, the variance estimator should be statistically adjusted for this dependence using methods for clustered or correlated data. If the focus is only on the incidence of first episodes, then a is the count of new outcome-positive individuals during the observation period, and Eq. (5) is valid. Because incidence rate is the count of new cases divided by units of person-time at risk, it follows that the units of the incidence are inverse time. For convenience, epidemiologic incidence rates are usually defined in terms of person-years and are multiplied by 100,000 when reporting the rate (e.g., "the rate of cardiovascular mortality was 35 per 100,000 person-years"). However, this is strictly convention, and in theory, any time unit or multiplier can used, provided that they are clearly specified. In fact, the actual numeric portion of incidence rate is meaningless unless the units and multiplier (if any) are clearly specified. Incidence rate is an instantaneous measure, analogous to the concept of speed as defined in physics. Note that speed also has units of inverse time.

As was noted previously, prevalence is a proportion and thus can never be less than zero or greater than one; it will soon become apparent that the same is true of incidence proportion. The incidence rate can also never be less than zero; however, unlike prevalence and incidence proportion, incidence rate has no upper bound, and can range up to infinity. This is consistent with the interpretation of incidence rate as an instantaneous measure. In fact, any incidence rate can be made to exceed one by simply rescaling the time unit. An incidence rate of 500 per 100,000 person-years, for example, is equivalent to 5 per 10 person-centuries and 5 per person-millennia.

The term "annual incidence rate" is in common use, but this term should be avoided. Use of this term masks the fact that incidence rate is an instantaneous measure. The confusion arises from that the fact that an incidence rate can often be readily computed by taking the annual number of new cases in a defined population and dividing this number by the size of the population at risk in the middle of the year. The resulting estimator is the average incidence rate (in units of person-years) for the year being studied. The size of the population at midyear has been multiplied by one to denote an average of 1-year of follow-up per individual. Though this procedure is numerically correct and a very reasonable method for estimating the average incidence rate, the mechanics of this computation have given rise to the unfortunate misperception that the resulting estimate is an "annual incidence rate," and as a result this rate is sometimes reported as being per

100,000 persons, instead of per 100,000 person-years. In fact, this estimator is the average incidence rate for the 1-year period under study, and its units are person-years, not persons. Although the distinction is subtle, this error confuses the correct interpretation of incidence rate.

Because the units of incidence rate are inverse time, the interpretation of the reciprocal of the all-cause mortality rate has a surprising interpretation, under certain conditions, as the average life expectancy of the population. These conditions include that the number of people exiting the population due to death or emigration is identical to the number of people entering the population due to birth or immigration. In practice, this condition is unusual.

Incidence Rate Example 1

David Savitz and Dana Loomis retrospectively constructed a cohort of workers employed in the electric industry to examine exposure to electromagnetic radiation and the incidence of leukemia and brain cancer. The cohort consisted of 138,905 workers who were employed in the industry between January 1, 1950 and December 31, 1988. Workers entered the cohort when they were hired into the industry and exited the cohort when they died. The total number of deaths from all-cause mortality was 20,733, of which 164 deaths were due to leukemia. The 138,905 workers accumulated 2,656,436 person-years of follow-up during the 38-year study period. The mortality rate from leukemia in these workers was therefore 6.2 per 100,000 person-years (95% CI: 5.2, 7.1).

Incidence Rate Example 2

The National Highway Traffic Safety Administration has estimated that the number of deaths from motor vehicle crashes in the United States in 2001 was 42,116. The total resident population of the United States in 2001 was 284,796,887. Thus, the incidence of motor vehicle fatality for 2001 was 14.8 per 100,000 person-years. Another way to express this incidence rate is as deaths per miles driven. During 2001, U.S. residents traveled 2781 billion miles in motor vehicles, with a fatality rate of 1.5 per 100 million vehicle miles traveled.

Incidence Proportion (Cumulative Incidence)

In addition to incidence rate, a widely used alternative measure of incidence is incidence proportion. The incidence proportion (also frequently referred to as cumulative incidence) is the proportion of the population that becomes outcome-positive over a defined time interval. As with incidence rate, the estimation of this measure uses data on the length of time for which the group was monitored, the number of people being monitored, and the number of new cases. The incidence proportion (IP) is estimated as

$$\mathrm{IP} = \frac{a}{m}, \tag{7}$$

where m is the size of the group being monitored at the beginning of the observation period and a is the number of individuals who become outcome-positive during the observation period. The standard error of incidence proportion is estimated as

$$\mathrm{SE}[\mathrm{IP}] = \sqrt{\frac{\mathrm{IP}(1-\mathrm{IP})}{m}}, \tag{8}$$

and a 95% confidence interval can be estimated as

$$\mathrm{IP} \pm 1.96 \cdot \mathrm{SE}[\mathrm{IP}]. \tag{9}$$

Again, exact techniques, such as those implemented in StatXact, should be used when a is small. Because incidence proportion is a proportion, it is bounded by 0 and 1, and has no units.

Two important considerations arise for incidence proportion. First, the length of the observation period must always be specified when reporting incidence proportion. This measure has no valid interpretation unless the observation period is given. Second, this measure does not take account of the attrition in the population due to causes of illness and death other than the outcome being studied. Unlike incidence rate, the denominator of incidence proportion is not adjusted for those who exit the population risk due to emigration or death from other causes (outcomes other than the outcome being studied, also referred to as competing causes). Nevertheless, the measure is important because it has strong intuitive appeal. The proportion of the population that will fall ill over a defined time period is closely related to the concept of risk that is central to demography and epidemiology. Incidence proportion can be thought of as average risk, subject to the absence of competing risks.

One frequent use of incidence proportion is to quantify the so-called case fatality rate, which is the proportion of outcome-positive individuals who die. Note that the case fatality rate is a proportion, not a rate. In terms of recurrent episodes of the outcome per individual, typically only the initial episodes are used in computing incidence proportion. That is, the incidence proportion would be the proportion of individuals with one or more episodes of the outcome during the observation period.

Incidence Proportion Example

The Centers for Disease Control (CDC) conducted an investigation of an outbreak of staphylococcal food poisoning at a private party in Florida in 1997. The CDC interviewed 98 people who attended the party; of these, 18 developed symptoms of food poisoning within 8 hours

after eating food at the party. The 8-hour incidence proportion of food poisoning for this group was therefore 19% (95% CI: 12, 27).

Relationship between Incidence Rate and Incidence Proportion

Intuitively, it would be expected that incidence rate multiplied by the observation period would equal incidence proportion. That is, if the average incidence rate is 1 per 1000 person-years, and the population is observed for 10 years, incidence proportion will be $10 \times 0.001 = 0.01$, or 1 per 100. In fact, this is an approximation that is true only under certain circumstances. The true relationship between incidence rate and incidence proportion is more complex. More formally, the intuitive expectation is

$$\mathrm{IP}_t = \mathrm{IR}_t \Delta t, \tag{10}$$

where Δt is the observation period and IR_t is the average incidence rate over this period. This equation is true when Δt is short, so that the size of the outcome-negative pool (the number of noncases) declines only slightly over the observation period. In addition, Eq. (10) assumes that the population is closed (no immigration or emigration) and there are no competing risks (outcomes other than the outcome being studied that remove individuals from the pool of noncases).

More generally, incidence rate may vary over the observation period, yielding

$$\mathrm{IP}_t = \sum_i \mathrm{IR}_i \Delta t_i, \tag{11}$$

where $\Delta t = \sum_i \Delta t_i$. Accounting for the fact that individuals leave the pool of prevalent noncases when they become incident cases requires using the methods of survival analysis. Specifically, application of the Kaplan–Meier survival estimator yields this relationship:

$$\mathrm{IP}_t = 1 - e^{\sum_i \mathrm{IR}_i \Delta t_i}. \tag{12}$$

As before, Eq. (12) assumes that the Δt_i values are short such that the number of noncases declines only slightly within each subinterval, that the population is closed (no immigration or emigration), and that there are no competing risks. Equation (12), commonly referred to as the exponential formula, describes the true relationship between incidence proportion and incidence rate (in a closed population without competing risks). However, Eq. (10) provides an excellent approximation when Δt is short, so that the number of noncases declines only slightly over the observation period. If Δt is not a short period, then Eq. (12) should be used in place of Eq. (10).

Relationship between Prevalence and Incidence

Now assume that the population is in a condition in which the flow of incident cases into the prevalence pool is exactly equal to the flow of cases out of the prevalence pool. This is the steady-state condition. The number of new cases entering the prevalence pool along the left arrow of Fig. 1 is perfectly balanced by removal of cases from the prevalence pool through death (bottom arrow) or recovery (right arrow), so that overall size of the prevalence pool is static over time. If the prevalence is <0.1, and there is no migration into, or out of, the population, then the relationship between the prevalence (P) and incidence rate can be approximated by the simple formula

$$P = \mathrm{IR} \cdot \bar{D}, \tag{13}$$

where $\bar{D}$ is the mean duration of the disease, i.e., the average time to death or recovery. This formula can be extended to apply to age-specific prevalence. If prevalence is ≥ 0.1, the relationship is only slightly more complex:

$$\frac{P}{(N-P)} = \mathrm{IR} \cdot \bar{D}, \tag{14}$$

where N is the size of the total population. The quantity $P/(N-P)$ is referred to as the prevalence odds. Note that the inflow into the prevalence pool is given by $\mathrm{IR}(N-P)\Delta t$ and the outflow is $P\Delta t/\bar{D}$. If these are equal (i.e., the steady-state condition), then the resulting equation can be solved to obtain Eq. (14).

Special Situations

Birth Outcomes

Suppose it is desired to ascertain some measure of the number of children born with a specific congenital complication, such as cleft palate. A question arises as to whether prevalence or incidence should be measured. On one hand, the goal could be to quantify the incidence of cleft palate births in the overall population, but this is clearly influenced by many factors, including fertility and fecundity. To remove the influence of these factors, it is usually considered desirable to determine the proportion of all live births with cleft palate. Note that this measure is technically the prevalence, not the incidence, of cleft palate in the population of live births (note that it is a proportion, not a rate). Thus, even though the cleft palate births are incident events in the overall population, within the population of live births, they are prevalent.

Chronic Outcomes

A number of chronic conditions, such as arthritis, depression, or musculoskeletal pain, have ill-defined onset times. This means it is difficult to determine their incidence. The standard approach in this situation is often to develop some measure of the prevalence of the outcome. For example, in a large community survey of osteoarthritis, it would be desirable to ask about current symptoms and diagnoses of this largely irreversible condition. Determining the incidence could be problematic because sufferers may be in an advanced stage before obtaining a diagnosis.

See Also the Following Articles

Attrition, Mortality, and Exposure Time • Risk and Needs Assessments

Further Reading

Benichou, J. (2000). Absolute risk. In *Encyclopedia of Epidemiologic Methods* (M. H. Gail and J. Benichou, eds.), pp. 1–17. Wiley, New York.

Benichou, J., and Gail, M. (1990). Estimates of absolute cause-specific risks in cohort studies. *Biometrics* **46,** 813–826.

Clingenpeel, J. M., and Marshall, S. W. (2003). Helmet rental practices at United States ski areas: A national survey. *Injury Prevent* **9,** 317–321.

Elandt-Johnson, R. C. (1975). Definition of rates: Some remarks on their use and misuse. *Am. J. Epidemiol.* **102**(4), 267–271.

Hennekens, C. H., and Buring, J. E. (1987). *Epidemiology in Medicine.* Little, Brown and Son, Boston, MA.

Keiding, N. (2000). Incidence-prevalence relationships. In *Encyclopedia of Epidemiologic Methods* (M. H. Gail and J. Benichou, eds.), pp. 433–437. Wiley, New York.

Klein, J. P., and Moschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer, New York.

Levy, P. S., and Lemeshow, S. (1999). *Sampling of Populations: Methods and Applications.* Wiley, New York.

Rothman, K. J., and Greenland, S. (1998). Measure of disease frequency. In *Modern Epidemiology*, 2nd Ed. (K. J. Rothman and S. Greenland, eds.), pp. 29–46. Wiley, New York.

Savitz, D. A., and Loomis, D. (1995). Magnetic field exposure in relation to leukemia and brain cancer and mortality among electrical utility workers. *Am. J. Epidemiol.* **141**(2), 123–134.

Traffic Safety Facts 2001. (2002). A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System (2002). DOT HS 809 484. National Highway Traffic Safety Administration, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, D.C.

Tsiatis, A. A. (2000). Competing risk. In *Encyclopedia of Epidemiologic Methods* (M. H. Gail and J. Benichou, eds.), pp. 239–249. Wiley, New York.

Ward, K., Hammond, R., Katz, D., and Hallman, D. (1997). Outbreak of staphylococcal food poisoning associated with precooked ham—Florida, 1997. *Morbid. Mortal. Wkly. Rep. (MMWR)* **46**(50), 1189–1191.

Primate Studies, Ecology and Behavior

Timothy D. Smith
Slippery Rock University, Slippery Rock, Pennsylvania, USA

Annie M. Burrows
Duquesne University, Pittsburgh, Pennsylvania, USA

Glossary

arboreal A tendency of inhabiting trees.
cathemeral A pattern in which waking activities occur at various times of both day and night.
folivore An animal that primarily relies on a diet of leaves.
frugivore An animal that primarily relies on a fruit diet.
polyandry A polygamous system in which a female mates with more than one male within the same general timeframe.
polygamy A mating strategy in which one sex mates with more than one individual of the other sex.
polygyny A polygamous system in which a male mates with more than one female within the same general timeframe.

Nonhuman primates inhabit much of Africa, Asia, South America, Central America, and some associated islands. Most nonhuman primates live in tropical, arboreal habitats but also some quite varied nontropical habitats. Within these varied environments, primates may walk quadrupedally (on the ground or on tree limbs), bipedally, or may move by swinging below or leaping among tree branches and trunks. Primates are active diurnally, nocturnally, or both, and thus occupy different temporal as well as spatial niches. Social systems range widely, from one adult to multi-male/multi-female groups. Studies on primate ecology and behavior have shifted in focus regarding causal mechanisms that influence the type of social systems used by primates. The nature and distribution of food resources and predation pressure are prime influences; these factors may have a cascade of influences on primate behavior, including mating strategies.

Introduction

Primate Origins

The Order Primates evolved in the Cretaceous period around 81.5 million years ago from an ancestral stock of mammals, possibly with insectivore-like characteristics and most likely nocturnal in activity pattern. Several scenarios have been advanced to account for the origin of primates (see Cartmill and Sussman for a more thorough discussion and for original references therein). The arboreal theory held that the first primates evolved from a tree-dwelling stock of insectivorous, nocturnal mammals. In such an arboreal environment, it was suggested that vision and tactile senses would have been relied on more heavily while olfaction would have become less important since these animals hunted insects in the dark and made their way along complex tree limbs. Over time, the orbits shifted from a lateral position (typical of many nonprimate mammals) to a more frontal position. This "orbital convergence" would have produced stereoscopic vision, important for a mammal subsisting in an arboreal setting. In this scenario, a division of labor between fore and hind limbs was seen as a product of arboreality. The hind limbs became most important in locomotion, while the fore limbs acquired added specializations for manipulating and grasping tree limbs and insect prey (e.g., the appearance of nails instead of claws). The arboreal theory eventually came under criticism since this suite of traits (enhanced vision and tactile senses, reduced olfaction, nails instead of claws) is not typical of arboreal mammals in general. It was argued that, had a switch to an arboreal lifestyle in the last primate ancestor been the driving force in the evolution of primates and their unique suite of

traits, then these traits should be apparent in most other arboreal nonprimate mammals. In fact, no other arboreal mammals display stereoscopic vision or nails.

An alternate theory suggests stereoscopic vision and grasping hands/feet were initially an adaptation to a very specific niche within the arboreal habitat. This hypothesis considers grasping hands and feet as a common feature of terminal branch feeders. This, combined with stereoscopic vision that primates have in common with mammals such as the carnivores, has led to the visual predation hypothesis. This view holds that the earliest primates inhabited the smaller branches of trees and shrubs, feeding on insects and fruit found in this habitat. This refinement explains the development of vision and grasping hands and feet with the simultaneous reduction of the nasal region (assumed to correlate with reduced olfaction) by focusing on the necessity for greater dexterity and accurate depth-perception in vision while foraging at the less supportive terminal branches.

Most recently, Sussman has suggested the earliest primates were omnivorous, consuming fruits and associated items (e.g., insects) in terminal branches. Sussman points out that the advent of primates corresponds to the radiation of angiosperms (flowering plants), which would have presented a variety of potential new food resources such as flowers, exudates, and fruits. Here, stereoscopic vision would have been important, but for discrimination between potential foods under the new lower light conditions in the canopies of forests at night. These animals would have been feeding on new resources that were relatively small and would have benefited from an increased visual acuity which would have also improved hand–eye coordination.

Whichever theory best reflects the origin of the order, the fossil evidence for the earliest known primates begins about 55–60 million years ago (late Paleocene/early Eocene). Fossil records indicate that the early primate inhabited woodlands and savannas in the Northern Hemisphere, zones that were uniformly warm throughout the year with little seasonality in temperature or rainfall. However, fossil evidence of primates in tropical or subtropical regions (the primary modern habitat) may simply be underrepresented as preservation and the number of archeological investigations in these geographic regions have not been ideal. It also should be emphasized that the common ancestor of primates likely predates the earliest fossil finds; some molecular evidence has indicated a much earlier date of origin in the Cretaceous, at about 81.5 million years ago.

Most extant nonhuman primates are found in southern Mexico, Central and South America, Africa, Asia, and associated islands (Table I). There are two extant taxonomic groups of primates (Table I), arranged into the Suborders Strepsirrhini (lemurs and lorises) and Haplorhini (tarsiers, monkeys, apes, and humans). The ensuing discussion focuses on nonhuman primates, except as specifically noted.

Habitats, Activity Patterns, and Locomotion

Extant nonhuman primates are widely distributed, being found in all continents except Antarctica and Australia. While most nonhuman primates live in tropical, arboreal habitats, present day habitats also reflect millions of years of dispersal to nontropical regions and nonarboreal substrates. In particular, African primates inhabit savannas, mountains, and limited desert regions, as well as tropical forests (ranging from higher elevations to swamps and flood forests). It may be generalized that the majority of primates spend most of their time in arboreal habitats (Table II). Some may spend nearly equal amounts of time on the ground (semiterrestrial; e.g., *Lemur catta*, Fig. 1A). Relatively fewer primates spend most of their day on the ground (e.g., Fig. 1D; although such primates may sleep and seek shelter in trees). It is noteworthy that most of these primates live in large groups (see below for adaptive reasons) and most are Old World haplorhines (Table II).

Within the variety of arboreal habitats, nonhuman primates use several locomotor strategies (Table II; Fig. 1). In arboreal quadrupedalism (Fig. 1C), fore- and hindlimbs are used relatively similarly to maintain a close contact with branches (although limbs are used differently for feeding). In leaping, primates more rapidly cover distances by using hindlimbs to propel themselves between discontinuous supports. This category simplifies locomotion to a great extent, and one example of this is leaping behavior coupled with a vertical posture when landing (vertical clinging and leaping), which occurs in tarsiers (Fig. 1B), sportive lemurs, sifakas, indris, wooly lemurs, and some lesser bushbabies. Finally, orangutans (Fig. 1E: genus *Pongo*), spider monkeys (genus *Ateles*), gibbons (genus *Hylobates*), and chimpanzees (genus *Pan*) all use a suspensory form of locomotion, which generally involves a more upright posture than seen in arboreal quadrupedalism. Gibbons and spider monkeys use a specialized suspensory behavior termed brachiation, in which the body is propelled via arm-swinging with hands alternating to grasp support branches. In terrestrial habitats, primates move either on all fours (Fig. 1D: terrestrial quadrupedalism, seen in some Old World haplorhines) or, more rarely, upright on two limbs (bipedalism, only used habitually by *Homo sapiens*, and occasionally in some other primates). Specialized forms of terrestrial quadrupedalism are used by chimpanzees and gorillas ("knuckle-walking") and orangutans ("fist-walking"), although suspensory locomotion is arguably as important for great apes (at least at some stages of ontogeny).

Table I Classification and Distribution of Living Primates

Taxonomic group[a] *(common names)*	*Distribution*
Suborder Strepsirrhini	
Infraorder Chiromyiformes	
Family Daubentonidae	Madagascar
Genus *Daubentonia* (aye-aye)	
Infraorder Lemuriformes	
Superfamily Lemuroidea	
Family Indriidae (wooly lemurs, sifakas, indris)	Madagascar
Genera: *Indri, Propithecus, Avahi*	
Family Megalapididae (sportive lemurs)	Madagascar
Genus *Lepilemur*	
Family Lemuridae (lemurs and bamboo lemurs)	Madagascar
Genera: *Lemur*, *Eulemur*, *Varecia*, *Hapalemur*	
Superfamily Cheirogaleoidea	
Family Cheriogaleidae (dwarf and mouse lemurs)	Madagascar
Genera: *Allocebus*, *Cheirogaleus*, *Microcebus*, *Mirza*, *Phaner*	
Infraorder Lorisiformes	
Family Galagonidae (bushbabies)	Africa
Genera: *Eoticus*, *Galago*, *Galagoides*, *Otolemur*	
Family Loridae (lorises, pottos)	Asia (India, SE Asia)
Genera: *Arctocebus*, *Loris*, *Nycticebus*, *Perodicticus*, *Pseudopotto*	
Suborder Haplorhini	
Infraorder Tarsiiformes	
Family Tarsiidae (tarsiers)	Asia (islands of SE Asia)
Genus *Tarsius*	
Infraorder Platyrrhini	
Superfamily Ceboidea	
Family Cebidae	S. America, Central America
Subfamily Callitrichinae (marmosets and tamarins)	
Genera: *Callimico*, *Callithrix*, *Cebuella*, *Leontopithecus*, *Saguinus*	
Subfamily Cebinae (capuchins, squirrel monkeys)	
Genera: *Cebus*, *Saimiri*	
Subfamily Aotinae (owl monkeys)	
Genus: *Aotus*	
Family Atelidae	S. America, Central
Subfamily Atelinae (howler monkeys, spider monkeys, wooly monkeys)	America, Mexico
Genera: *Alouatta*, *Ateles*, *Brachyteles*, *Lagothrix*	
Subfamily Pitheciinae (sakis, uacaris)	
Genera: *Cacajao*, *Chiroptes*, *Pithecia*	
Subfamily Callicebinae (titi monkeys)	
Genus: *Callicebus*	
Infraorder Catarrhini	
Superfamily Cercopithecoidea	
Family Cercopithecidae	Africa, Asia
Subfamily Cercopithecinae (macaques, baboons, mandrills, geunons, patas monkeys, swamp monkeys, vervets)	
Genera: *Allenopithecus*, *Cercocebus*, *Cercopithecus*, *Chlorocebus*, *Erythrocebus*, *Lophocebus*, *Macaca*, *Mandrillus*, *Miopithecus*, *Papio*, *Theropithecus*	

continues

Table I *continued*

Taxonomic group[a] *(common names)*	*Distribution*
Subfamily Colobinae (colobus monkeys, langurs, proboscis monkeys, snub-nosed monkeys)	
Genera: *Colobus*, *Kasi*, *Nasalis*, *Piliocolobus*, *Presbytis*, *Procolobus*, *Pygathrix*, *Rhinopithecus*, *Semnopithecus*, *Simias*, *Trachypithecus*	
Superfamily Hominoidea	
Family Hylobatidae (gibbons and siamangs)	SE Asia and associated islands
Genus: *Hylobates*	
Family Pongidae (orangutans)	SE Asian islands
Genus: *Pongo*	
Family Hominidae (bonobos, chimpanzees, gorillas, humans)	Africa (humans: worldwide)
Genera: *Gorilla, Pan, Homo*	

[a] The taxonomic groupings followed here reflect suborders used by Groves and other researchers. Taxa within Strepsirrhini also are grouped according to Groves. All other taxonomic divisions follow Fleagle. It should be noted that many authors believe platyrrhines and catarrhines comprise a clade, Anthropoidea.

Table II Habitat and Locomotory Patterns of Living Primates

Taxonomic group (common names)	*Habitat*	*Locomotion*
Suborder Strepsirrhini		
Family Daubentonidae	A	AQ
Family Indriidae	A	L (B on ground)
Family Lepilemuridae	A	L
Family Cheirogaleidae	A	AQ (also L for mouse lemurs)
Family Lemuridae	A, S-T	TQ, AQ, L
Family Galagidae	A	AQ, L
Family Lorisidae	A	AQ
Suborder Haplorhini		
Family Tarsiidae	A	L
Family Cebidae		
Subfamily Callitrichinae	A	AQ, L
Subfamily Cebinae	A	AQ, L
Subfamily Aotinae	A	AQ, L
Subfamily Atelinae	A, occasionally T	AQ, S
Subfamily Pitheciinae	A, occasionally T	AQ, L
Subfamily Callicebinae	A	AQ, L
Family Cercopithecidae		
Subfamily Cercopithecinae	A, T, or S-T	AQ, TQ or both (some guenons are also capable leapers; Patas monkeys are most highly terrestrial)
Subfamily Colobinae	A, or S-T	AQ, L (Hanuman langurs use more TQ than any colobine)
Family Hylobatidae	A	S (B on ground)
Family Pongidae	A, T or S-T	S, TQ (fist-walking)
Family Hominidae	A, T, or S-T	S, TQ (knuckle-walking), B (habitually, only in humans)

Note: Habitats: A, arboreal; T, terrestrial; S-T, semi-terrestrial. Locomotion: AQ, arboreal quadrupedalism; TQ, terrestrial quadrupedalism; L, leaping; S, suspensory (including brachiation); B, bipedalism. This table is based on locomotory behavior described for each taxa in the work of Fleagle.

Primates have varied patterns of temporal activity. Activity patterns influence access to resources (e.g., access to prey that may be available for restricted parts of the day). Primates may be active primarily during the day (diurnal), night (nocturnal), or at various times of both day and night (cathemeral). The most diverse primates in regard to activity patterns (or cycles) are the Malagasy strepsirrhines (Table III). Most of these animals are strictly nocturnal (e.g., the cheirogaleids) or cathemeral (most true lemurs). Ring-tailed lemurs are diurnal and bamboo lemurs, though active at various times, are active mostly at dawn and dusk (crepuscular). Whereas most strepsirrhines are nocturnal, only one primate is nocturnal among anthropoids

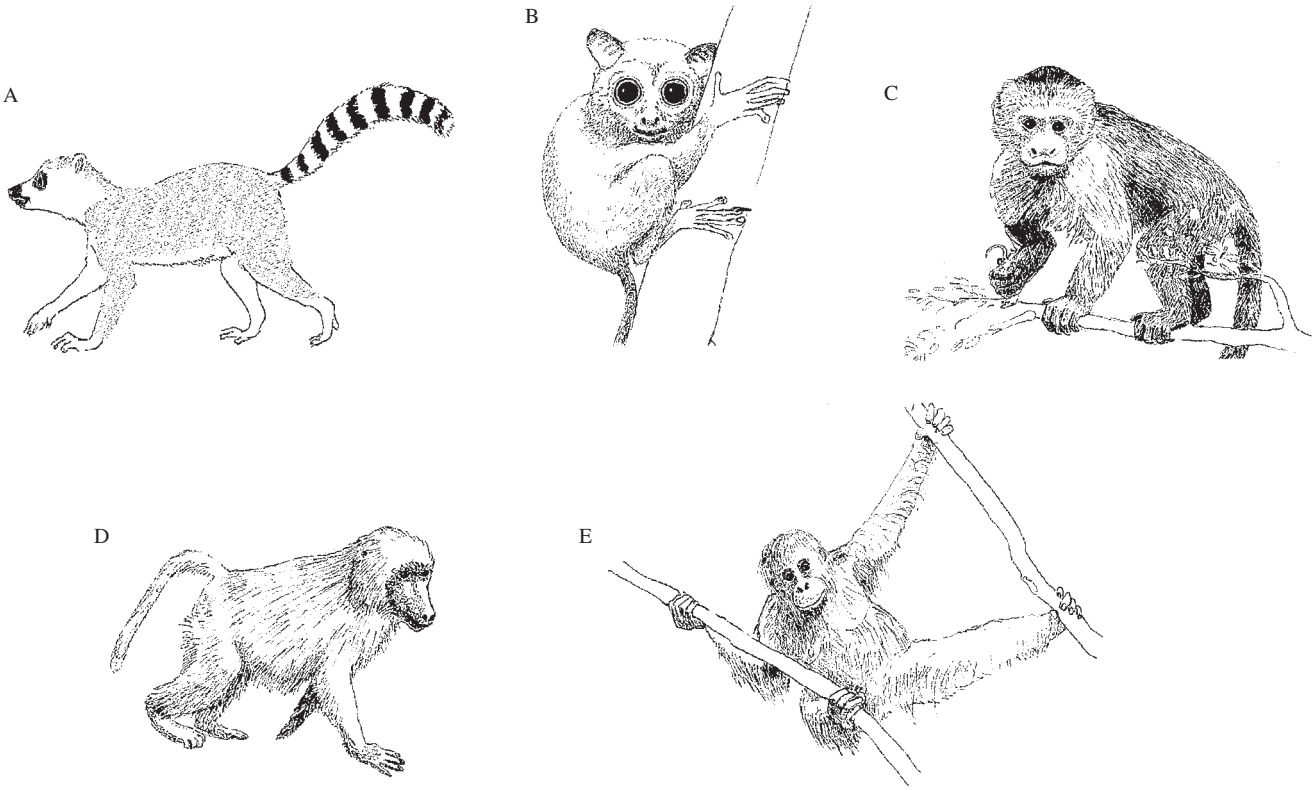


Figure 1 Examples of primates and their different locomotor behaviors. The ring-tailed lemur, *Lemur catta* (A) is semi-terrestrial, whereas most strepsirrhines primarily rely on arboreal quadrupedalism. The tarsier (B) (*Tarsius syrichta* is shown) and some strepsirrhines use leaping to move about tree limbs and adopt a vertical posture after landing. All New World monkeys, such as *Cebus capuchinus* (C), are primarily arboreal quadrupeds. Some Old World monkeys, such as baboons (D) (*Papio hamadryas* is shown), are mostly terrestrial quadrupeds. The great apes utilize a variety of locomotor patterns, including suspensory (the (E) orangutan, *Pongo abellii* is shown). All drawings copyright 2003, T. D. Smith.

(monkeys, apes, and humans), the New World owl monkey (genus *Aotus*). Particular social and ecological implications of these activity patterns are discussed below.

The varying habitats of Asia, Central and South America, and Africa house primates with numerous locomotor patterns. Dispersal beyond forested habitats necessitated adaptations for terrestrial life. Terrestrial locomotion may be employed by any primate for at least short periods of time (e.g., arboreal primates may transiently feed on the ground). However, numerous Old World monkeys and apes (e.g., macaques, baboons, chimpanzees) and one strepsirrhine primate (ring-tailed lemur) spend the majority of their days (all are diurnal) on the ground. Whereas arboreal primates may use arm-swinging or vertical clinging/leaping as the most specialized modes, the most prevalent pattern is arboreal quadrupedalism (Table I). Nocturnal arboreal quadrupeds have small group sizes, sometimes solitary, and this may well describe ancestral primates.

Primate Groups

Social Systems (Table III)

Traditionally, primates have been described according to mating systems (monogamy, polygyny, etc.), but recent authors group primates according to social systems or social groups. The reason for this altered trajectory is an understanding that reproductive behavior is not the sole factor dictating group composition in primates (a more thorough discussion is found in the work of Sussman). Although we use the term "social system" herein, it corresponds conceptually with the "social groups" identified by Sussman. Sussman's definition of

Table III Ecological and Behavioral Aspects of Living Primates (excluding humans)

Genus	*Diet*	*Activity*	*Group size*	*Social system(s)*
Daubentonia	A, Fr, S, N	N	1–2	One adult
Indri	L*, Fr, S	D	2–6	One male/one female
Propithecus	L, S, Fr, Fl	D	2–12	Multi-male/multi-female
Avahi	L*, Fl, Fr, B	N	2–5	One male/one female
Lepilemur	L*, Fr, B, S, Fl	N	1–3	One adult
Allocebus	N?	N	1–6	One adult; one male/one female
Cheirogaleus	Fr, L, Fl, A, Ex, N	N	1–5	One adult
Microcebus	Fr, A, Fl, Ex, N	N	1–5	One adult
Mirza	Fr, A, Fl, Ex, N	N	1–3	One adult; one male/one female
Phaner	Ex*, Fr, A, Fl	N	1–4	One adult
Lemur	Fr*, L, Fl, B	D	5–30	Multi-male/multi-female
Eulemur	L, Fr, Fl, P, A, N[1], S, B	C	2–18	One male/one female; multi-male/multi-female
Varecia	F, S, L, N	D	5–16	One male/one female; multi-male/multi-female
Hapalemur	Bamboo*, Fr, Fl, L	D, C, or Cr	2–12	One male/one female; one male/multi-female; multi-male/multi-female
Eoticus	Ex*, A, Fr	N	1–7	One adult
Galago	A[2], Fr, Ex, S	N	1–4	One adult
Galagoides	A*, Fr, Ex, L	N	1–6	One adult
Otolemur	Ex[3], A, F	N	1–6	One adult
Arctocebus	A*, Fr	N	1–2	One adult
Loris	A*, L, Fl	N	2–4	One adult
Nycticebus	Fr*, A, Ex, L	N	1–?	One adult
Perodicticus	Fr*, Ex, A	N	1–2	One adult
Tarsius	A*	N	1–6	One adult; one male/one female; multi-male/multi-female
Callimico	Fr, A, Ex	D	2–8	One male/one female; multi-male/multi-female
Callithrix	Fr*, A, Ex[4]	D	3–15	One male/multi-female; one female/multi-male; multi-male/multi-female
Cebuella	Ex*, Fr, N, A	D	1–15	One male/one female
Leontopithecus	Ex, N, Fr, A	D	2–16	One male/one female; multi-male/multi-female[8]
Saguinus	Fr, Ex, N, A	D	2–16	One male/multi-female; one female/multi-male; multi-male/multi-female
Cebus	Fr*, N, S, Fl, Ex, A	D	2–20	Multi-male/multi-female
Saimiri	A*, Fr, Fl, N, L	D	10–70	Multi-male/multi-female
Aotus	Fr*, Fl, L, N, A	N	2–5	One male/one female
Alouatta	L*, Fr, Fl	D	2–45	Multi-male/multi-female
Ateles	Fr*, A, N, S, Fl, L[5]	D	2–35	Multi-male/multi-female
Brachyteles	Fr*, S, N	D	7–42	Multi-male/multi-female
Lagothrix	Fr*, L, A, S, Ex	D	2–70	Multi-male/multi-female
Cacajao	S*, Fr, Fl, A	D	5–100	One male/multi-female; multi-male/multi-female
Chiroptes	Fr*, S, A[6]	D	4–20	Multi-male/multi-female
Pithecia	Fr*, S[7], L, Fl, A, B	D	1–5	Multi-male/multi-female

continues

Table III *continued*

Genus	*Diet*	*Activity*	*Group size*	*Social system(s)*
Callicebus	Fr*, L, Fl, A, S	D	2–7	One male/one female
Allenopithecus	Fr*, Fl, A, N, roots	D	?	Multi-male/multi-female
Cercocebus	Fr, S, F1, L, A	D	14–95	Multi-male/multi-female
Cercopithecus	Fr*, A, L, Ex, Fl, S	D	2–60	One male/one female; one male/multi-female; multi-male/multi-female
Chlorocebus	Fr*, S, L, A	D	5–76	Multi-male/multi-female
Erythrocebus	Fr*, S, A, grass	D	5–34	One male/multi-female
Lophocebus	Fr*, L, Fl, A, B	D	6–28	One male/multi-female; multi-male/multi-female
Macaca	Fr*, S, L, Fl, A	D	4–200	One male/multi-female; multi-male/multi-female
Mandrillus	Fr, S, A, roots	D	2–1350	One male/multi-female; multi-male/multi-female
Miopithecus	Fr*, L, Fl, A	D	40–112	Multi-male/multi-female
Papio	Fr, S, L, Fl, A, roots	D	25–750	Multi-male/multi-female
Theropithecus	grass*, S, L, A, bulbs	D	3–20	Multi-male/multi-female
Colobus	S, L, Fr, Fl	D	2–50	One male/multi-female; (occasionally 2-male/multi-female)
Kasi	L*, Fr, Fl, A	D	3–25	One male/multi-female; multi-male/multi-female
Presbytis	L*, Fl, Fr, S	D	2–21	One male/multi-female; multi-male/multi-female
Procolobus	L*, Fl, Fr	D	5–20	One male/multi-female; multi-male/multi-female
Piliocolobus	L*, Fl, Fr	D	5–80	One male/multi-female; multi-male/multi-female
Pygathrix	L, Fl, Fr, S	D	4–40	Multi-male/multi-female
Nasalis	L*, S, Fr, Fl, A	D	4–20	One male/multi-female
Rhinopithecus	Fr, L, S, Fl, lichen	D	3–600	One male/multi-female; multi-male/multi-female
Semnopithecus	L*, Fr, Fl, Ex, A	D	5–100	One male/multi-female; multi-male/multi-female
Simias	L*, Fr	D	2–20	One male/one female; one male/multi-female; multi-male/multi-female
Trachypithecus	L*, Fr, Fl, A	D	3–40	One male/multi-female; multi-male/multi-female
Hylobates	Fr*, L, Fl, A	D	2–12	One male/one female
Gorilla	L*, Fr, Fl, A, roots	D	3–21	One male/multi-female
Pan	Fr*, L, A	D	6–200	Multi-male/multi-female
Pongo	Fr*, L, Fl, B, A	D	1–3	One adult

Note: This table is primarily based on information from Rowe with updates based on Sussman. *Homo sapiens* are excluded above, since they are more variable in all categories than other primates. Diet: A, animal matter or prey; B, bark; Ex, exudates; Fl, flowers; Fr, fruit; L, leaves; N, nectar; S, seeds; *, appears to constitute the largest percentage part of the diet among all components (note that some primates may vary the major dietary component based on seasonal availability, e.g., *Saguinus* spp.); Activity: C, cathemeral; Cr, crepuscular; D, diurnal; N, nocturnal.

[1] Preferred by *E. mongoz*.

[2] In *Galago*, some species appear to primarily feed on animal matter (*G. senegalensis*) while others feed primarily on fruit (*G. alleni*).

[3] *Otolemur crassicaudatus* feeds primarily on exudate, and *O. garnettii* relies equally on fruit and animal matter.

[4] *C. flaviceps, C. jacchus,* and *C. penicillata* consume more exudates than other *Callithrix* spp.

[5] Some species may not consume significant quantities of animal resources.

[6] *C. satanas* has been reported to utilize seeds as primary food.

[7] Seeds are another major dietary element and are the primary food for at least one species.

[8] It is likely that a dominant pair may bond within multi-male/multi-female groups.

a social group is helpful at the outset: ". . . individuals who interact socially more frequently among themselves than with other individuals; group members exhibit different behavior toward nongroup members and occupy the same home range." Below, we provide categories of social systems used by Falk.

1. Multi-male/multi-female (a.k.a., multi-male group). This system includes more than one breeding adult of each sex and offspring and varies greatly in size. In these groups, either males or females may mate with multiple partners (some authors use the term polygamy to describe this practice, rather than its more general definition in the glossary above). Although this appears to refer to overall promiscuity, in most groups there are hierarchical male or female subunits that may limit/enhance opportunities for individuals of either sex. Furthermore, even in multi-male/multi-female groups, it may be the case that one individual male mates with multiple females (polygyny) or one female mates with multiple males of a subgroup (polyandry). Polygynous subgroups occur in some baboons and langurs. Multi-male/multi-female systems are more prevalent in haplorhines than strepsirrhines (Table III) and appear to occur in the context of relatively abundant resources, a variable degree of predation pressure, and diurnal activity patterns (see below for further discussion).

2. One male/multi-female group (a.k.a., one-male group). This system includes one dominant, breeding adult male that mates with multiple females in the group, reflecting a polygynous mating system. In these species, such as the gorilla, juvenile males stay with the natal group until they are of reproductive age. They then typically migrate out of the group to locate new groups where they may have access to females.

Certain complexities should be noted. For instance, one male/multi-female subunits may exist *within* larger multi-male/multi-female groups (e.g., some baboons). In addition, all male groups may occur in species that primarily form one male/multi-female social systems. Such groups may be temporary alliances that create opportunities to oust a male that currently holds tenure in a one male/multi-female group (see Strier for further discussion and references therein to the work of Hrdy). These are not indicated in Table III, but they are seen in proboscis monkeys and langurs, for example. Infanticidal attacks by competing males may occur in one male/multi-female groups (e.g., in gorillas, at least some langurs), a strategy that quickly makes new reproductive opportunities available.

3. One female/multi-male (a.k.a., cooperative polyandrous) group. This system includes one breeding adult female that mates with two or more adult males in the group (polyandry), along with offspring of various ages. This is a stable family group, that is, males do not leave the group and mate with other females. This typifies numerous callitrichines. Interestingly, these primates are quite comfortably monogamous in captive settings, and it was not until field data were available that group composition became clearer.

The number of males per group in primate social systems appears to depend on multiple factors, some of which are discussed further below. In the case of callitrichines, males may tolerate one another because paternity may be uncertain and twinning and relatively high infant body mass make care-giving highly taxing. In other words, the reproductive benefits gained by receiving help with infant care outweigh costs associated with uncertain paternity.

4. All male group. In some primates, males may form sex-exclusive, temporary associations that may cooperate in gaining access to resources and/or females. Many langurs have stable one-male, multi-female systems that must be maintained against such groups (see above). More long-term alliances occur in other primates such as chimpanzees and bonobos, and these alliances appear to increase access to resources for group members.

5. One male/one female (a.k.a, pair bonds). This system consists of small family groups, with one breeding adult female, one breeding adult male (i.e., monogamous pair), and their offspring. This system is common in some callitrichines, and in gibbons, siamangs, and bamboo lemurs. The circumstances and benefits of this social system are discussed more below. It should be noted that there are instances of pair bonded animals within multi-male/multi-female groups (e.g., lion tamarins).

6. One adult (a.k.a., solitary but social). This system is found in most nocturnal primates and orangutans. Adult males mate with more than one adult female but do not participate in the care of offspring. The solitary behavior is mostly foraging. Whereas adult males avoid one another, males associate with multiple females. Adult nesting groups exist in some species, although most nesting groups consist of adult females and their subadult offspring. This system is present in most nocturnal strepsirrhines (Table III).

Influences on Primate Social Systems

The distribution of resources influences foraging strategies for female primates, who have high energetic costs associated with their own metabolic needs, especially during pregnancy and lactation (see below for more discussion). The way in which female primates are distributed (and perhaps dispersed) hinges on the availability and distribution of resources (food, habitat). Male reproductive options, in turn, are limited by the resulting spatial distribution of females.

The approach to obtaining resources and/or maximizing reproductive fitness, may be further delineated as female and male social strategies. It seems such strategies

are most frequently about access to resources first, and then about reproduction. If all of the above factors allow, primates appear more likely to form groups than not. In such contexts, male strategies for acquiring resources may become extremely intense, involving more direct and frequent confrontations with other males or other groups. For males, the ability to monopolize access to females dramatically maximizes their reproductive fitness (provided they are effective in doing this). When reproductive prospects are low (for migrating males or all male groups), another seemingly extreme male strategy is infanticide. This is quite common among one male/multi-female or multi-male/multi-female groups, and has a potentially high payoff since the mother could be fertilized again by a new male much sooner than otherwise possible. Also in such groups, female primates may employ more frequent polyandrous matings with invading males as a counter strategy. By keeping paternity uncertain, polyandrous matings can have a high payoff in either protecting infants from infanticidal males or, in the case of callitrichines, recruiting added parental care (e.g., many callitrichines).

In large primate groups, male or female alliances and hierarchies represent a major important social strategy. Female alliances are another important influence on the likelihood of infant survival in multi-male/multi-female groups. Male–female relationships may even dictate duration of male tenure in one-male, multi-female groups. The dynamics of these relationships have a major role in access to food or reproductive opportunities for each sex, as discussed in much more detail by Strier.

Communication

Communication between primates may be grouped under broad categories of visual, acoustic, and chemical forms. Primate groups use all forms of communication, with varying emphases. The term "solitary," used in some contexts to describe primates of the one adult social system, may be misleading. The one-adult social system is actually characterized by numerous interactions if one considers the different nature of communication used by such primates (mostly nocturnal strepsirrhines—Table III). In the nocturnal activity pattern, chemical (either scent and/or pheromonal) cues may be much more effective long range signals than the acoustic signals preferred by diurnal primates. These signals may carry valuable information ranging from territory boundaries to reproductive status, i.e., information that might be more readily acquired by other special senses in diurnal activity. Broadly speaking, nocturnal primates rely more heavily on chemical signals than diurnal primates (e.g., see the work of Hrdy and Whitten regarding reproductive signals), although diurnal primates certainly use olfactory signals. Among primates, by far the best evidence for a functional "accessory olfactory" sense, via the vomeronasal organ, exists for strepsirrhines. In at least some strepsirrhine primates, the vomeronasal organ likely is sensitive to pheromones conveying reproductive information, whereas the precise function of this organ is more uncertain in platyrrhines, and probably negligible in catarrhines that possess it.

Diurnal primates in general tend to have relatively greater emphasis on visual/acoustic signals than nocturnal species. Communication in callitrichine primates is highly interesting in that the use of scent marks is extremely frequent compared to most other haplorhines. It is likely that scent may convey complex information that even may allow individual identification of primates. Chemical communication may be of a more subtle nature in other instances, where pheromones appear to play a role. Pheromones are substances released by one individual that may have a neuroendocrine effect on another individual (in other words, these signals may alter the physiology of a conspecific). One notable putative pheromonal effect in callitrichines is reproductive suppression of one female by the dominant female in a group. Interestingly, this behavior appears to not be pheromonally influenced in certain callitrichines (lion tamarins), which instead intimidate subordinate females (for further discussion, see Dixson).

Communication within multi-male/multi-female groups is highly complex. For instance, females in this system have the most obvious behavioral (e.g., presentation of hind-quarters) or anatomical (reddening of genitalia) cues of proceptivity among primates. Interesting exceptions exist regarding anatomical cues, notably in callitrichines that may have large group sizes, but have polyandrous subgroups. Multi-male/multi-female groups may use a highly complex repertoire of vocalizations to communicate about food resources and predators over large distances. Facial displays can be very elaborate in members of these groups, communicating information among different animals in hierarchies, for example.

Ecological and Behavioral Influences on Primate Social Systems

Temporal, spatial, and geographic availability of resources and predators have critical influences on the geographical distribution of primates. These same factors may be strong selective influences on the activity cycles of a species which, in turn, can limit the prevalent type(s) of communication. This cascade of influences has a profound importance to social systems of primates, as described below. More subtle determinants of group dynamics are not considered below, such as conflicting female and male strategies regarding reproduction and

meeting metabolic needs (readers are referred to Strier for further discussion).

Diet and Feeding Behavior

Perhaps the greatest driving force in the life of any primate is the need to obtain enough calories on a daily basis in order to survive. Arguably, securing enough food of a high enough quality is the most important waking activity. Of course, different primates consume different foods depending upon their size, teeth, gastrointestinal tract, etc. As a general rule of thumb, smaller primates tend to consume high-quality protein-rich foods that are relatively easy to digest while larger primates tend to consume lower quality foods that take a longer time to digest (see Table III). For example, some species of the diminutive tarsiers consume only large insects and small vertebrates such as snakes and birds while the massive gorilla consumes mostly tough mature leaves. This dietary phenomenon is closely associated with basal metabolic rate (BMR). Relatively small primates, such as the tarsier, have relatively high BMRs and require a high amount of energy for routine daily physiologic functions. In order to support these functions, small primates require a diet of easily digestible calorie-rich foods that do not need a great deal of mechanical effort and time to collect them. Such a diet would typically be represented by other small animals (large insects, reptiles, etc.) and perhaps supplemented by fruits and/or exudates. This is the case seen in many small primates such as the tarsiers, bushbabies, and marmosets. Large primates, such as orangutans and gorillas, have relatively low BMRs and are not under such intense pressure to consume high-quality easily digestible diets. Such large-bodied primates tend to be folivorous and consume lower quality but highly available foods including tough mature leaves, roots, and wood with supplementation from fruits. In addition to orangutans and gorillas, other large primates such as gibbons and baboons pursue this dietary niche. Mature leaves represent a low-quality diet that is difficult to digest but one that is generally available in large predictable quantities and varies little in seasonal availability (imagine the amount of insects a gorilla would need to consume on a daily basis in order to gain enough calories!). Large primates with a low BMR can tolerate the increased time necessary to digest such a poor-quality diet while smaller primates with a high BMR must consume food that can be relatively quickly broken down.

With this dietary quality/quantity information, we can now examine how dietary needs have a role (perhaps the largest role) in shaping primate social systems. The three major food categories for primates are fauna (including large insects and small vertebrates), fruits, and leaves. Other sources such as roots, exudates, and flowers are not considered to be major sources of calories but, for some primates such as the needle-clawed bushbaby (*Eoticus* spp.) and the mongoose lemur (*Eulemur mongoz*), are the major source of calories (see Table III for their occurrences across genera). Distribution of these resources occurs spatially (in clumps or more evenly spread out) and temporally (seasonal occurrences or regular availability). In foraging spatially for a particular food, primates must consider the potential quality of that food as well as what it will cost them to obtain enough of that food in order to meet their caloric needs. For example, is the tree which bears the desired fruit found in clumps with others (requiring little travel effort and time to harvest the fruit) or does it occur in only isolated locales (requiring long travel distances and potentially greater exposure to predators)? Does the desired insect swarm in a particular location or are they found only singly here and there on tree branches?

Most foods occur in clumps, whether they are small clumps or much larger ones. By and large, it is the size of the clump that determines how large a primate group can feed there at the same time. A food clump that is relatively large, such as a savanna of grasses or leaves from trees in a rainforest, can support a large number of individuals feeding there at the same time. Primate folivores thus tend to occur in large groups or troops, such as the hamadryas baboons of the east African savannas (up to 750 individuals per troop, Table III) and the rhesus macaques of the southeastern Asian forests (up to 200 individuals per troop, Table III). A large number of individuals feeding at a clump will indeed compete with one another for the food but there are potential benefits that outweigh such a cost. More individuals in one area can defend the clump of food against other species or other conspecifics from different troops as well as provide increased predator detection. Increased predator detection may provide increased survival chances for one's offspring, siblings, and extended kin, individuals who have a high likelihood of having the same genes. Thus, most primate folivores occur in large groups or troops (see Table III). Notable exceptions include the most massive primates, gorillas. These animals are indeed folivores but occur in much smaller group sizes, around 20 individuals. Whereas their food resources are relatively abundant and found in clumps, such a large primate is not subject to predator pressure (humans excluded). This relaxation of predator pressure may influence the relatively small group size in these folivores.

Frugivorous primates do indeed most commonly occur in groups but group size is typically far smaller than in folivores. Fruits tend to grow on trees and are present in clumps but they are relatively less abundant than the leaves that grow on trees and are often more seasonal in availability. In addition, because the fruit is less clumped than leaves or grasses, a frugivorous primate often requires high travel time and distance in order to

obtain enough calories. In addition to spatial clumping of fruits, frugivorous primates must deal with the temporal clumping of fruits. Many ecological zones, such as rainforests, produce fruits in a highly seasonal pattern. In these circumstances, frugivorous primates feed almost exclusively on fruits when in season but supplement their diets with such resources as nectar and exudates during the off season. Such a drastic divergence in temporal availability of fruits along with their spatial clumping patterns may not, then, support the large number of individuals in a group that are seen in folivores. Indeed, many frugivorous primates such as the lion tamarins, spider monkeys and mangabeys occur in groups from about 15 to 60 individuals.

Faunal resources such as large insects and small vertebrates tend to be spread out more randomly and at greater geographic intervals than leaves and fruits. Thus, a faunivorous dietary niche usually cannot support a large number of individuals feeding on them at one time. For example, large insects and snakes tend to occur singly or in small groups, not in densely concentrated clumps. These small irregular patches of food could not support more than one or a few individuals feeding at a time together. Faunivorous primates, such as tarsiers, tend to occur singly or in small groups of up to three individuals.

Territoriality is in part a direct consequence of how dense each food patch is. If a food is densely concentrated in a patch it may be easily defensible in a primate's home range. However, the availability and quality of food plays a part in determining the worth of defending it. Evenly spaced, abundant, low-quality foods such as grasses would probably not be worth defending but clumped, higher quality foods such as fruits or large insects may be worth the potential costs of defending.

In addition to spatial distribution, food availability can vary temporally. Seasonal variation in sunlight, rainfall, and temperature affect plant productivity and faunal availability. Primates typically occur in ecological zones that have a distinct rainy season, such as tropical forests. During the rainy season there is great availability of high-quality foods such as fruits and young leaves. Consequently, primates tend to feed exclusively on these abundant foods during the rainy season. While these foods are abundant, individuals tend to focus on these foods exclusively and form smaller feeding parties, splitting up the larger group or troop. During the dry season, high-quality foods tend to become more scarce and most primates are forced to diversify their diets to include less desirable, low-quality foods such as mature leaves and roots. At these times individuals tend to reaggregate into the larger group or troop.

In primate species that group into large numbers, such as the hamadryas baboon and the rhesus macaque, there are, consequently, large numbers of females concentrated in a relatively small area. Female primates will be found where the food is. There is, then, the potential for one or a few males to control access to all of the reproductive females. While females go where the food is, males go where the females are. A male who is strong enough, showy enough, fast enough, or has enough allies of these characteristics can control other males' access to females of reproductive age. Thus, polygyny (one male having more than one female mate) tends to be seen in primate species that group in large numbers. Primates such as baboons, macaques, guenons, langurs, and others that feed on abundant low-quality foods (such as leaves and grasses) live in the most extreme polygamous multimale/multi-female groups (Table III). In these groups, one or a few males guard and mate with a large number of females or "harems." It is only these strong, showy, dominant males that usually mate with the females. Other males tend to migrate out of the natal group into new groups while females tend to stay with their natal group for life.

Frugivorous primates are also polygamous multi-male/multi-female groups but tend to be less extreme than the folivorous primates. In these species one male may control access to female, but there are typically fewer females controlled by the male. In such groups, females may have more access to reproductive opportunities with other males.

Primate species that feed on higher quality foods that tend to be relatively rare and spread out are usually solitary or occur only in small groups that consist of a mother and her young offspring. These species, such as the nocturnal lemurs and bushbabies, tend to be promiscuous, although at least some tarsiers occur in monogamous pairs (Table III).

In those species that consume foods that are relatively high quality and are regularly available with little travel time between clumps, such as exudates, some fruits, and bamboo, monogamy may be seen (Table III). If resources are predictable and relatively abundant, females may not gain an advantage by grouping together. In addition, concentrated clumps of high-quality food may favor the development of territoriality in these species, such as gibbons, siamangs, and the pygmy marmoset. While the male of these species gives up a potentially high reproductive rate by bonding with only one female, he does gain a higher paternal certainty and offspring survival rate, since he provides a great amount of parental care.

Predation Pressure

Virtually all primate species are subjected to predator pressure and some are themselves predators. Predators such as hawks, eagles, and other primates (excluding humans) can have a profound influence on social systems. Primates that exist in regions where predators are present in considerable numbers and who are small enough to be prey tend to group in relatively large numbers. Species

taken as prey, such as the red colobus monkey and guenons, tend not to occur in groups as enormous as the larger-bodied grass/leaf eaters (the baboons and macaques), but they do aggregate into fairly large groups (up to 35 individuals).

In addition to intraspecific associations, many small- to medium-sized primates subject to predator pressure associate with other primate species under similar pressure during feeding and sleeping bouts. While such an association during feeding may decrease foraging efficiency and total calorie intake, it may provide increased predator detection and, ultimately, increase survival. For example, red colobus and Diana monkeys inhabit tropical forests of western Africa and both feed on leaves in similar geographic areas. However, these species often forage together when one of their main predators, the chimpanzee, is near. When these species forage together they change the composition of their diets but gain the potential benefit of increased predator detection.

Larger bodied primates typically have low predation pressure. These species tend to be more stable in their associations with other primate species. However, some of these primates, especially male chimpanzees and baboons, are themselves predators. Chimpanzee males are significant and regular predators of red colobus monkeys in western Africa. Male chimps regularly form "hunting parties" that prey upon several primate species. The parties typically consume the meat themselves without sharing it with females or other males.

Activity Patterns

Predation pressure is one causal mechanism of nocturnality in primates, and also may influence body size (nocturnal primates are generally smaller) and group size (see Table II). Predator avoidance is not the sole cause of nocturnality or small group size in mammals. As illustrated by some solitary, nocturnal predators of primates, access to resources is another important influence. Although not generally considered a direct influence on social systems, activity patterns themselves impose important limitations on social interactions. For example, nocturnal mammals in general rely more heavily on nonvisual (e.g., auditory, olfactory) and/or longer lasting (e.g., scent-marking) modes of communication compared to diurnal mammals. Nocturnality itself reinforces small group size in primates by diminishing the impact of close range modes of communication.

Conclusion: Influences on Primate Social Systems

Primate social systems vary widely, from one adult to multi-male/multi-female groups. Recent studies on primate ecology and behavior focused on the nature and distribution of food resources and predation pressure as prime influences on primate social systems, influencing group size and structure. More subtle dynamics of primate interactions may also be attributed to such factors, such as the often conflicting female and male strategies regarding reproduction and meeting metabolic needs. Thus these ecological factors may have a cascade of direct and indirect effects, influencing activity patterns (nocturnality, diurnality, etc.) and even types of interactions among group members.

See Also the Following Article

Behavioral Psychology

Further Reading

Cartmill, M. (1992). New views on primate origins. *Evolution. Anthropol.* **1,** 105–111.

Dixson, A. F. (1998). *Primate Sexuality.* Oxford University Press, Oxford.

Falk, D. (2000). *Primate Diversity.* Norton, New York.

Fleagle, J. G. (1999). *Primate Adaptation and Evolution,* 2nd Ed. Academic Press, San Diego.

Goodall, J. (1996). *The Chimpanzees of Gombe: Patterns of Behaviour.* Belknap Press (Harvard University Press), Cambridge, MA.

Groves, C. P. (2001). *Primate Taxonomy.* Smithsonian Institution Press, Washington, DC.

Heymann, E. W. (2000). The number of adult males in callitrichine groups and its implications for callitrichine social evolution. In *Primate Males. Causes and Consequences of Variation in Group Composition* (P. M. Kappeler, ed.), pp. 64–71. Cambridge University Press, Cambridge, UK.

Hrdy, S. B., and Whitten, P. L. (1987). Patterning of sexual activity. In *Primate Societies* (B. B. Smuts, D. L. Cheney, R. M. Seyfarth, R. W. Wrangham, and T. T. Struhsaker, eds.), pp. 370–384. University of Chicago Press, Chicago.

Leigh, S. R. (1994). Ontogenetic correlates of diet in anthropoid primates. *Am. J. Phys. Anthropol.* **94,** 499–522.

Rowe, N. (1996). *The Pictorial Guide to the Living Primates.* Pogonias Press, Charlestown, RI.

Strier, K. B. (2000). *Primate Behavioral Ecology,* 1st Ed. Allyn & Bacon, Boston.

Sussman, R. W. (1999). *Primate Ecology and Social Structure. Vol. 1: Lorises, Lemurs and Tarsiers.* Pearson, Needham Heights, MA.

Sussman, R. W. (2000). *Primate Ecology and Social Structure. Vol. 2: New World Monkeys.* Pearson, Needham Heights, MA.

Tavaré, S., Marshall, C. R., Will, O., Soligo, C., and Martin, R. D. (2002). Using the fossil record to estimate the age of the last common ancestor of extant primates. *Nature* **416,** 726–729.

van Schaik, C. P., van Noordwijk, M. A., and Nunn, C. L. (1999). Sex and social evolution in primates. In *Comparative Primate Socioecology* (P. C. Lee, ed.), pp. 204–231. Cambridge University Press, Cambridge, UK.

Probit/Logit and Other Binary Models

William D. Berry
Florida State University, Tallahassee, Florida, USA

Glossary

additive model A model in which the effect of each independent variable on the dependent variable is the same regardless of the value of the other independent variables.
binary variable A variable that has only two possible values (typically labeled zero and one).
instantaneous effect The slope of a probability curve for an independent variable.
latent variable An unobserved variable presumed to measure an observed variable.
linear model A model in which the effect of each independent variable on the dependent variable is the same regardless of the value of that independent variable.
logit of a binary dependent variable The log of the odds that a binary dependent variable equals one (rather than zero).
odds of an event The probability that the event will occur relative to the probability that it will not.
probability curve A graph of the relationship between an independent variable and the predicted probability that a binary dependent variable equals one (rather than zero) when the remaining independent variables are held constant at specified values.

Probit and logit are techniques for estimating the effects of a set of independent variables on a binary (or dichotomous) dependent variable. When ordinary least squares is used to estimate a binary dependent variable model, the model is often called a linear probability model (LPM). Probit and logit avoid several statistical problems with LPMs and generally yield results that make more sense.

Introduction

Probit and Logit: An Alternative to Regression When the Dependent Variable Is Binary

Multiple regression is the most widely used technique in the social sciences for measuring the impacts of independent (or explanatory) variables on a dependent variable. Regression—more technically, ordinary least squares (OLS) regression—generally assumes that the dependent variable is continuous. Yet many of the dependent variables social scientists wish to study are not continuous. Indeed, many have only two possible values (an event did, or did not, occur), and are termed binary (or dichotomous) variables: for example, whether a nation is at war, a candidate wins an election, an organization adopts some innovative practice, or an adult is married. Other statistical methods are more appropriate when the dependent variable is binary, and probit and logit are the most common.

A Review of Regression

Consider a regression model with a continuous dependent variable Y and two independent variables assumed to influence Y: X, which is continuous, and D, which is dichotomous (scored either zero or one):

$$Y = \alpha + \beta X + \delta D + \varepsilon. \quad (1)$$

The equation expresses the value Y for any observation as a function of its values for X, D, and an error (or disturbance) term, ε. ε represents those variables influencing Y (in addition to X and D) that have not been measured and included in the model. α, β, and δ are termed parameters and are constants; the latter

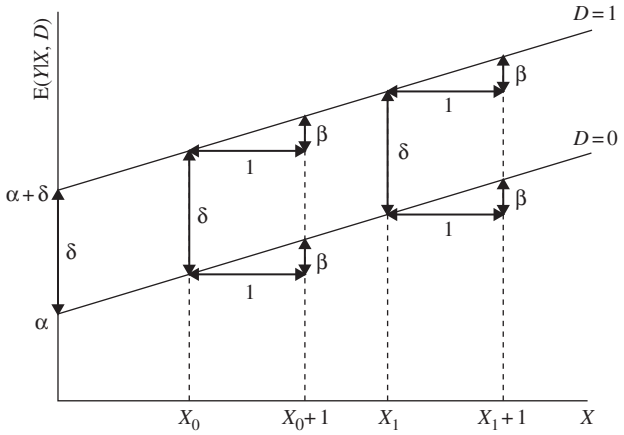


Figure 1 The population regression function (PRF) [Eq. (3)] for a linear additive regression model (LARM) [Eq. (1)].

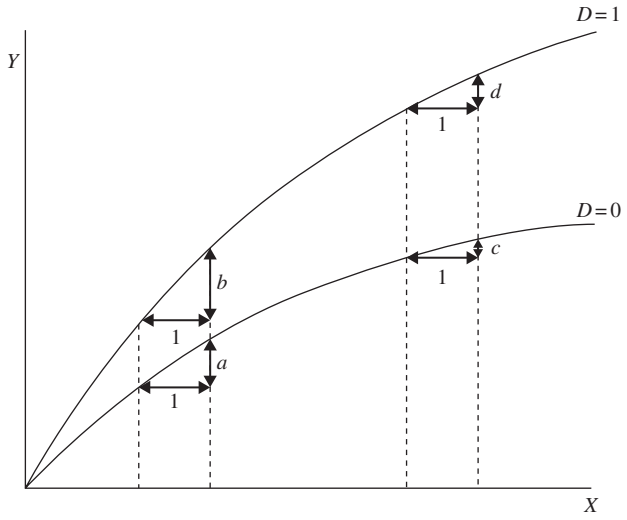


Figure 2 A nonlinear, non-additive model.

two—the parameters for the independent variables—are called partial slope coefficients.

It is generally assumed that the error term in Eq. (1) has a conditional mean (or expected value) of zero. If so, the regression model of Eq. (1) implies a population regression function (PRF) of the form

$$\mathrm{E}(Y \mid X, D) = \alpha + \beta X + \delta D. \qquad (2)$$

The left side of Eq. (2) is a conditional mean and is read "the expected value of Y, given X and D." The PRF writes the expected (or mean) value of Y (given specific values for X and D) as a function of the values of X and D. Notice that the error term ε is not in the PRF; while this variable influences the value of Y for a particular observation, if ε has a mean of zero for any fixed values of X and D, then the mean of Y for a case is fully determined by X and D, and is not influenced by the value of ε. The PRF is graphed in Fig. 1 in a format showing the relationship between X and the expected value of Y for both values of D (0 and 1).

In this model, the slope coefficient β characterizes the impact of X on Y: it tells the change in the expected value of Y resulting from an increase of one in X when the other independent variable, D, is held constant. For example, Fig. 1 shows that when D is held constant at one (see the higher line), and the value of X rises by one from x_0 to $x_0 + 1$, the expected value of Y grows by β (capital letters indicate variables, and lower case letters indicate specific values that a variable assumes). Similarly, δ reflects the impact of D, as it indicates how much Y can be expected to change if D is increased from zero to one while X remains fixed. In the graph, δ is the vertical distance between the lines reflecting cases for which D equals 0 and D equals 1.

Linearity and Additivity

Equation (1) can be called the linear additive regression model (LARM). A model is termed linear if it assumes that the effect of each independent variable on the dependent variable is the same regardless of the value of that independent variable. The linearity of the equation is evident from the fact that both curves in Fig. 1 take the form of a straight line: the expected response of Y to an increase of one in X is the same whether we begin at x_0, at x_1, or at any other value of X. A model is called additive if the effect of each independent variable on the dependent variable is the same regardless of the value of the other independent variables in the model. The additivity of Eq. (1) is reflected in the two lines in Fig. 1 being parallel: X has the same effect on Y whether D is zero or one, and D has the same impact on Y at any fixed value of X.

But consider Fig. 2. This graph reflects a model that is neither linear nor additive. The model is non-additive because at any fixed value of X, the effect of X on Y is different when $D = 0$ than when $D = 1$: at any value of X, Y responds more to a unit change in X when $D = 1$ than when $D = 0$. (For example, in Fig. 2, $b > a$ and $d > c$.) The model is nonlinear because at either value of D, the impact of X on Y depends on the value of X: this impact declines gradually as X gets larger. (For example, $c < a$ and $d < b$.) Thus, Fig. 2 presents a nonlinear, non-additive model, in which the effect of each of the independent variables depends on the value of both.

The Linear Probability Model: Using OLS Regression with a Binary Dependent Variable

When one ignores the typical assumption made in OLS regression that the dependent variable is continuous, and employs it with a binary dependent variable, the model is frequently labeled the linear probability model (LPM).

Consider the PRF for an LPM with a single continuous independent variable, X, and a dependent variable, Y, with two possible values, zero and one.

$$E(Y \mid X) = \alpha + \beta X. \tag{3}$$

This equation is graphed in Fig. 3 along with the X and Y values for some hypothetical observations. The fact that the dependent variable can be only zero or one means that all observations in the graph fall on one of two horizontal lines ($Y = 0$ or $Y = 1$).

The Meaning of Slope Coefficients in a Linear Probability Model

To understand the nature of the slope coefficient β in the LPM of Eq. (3), consider the conditional mean on the left side. Since Y can assume only the two values zero and one, the expected value of Y given the value of X is equal to 1 multiplied by "the probability that Y equals 1 given the value of X" plus 0 multiplied by "the probability that Y equals 0 given the value of X." Symbolically,

$$\mathrm{E}(Y \mid X) = (1)[\mathrm{P}(Y = 1 \mid X)] + (0)[\mathrm{P}(Y = 0 \mid X)]. \tag{4}$$

Of course, one multiplied by any value is that value, and zero multiplied by any number remains zero; so Eq. (4) simplifies to

$$\mathrm{E}(Y \mid X) = [P(Y = 1 \mid X)]. \tag{5}$$

Given this equality, $\mathrm{P}(Y = 1 \mid X)$ can be substituted for $\mathrm{E}(Y \mid X)$ in Eq. (3), obtaining

$$\mathrm{P}(Y = 1 \mid X) = \alpha + \beta X. \tag{6}$$

This revised form of the PRF for the LPM—with the probability that Y equals 1 as the dependent variable—implies that the slope coefficient β indicates the change in the probability that Y equals 1 resulting from an increase of one in X.

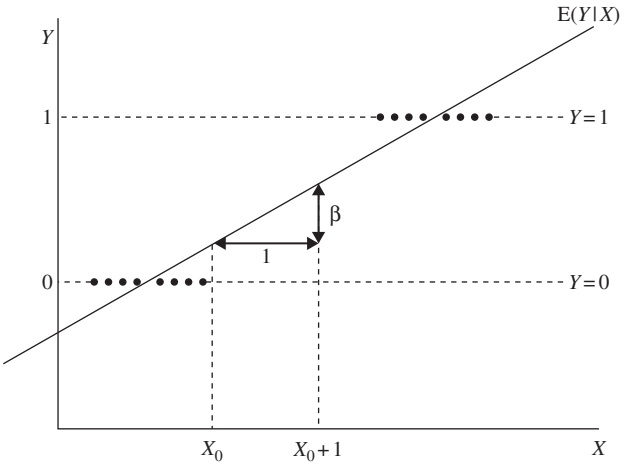


Figure 3 The population regression function (PRF) [Eq. (3)] for a linear probability model (LPM).

In the more general LPM with multiple independent variables (say k of them—$X_1, X_2, \ldots, X_k$), the PRF takes the form

$$\begin{aligned}\mathrm{E}(Y \mid X_1, X_2, X_3, \ldots, X_k) &= \mathrm{P}(Y = 1 \mid X_1, X_2, X_3, \ldots, X_k) \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &\quad + \cdots + \beta_k X_k.\end{aligned}$$

If (bold italicized) $\mathbf{X}$ is used as a shorthand for all the independent variables, this equation simplifies to

$$\begin{aligned}\mathrm{E}(Y \mid \mathbf{X}) = \mathrm{P}(Y = 1 \mid \mathbf{X}) &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &\quad + \cdots + \beta_k X_k.\end{aligned} \tag{7}$$

The slope coefficient, β_i, for independent variable X_i (where i can be 1, 2, 3, . . . , k) can be interpreted as the change in the probability that Y equals 1 resulting from a unit increase in X_i when the remaining independent variables are held constant.

An Illustration of the Linear Probability Model: Class Performance

Data collected by Spector and Mazzeo (1980) on the performance of students taking a course in macroeconomics can be used to illustrate the LPM. The dependent variable, denoted GRADE, is a student's class performance, measured dichotomously. For convenience, the two outcomes will be called "success" (GRADE = 1) and "failure" (GRADE = 0). There are three independent variables in the LPM: (1) the student's entering grade point average, denoted GPA, and ranging from 2.06 to 4.00 in the sample, (2) an exam score indicating the student's knowledge at the beginning of the course, denoted BEGIN, and ranging from 17 to 29 in the sample, and (3) a dichotomous variable distinguishing two teaching methods, an experimental approach called PSI and the traditional method. The last variable is labeled METHOD:PSI and is scored 1 for PSI and 0 for the traditional approach.

Using OLS regression with Spector and Mazzeo's data yields the coefficient estimates in column 1 of Table I. The slope coefficient for GPA is 0.46, indicating that when a student's pre-course knowledge and teaching method are held constant, an increase of one in a student's entering grade point average—a little over half the range from 2.06 to 4.00—is estimated to produce an increase of 0.46 in the probability that the student will succeed in the class. In particular, the estimated probability of success for a student with GPA of 2.5, a BEGIN value of 21.94 (the mean value in the sample), and who is taught by the PSI method (i.e., METHOD:PSI = 1) is 0.27 [calculated as $-1.4980 + 0.4639(2.5) + 0.0105(21.94) + 0.3786(1)$].

Table I Coefficient Estimates for Binary Dependent Variable Class Performance Models

Independent variable	*(1) OLS coefficient estimate*	*(2) Probit coefficient estimate*	*(3) Logit coefficient estimate*	*(4) Instantaneous effect (with all variables at their mean)*[a]
GPA	0.4639[b]	1.6258[b]	2.8261[b]	0.5333
(s.e.)	(0.1620)	(0.6939)	(1.2629)	
Z	2.864	2.343	2.238	
BEGIN	0.0105	0.0517	0.0952	0.0167
(s.e.)	(0.0195)	(0.0839)	(0.1416)	
Z	0.538	0.616	0.672	
METHOD:PSI	0.3786[b]	1.4263[b]	2.3787[b]	0.4679
(s.e.)	(0.1392)	(0.5950)	(1.0646)	
Z	2.720	2.397	2.234	
Intercept	−1.4980	−7.4523	−13.0214	
Likelihood ratio statistic[c]		15.55	15.40	

[a] Based on probit estimates in column 2.
[b] Different from zero at 0.05 level of significance (1-tail test).
[c] Distributed as chi-square with 3 degrees of freedom.

In contrast, the predicted chance of success for a student with GPA one unit higher, at 3.5, but the same values for GPA and BEGIN is 0.73: 0.46 higher. The slope coefficient for METHOD:PSI is 0.38. This means that the predicted probability of success of a student taught by the experimental PSI method is 0.38 greater than the chance of success of a student trained using the traditional method but having the same GPA and BEGIN values.

Weaknesses of the Linear Probability Model

Non-normally Distributed Error Term

An assumption that the error term, ε, is normally distributed, in combination with other assumptions of OLS regression, can be used to justify various techniques of statistical inference (e.g., hypothesis testing). However, the assumption that the error term is normally distributed cannot hold for an LPM: given any fixed values for the independent variables, ε can assume only two values, and a dichotomous variable is not normally distributed.

Heteroscedasticity

OLS regression assumes that the error term, ε, is homoscedastic, i.e., that the variance of ε is constant. In an LPM, because the dependent variable is constrained to two possible values, this assumption is violated; the variance of the error term is larger for some values of the independent variables than others, a condition called heteroscedasticity. Goldberger (1964) demonstrated that this problem can be overcome by using a weighted least squares (WLS) procedure for estimation instead of OLS. But even when heteroscedasticity is avoided by this means, other more serious problems remain.

Predicted Probabilities with Nonsensical Values

We have seen that the dependent variable in the PRF for an LPM can be interpreted as the probability that Y equals 1 [i.e., $P(Y = 1 \mid \mathbf{X})$], and thus should be no less than zero and no greater than one. Unfortunately, there is nothing that constrains predicted values that Y equals 1—based on either OLS or WLS estimation of an LPM—to the range between zero and one. For example, the OLS estimates of the class performance model in column 1 of Table I imply that the predicted probability of success for a student taught by the traditional method (i.e., METHOD: PSI = 0) with GPA at 2.5 and a BEGIN value of 21.94 (the mean in the sample) is $-0.11[= -1.4980 + 0.4639(2.5) + 0.0105(21.94) + 0.3786(0)]$. This prediction is nonsensical, and illustrates a significant flaw in the LPM.

Linearity and Additivity Cannot Hold

Just like the LARM, the LPM is both linear and additive. But the linearity and additivity of the LPM are at odds with the fact the dependent variable in the PRF for an LPM can be viewed as the probability that Y equals 1 [i.e., $P(Y = 1 \mid \mathbf{X})$]. Consider any one of the independent variables in an LPM: say X_i. The linearity of the model implies that a unit increase in X_i results in the same change—β_i—in the probability that Y equals 1, regardless of the value of X_i. However, if every unit increase in X_i leads to a change of β_i, no matter the size of β_i, eventually

X_i will become large enough to push the probability that Y equals 1 out of its acceptable range (between 0 and 1). Beyond the mathematical imperative, the linearity and additivity of the LPM often makes no sense substantively. Returning to the class performance model, assume that at certain values of the independent variables (BEGIN, GPA, and METHOD:PSI), the probability that a student will succeed is 0.50. It is certainly conceivable that for students with these independent variable values an increase in GPA (or any of the other independent variables) could lead to a substantial increase in the probability of success. In contrast, assume that at different values of the independent variables, the probability of success is 0.99. Given this very high chance of success at the outset, no increase in GPA (or any other variable) could appreciably increase the probability of success. Thus, a nonlinear, non-additive model—allowing the effects of independent variables on the probability of success to differ depending on the values of the independent variables—would be superior.

Other Binary Dependent Variable Models: Probit and Logit

Probit and logit assume that the effects of independent variables on the probability that Y equals 1 are nonlinear and non-additive. In particular, these models assume that the effects of independent variables decline in magnitude as $P(Y=1\,|\,\mathbf{X})$ approaches either zero or one. For most applications involving binary dependent variables (BDVs), this assumption will be quite sensible. Figure 4 presents a graph of the probit model with a single independent variable. The graph of the logit model is very similar. Indeed, the shapes of the curves are so close that probit and logit can be viewed as functional equivalents in applied research settings. With both probit and logit, the gradual flattening of the S-shaped curve as X gets either small or large makes clear that as X becomes large or small, the effect of X on $P(Y=1\,|\,X)$ diminishes in strength.

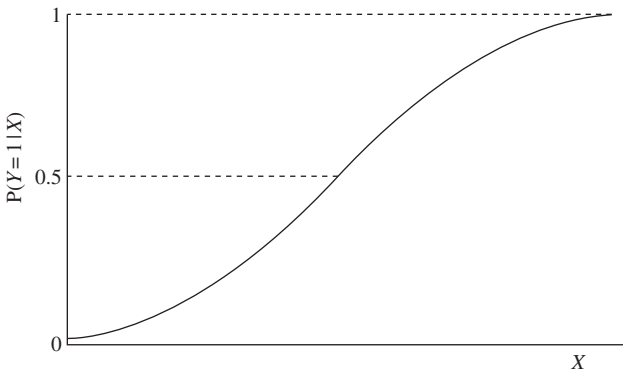


Figure 4 The probit model with a single independent variable, X.

A Latent Variable Model: One Derivation of Probit and Logit

Assume a continuous dependent variable, Y^c, that cannot be observed directly. But a binary indicator, Y, of Y^c is observable. In particular,

$$Y = 1 \quad \text{if} \quad Y^c \geq T^*$$

and

$$Y = 0 \quad \text{if} \quad Y^c < T^*,$$

where T^* is some threshold value for Y^c. In effect, T^* is a cut-point that divides values of Y^c into two groups: lower than T^*, and greater than or equal to T^*. Y^c is referred to as a latent variable (or unobserved indicator) for Y.

Assume that a set of independent variables have linear and additive effects on Y^c:

$$Y^c = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon. \quad (8)$$

Although Eq. (8) takes the form of a LARM, it cannot be estimated using OLS regression because Y^c is not observable. But the model can be estimated using maximum likelihood (ML) estimation, as long as some arbitrary assumptions about the value of T^* (typically that it equals zero) and the distribution of the error term, ε, are made. Specifically, both probit and logit assume that $E(\varepsilon\,|\,\mathbf{X}) = 0$. In probit, it is assumed, in addition, that ε is normally distributed (for all possible values of the independent variables) with a constant variance of one. In logit, it is assumed that ε has a logistic distribution (for all values of the Xs) with a constant variance of $\pi^2/3$.

An Alternative Derivation of Probit and Logit: Transforming $P(Y=1\,|\,\mathbf{X})$ to an Unbounded Variable

One can derive probit and logit without introducing the notion of a continuous latent variable. We start with the PRF for the LPM:

$$E(Y\,|\,\mathbf{X}) = P(Y=1\,|\,\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k. \quad (9)$$

It was previously shown that a problem with this model is that $E(Y\,|\,\mathbf{X})$ is the probability that Y equals 1, yet the linearity and additivity of the model ensures that $E(Y\,|\,\mathbf{X})$ is not constrained to be in the range between zero and one, the only possible values for a probability. The solution

is to use a mathematical function to transform (1) $P(Y=1\,|\,\mathbf{X})$ from the range between zero and one to the range between negative infinity and positive infinity so that it is consistent with the range of $\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots+\beta_kX_k$, or (2) the unbounded value, $\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_kX_k$, to the range between zero and one so that it is consistent with the range of $P(Y=1\,|\,\mathbf{X})$. There are countless mathematical functions that accomplish this transformation.

With logit, $P(Y=1\,|\,\mathbf{X})$ is transformed first into the odds that Y equals 1 (i.e., the probability that Y equals 1 relative to the probability that Y does not equal 1):

$$\text{the odds that } Y=1 \text{ given } \mathbf{X} = \frac{P(Y=1\,|\,\mathbf{X})}{1-P(Y=1\,|\,\mathbf{X})} = \frac{P(Y=1\,|\,\mathbf{X})}{P(Y=0\,|\,\mathbf{X})}.$$

The odds (being the ratio of two non-negative values) cannot be negative, but as $P(Y=1\,|\,\mathbf{X})$ approaches one, the ratio approaches infinity, so the constraint to values less than one has been eliminated. Taking the logarithm of the odds removes the constraint to values greater than zero, arriving at a value that can range from negative infinity to positive infinity. The logit model assumes that the log of the odds that Y equals 1—also called the logit of Y— is a linear and additive function of the Xs:

$$\ln\left[\frac{P(Y=1\,|\,\mathbf{X})}{1-P(Y=1\,|\,\mathbf{X})}\right] = \alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_kX_k. \quad (10)$$

With probit, $\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\ldots+\beta_kX_k$ is transformed to the range between zero and one using the cumulative distribution function (CDF) for the standard normal distribution.

Assessing the Impacts of Independent Variables

What Probit and Logit Coefficients Tell Us Directly

Since both probit and logit can be derived by assuming that independent variables have linear and additive effects on an unbounded latent variable, Y^c [see Eq. (8)], the probit or logit coefficient for independent variable X_i can be interpreted as the expected change in the latent variable associated with a unit increase in X_i when all remaining independent variables are held constant. Just as in the case of regression, a positive value for β_i indicates that when other independent variables are held constant, an increase in X_i tends to lead to an increase in the latent variable, whereas a negative coefficient means that an increase in X_i tends to prompt a decrease in Y^c. For example, the probit and logit ML coefficient estimates for GPA are positive (see columns 2 and 3 of Table I). This means that when BEGIN and METHOD:PSI are held constant, increases in a student's GPA tend to produce increases in the latent variable that the observed dichotomous variable, GRADE, measures—presumably a continuous measure of course performance. The positive coefficient estimates for METHOD:PSI imply that a student exposed to the experimental method (METHOD:PSI = 1) is expected to outperform one trained through traditional means (METHOD:PSI = 0) yet having the same values for BEGIN and GPA. Additionally, the standard errors of the coefficient estimates for METHOD:PSI and GPA suggest that these two variables have effects on class performance that are statistically significant.

Thus, probit and logit coefficients tell us the direction (positive or negative) of estimated effects of independent variables, and whether these effects are statistically significant. However, since the measurement scale for the latent variable is unknown, probit and logit coefficient estimates offer little directly interpretable information about the strength of the effects of independent variables. Fortunately, probit and logit coefficients can be used to calculate other statistics that can be very valuable for assessing the magnitudes of impacts.

Analyzing Changes in Predicted Probabilities that Y Equals 1

The best way to clarify the magnitude of the impact of an independent variable on a dependent variable is to estimate how much the dependent variable shifts when the independent variable is changed by a specified amount while the remaining independent variables are held constant. Armed with either probit or logit coefficient estimates for a model, one can predict the probability that Y equals 1 for any set of values for the independent variables. (Using software called CLARIFY, recently developed by Gary King and associates, one can also easily compute a confidence interval for this predicted value.) Such predicted probabilities permit a characterization of the magnitude of the impact of any independent variable, X_i, on $P(Y=1\,|\,\mathbf{X})$ through the calculation of the change in the predicted probability that Y equals 1 that results when X_i is increased from one value to another while the other independent variables are fixed at specified values. With probit and logit, since the effects of the independent variables on $P(Y=1\,|\,\mathbf{X})$ vary depending on the values of all independent variables, characterizations of the impact of a variable must specify the values at which the other independent variables are fixed. Most often, researchers report impacts when other independent variables are held at central values (i.e., their mean or their mode), but sometimes other values of theoretical interest are analyzed.

When assessing the strength of the impact of X_i, a variety of types of changes in X_i may be considered. For

a variable with a readily interpretable measurement scale, a change representing an easily described small increment is appropriate: for example, $1000 in annual income, 1 year of age, or 10 points on an IQ test. Studies have also reported the response of the predicted value of $P(Y = 1 \mid \mathbf{X})$ to a change based on specified locations within an independent variable's distribution: for instance, a change in the variable (1) from the lowest value in the sample to the highest, or (2) from one or two standard deviations below the mean to a comparable amount above the mean. With a binary independent variable, the choice of increment to analyze is simple, since the only possible change in the value of a binary variable is from one value to the other.

The need to select a single specific increment in X_i can be avoided by constructing a probability curve: a graph showing the relationship between X_i and the predicted value of $P(Y = 1 \mid \mathbf{X})$ over the range of X_i values in the sample, when the other independent variables are fixed at specified values. Such a graph can also include vertical "bars" around predicted $P(Y = 1 \mid \mathbf{X})$ values that depict confidence intervals for the predicted values. Figure 5 presents two probability curves reflecting the impact of a student's grade point average on class success, constructed using the probit estimates. (Because the graph is designed to show more than probability curves, bars indicating confidence intervals are excluded in the interest of visual clarity.) Both probability curves show the relationship between GPA and the predicted probability of class success assuming that knowledge at the beginning of the course (BEGIN) is held constant at its mean value in the sample (21.94). The upper curve shows the relationship for students taught by the experimental method (METHOD:PSI = 1); the other shows the relationship for students taught using the traditional approach (METHOD:PSI = 0). It can be seen, for example, that the predicted probability of success for a student with GPA equal to 2.5, with a BEGIN value at the mean, and who is taught by the traditional method is about 0.03. (Compare this to the nonsensical prediction produced by the LPM discussed earlier.)

Both curves have a substantial positive slope over a wide range of values of GPA, indicating that, overall, the effect of GPA on the probability of class success is strong among both groups of students. But for students with low GPAs and taught traditionally, even when they have a BEGIN value at the mean, the effect of GPA is quite weak. An increase in GPA from its lowest value in the sample (2.06) to the 25th percentile (2.80) leads to an increase in the predicted probability of success from 0.017 to 0.060, for a probability difference of only 0.043. Also note that, overall, the effect of GPA on the probability of success is stronger among students taught by the experimental approach than among students taught by traditional means. We can see this by calculating the predicted response of the probability of success to an increase in GPA from the lowest value in the sample (2.06) to the highest (4.00) among students taught by traditional methods to the same response in the PSI group. In the PSI group, the probability of success increases by 0.79 (from 0.11 to 0.90) as GPA increases from 2.06 to 4.00; in the traditional group, the same increase in GPA prompts an increase of 0.54 in the probability of success (from 0.02 to 0.56). The "difference in increases," 0.25 ($= 0.79 - 0.54$), is substantial. But because the sample

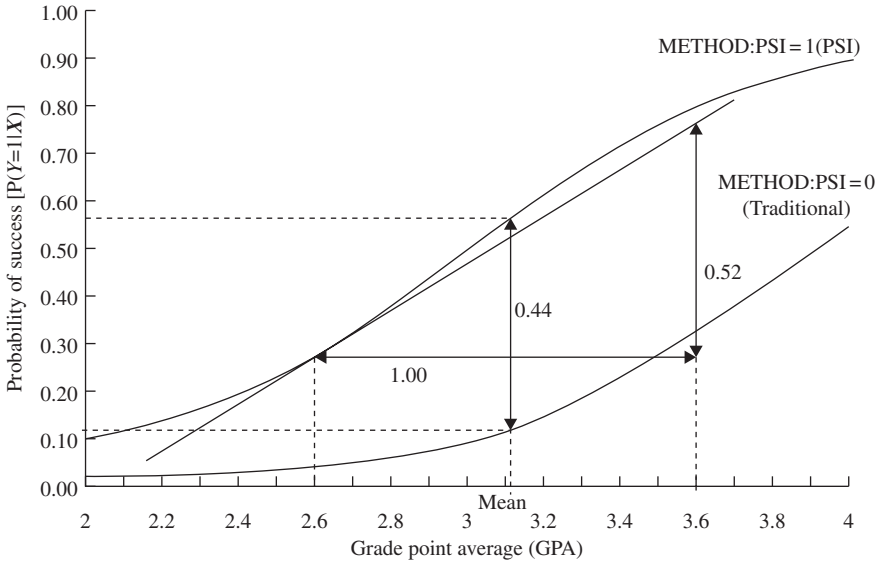


Figure 5 Probability curves for the class success model (based on the probit results in column 2 of Table 1). Note that this graph assumes that BEGIN is fixed at 21.94, its mean in the sample.

is very small ($n = 32$), the 95% confidence interval for the probability difference is quite wide: from just above 0.00 to 0.57.

We can also analyze predicted probabilities to assess the nature of the effect of method of instruction on student success. The impact of method of instruction is indicated in Fig. 5 by the vertical distance between the probability curves. When both GPA and BEGIN are set at their mean (for GPA, 3.117), students exposed to the traditional teaching method have a predicted probability of success of 0.12, but PSI-taught students are predicted to have a 0.56 chance of success, for a probability difference of 0.44 ($= 0.56 - 0.12$). Again, the 95% confidence interval for this probability difference is wide: from 0.09 to 0.74. At lower levels of GPA, the effect of teaching method is weaker, but the 95% confidence interval for the probability difference stays above zero over the entire range of GPA values in the sample.

Calculating Instantaneous Effects

Given a probability curve showing the estimated relationship between an independent variable (say X_i) and $P(Y = 1 \mid \mathbf{X})$ when the remaining independent variables are held constant, one could determine the slope of the curve at any value of X_i, x_i. (The slope of the curve at $X_i = x_i$ is defined as the slope of the line tangent to the curve at this point.) For example, Fig. 5 shows the tangent to the upper curve when GPA = 2.6, and indicates that the slope at this value is 0.52. Just as a partial slope coefficient in a LARM can be interpreted as a measure of the magnitude of the effect of an independent variable, so too can the slope of a probability curve. Indeed, the slope of a probability curve for an independent variable is often described as a measure of the instantaneous or marginal effect of the variable. (Note that the instantaneous effect of X_i is the derivative of $P(Y = 1 \mid \mathbf{X})$ with respect to X_i.)

For a LARM, a partial slope coefficient can be used to determine the expected response of the dependent variable to a change of any amount in the independent variable. If the partial slope coefficient for an independent variable, X_i, is β_i, then a k unit increase in X_i results in a change of $k\beta_i$ in the expected value of Y, no matter the value of k. In contrast, because the probability curve is not straight, the slope of a probability curve at a particular value of GPA does not give sufficient information to calculate the response of $P(Y = 1 \mid \mathbf{X})$ to any discrete change in GPA. In fact, because in probit and logit the effects of variables on $P(Y = 1 \mid \mathbf{X})$ are nonlinear and non-additive, the instantaneous effect of an independent variable depends on the value of that variable and all other independent variables. When BEGIN is at its mean and METHOD:PSI = 1, GPA has the greatest instantaneous impact (0.65) when GPA takes the value 3.01; this is the value of GPA at which the probability curve has its steepest rate of ascent.

Probably the most common practice when reporting the instantaneous effect of a variable is to calculate it assuming all independent variables are held at central values. Column 4 of Table I reports the instantaneous effects of independent variables in the class performance model (based on the probit results) when all independent variables are held at their mean. Instantaneous effect estimates are most meaningful when all independent variables are continuous, since "holding all independent variables at their mean" nicely reflects a hypothetical case with typical values. However, if one of the independent variables is dichotomous (like the method of instruction in the class performance model), holding the variable at its mean yields a value that does not exist in the sample, making it difficult to glean substantive meaning from an instantaneous effect estimate.

Calculating Changes in the Odds (with Logit)

Another approach to interpreting the effects of variables in a BDV model is estimating the change in the odds that Y equals 1 associated with a given increase in an independent variable. (Recall that the odds that Y equals 1 is the ratio of the probability that Y equals 1 to the probability that Y does not equal 1.) This method is applicable when logit is used, but not probit. Assume that the logit coefficient estimate for an independent variable, X_i, is β_i. One hundred multiplied by $[\exp(\beta_i) - 1]$ yields the percentage change in the odds that Y equals 1 resulting from a unit increase in X_i when the remaining independent variables are held constant. Equivalently, it can be said that a unit increase in X_i changes the odds that Y equals 1 by a multiplicative factor of $\exp(\beta_i)$. This factor is independent of both the starting value for X_i and the values at which the other independent variables are fixed, and thus $\exp(\beta_i)$ is a single value that can summarize the effect of X_i. This is an advantage of $\exp(\beta_i)$ over characterizations of independent variable impacts based on predicted probabilities—which are specific to particular values for the independent variables.

Consider the logit coefficient estimate for METHOD: PSI in our exam success model: 2.38. The antilog of 2.38, exp(2.38), is 10.80. This means that a shift from a student being taught traditionally to being taught by the PSI method increases her predicted odds of success by a factor of 10.80. [If her odds of success when being taught traditionally were 1.5, her odds would be (1.5) (10.8) = 16.2 under PSI.] Subtracting 1 from 10.80 yields 9.80, which when multiplied by 100 becomes 980. Thus, we can also say that the change in method of instruction increases the odds of success by 980%. [1.5 increased by 980% = 1.5 + (9.8) (1.5) = 16.2.] If for some independent variable $\beta_i < 0$, this implies that an increase in X_i decreases the odds that Y equals 1. For example, if $\beta_i = -0.50$,

then $\exp(\beta_i) = 0.61$, and $0.61 - 1.00 = -0.39$. This means that when X increases by one, the odds that Y equals 1 decreases by 39%; equivalently, the odds that Y equals 1 changes by a multiplicative factor of 0.61.

Statistical Inference and Goodness of Fit

Many of the possibilities for statistical inference with probit and logit mirror the options available in regression. Software for estimating probit and logit models generally reports standard errors for individual coefficient estimates. The standard errors can be used, just as with a regression model, to determine for each independent variable whether the estimated effect of the variable is statistically significant. The running example has shown that the effects of both GPA and method of instruction on class performance are statistically significant (at better than the 0.05 level). One can also conduct joint hypothesis tests for a subset of coefficients (e.g., that $\beta_1 = \beta_2 = 0$), or a test of the null hypothesis that the coefficients for all independent variables are zero (which serves as a measure of the goodness of fit of the model). Both of these tests are based on a likelihood ratio statistic that approximates a chi-square distribution. In the class performance illustration, the chi-square statistic for the goodness of fit of the model is a bit over 15 (with 3 degrees of freedom) for both the probit and logit analyses, indicating significance at better than the 0.01 level.

Probit and logit do not yield the R^2 coefficient produced by regression. A number of "pseudo-R^2" coefficients that range between zero and one have been developed and can serve as goodness-of-fit measures for a probit or logit model. However, none of these statistics has achieved wide acceptance as a standard measure of fit.

Another common goodness-of-fit measure for probit and logit is the so-called percent correctly predicted (PCP). This is computed by using the coefficient estimates along with the values of independent variables for cases in the sample to predict $P(Y = 1 \mid \mathbf{X})$ for each case. Every case for which $P(Y = 1 \mid \mathbf{X})$ is estimated to be greater than 0.5 is predicted to have a Y value of 1; all cases for which $P(Y = 1 \mid \mathbf{X})$ is less than 0.5 are predicted to have $Y = 0$. Then, these predicted Y values are compared with the observed Y values for cases, and the proportion of the cases for which the two agree is calculated. The higher this value, the better the presumed fit of the model. The value of this statistic declines, however, as the distribution of the dependent variable becomes skewed. This is because the PCP cannot drop lower than the proportion of cases in the modal category (PMC) of the dependent variable. The proportional reduction in error (PRE) is a statistic that overcomes this weakness of the PCP by indicating how much a probit or logit model reduces prediction error over the model that predicts that all cases have the modal value of the dependent variable. The PRE can be calculated by

$$\text{PRE} = (\text{PCP} - \text{PMC})/(1 - \text{PMC})$$

For the class performance model, the PCP is 0.813 with either probit or logit. In the data set, the modal category of GRADE is "failure" (0), with 65.6% of the cases. Thus, $\text{PRE} = (0.813 - 0.656)/(1 - 0.656) = 0.157/0.344 = 0.456$. In effect, the naïve model assuming that all cases are in the modal category of the dependent variable is in error for 34.4% (100% − 65.6%) of the cases. Since the PCP by the probit/logit model is 15.7% (i.e., 81.3% − 65.6%) greater than the PCP by the naïve model, the probit/logit model reduces the error in prediction by 45.6% = (15.7%/34.4%).

See Also the Following Articles

Contingency Tables and Log-Linear Models • Ordinary Least Squares (OLS)

Further Reading

Aldrich, J. H., and Nelson, F. D. (1984). *Linear Probability, Logit and Probit Models.* Sage, Newbury Park, CA.

Golberger, A. S. (1964). *Econometric Theory.* John Wiley, New York.

Greene, W. H. (2000). *Econometric Analysis.* Prentice Hall, Upper Saddle River, NJ.

Hagle, T. M., and Mitchell, G. E., II. (1992). Goodness-of-fit measures for probit and logit. *Am. J. Polit. Sci.* **36,** 762–784.

Hanushek, E. A., and Jackson, J. E. (1977). *Statistical Methods for Social Scientists.* Academic Press, Orlando, FL.

King, G. (1989). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Cambridge University Press, Cambridge, UK.

King, G., Tomz, M., and Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *Am. J. Polit. Sci.* **44,** 341–355.

Liao, T. F. (1994). *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models.* Sage, Thousand Oaks, CA.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage, Thousand Oaks, CA.

Spector, L. C., and Mazzeo, M. (1980). Probit analysis and economic education. *J. Econ. Ed.* **11,** 37–44.

Tomz, M., Wittenberg, J., and King, G. (2001). *CLARIFY: Software for Interpreting and Presenting Statistical Results.* Version 2.0. Harvard University, Cambridge, MA.

Problem-Solving Methodologies

Ton de Jong
University of Twente, Enschede, The Netherlands

Glossary

card sorting Assessment technique for charting the structure of knowledge in the knowledge base of the problem solver. Subjects have to arrange cards with information. The content of this information depends on the goal of the assessment.
problem A perceived gap between a current situation and a desired situation that the problem solver does not immediately know how to bridge.
problem representation The mental representation of a certain state of a problem.
problem-solving process The transformation of the start state of a problem to the goal state.
thinking aloud Assessment technique in which subjects solve a problem while concurrently reporting aloud on their problem-solving process and the evolving problem representation.
transfer problems Problems different from originally learned problems that can be used to measure the range of applicability of problem-solving knowledge.

Problem solving is the process in which a person wants to reach a desired situation from a present one and does not know immediately what to do. Problems can be classified on two dimensions: the amount of domain knowledge needed to solve the problem and the ways the start state, goal state, and operators of the problem are defined. In the problem-solving process, a distinction is made between states of a problem, the processes involved in taking the problem from the start state to the goal state, and the relevant knowledge base of the problem solver. A number of assessment techniques are available for measuring the problem states, the problem-solving processes, and the knowledge base.

Problems and Problem Solving

Problem solving is a widely esteemed capability and much of our education aims to teach learners problem solving in some kind of domain. In 1945, Duncker defined a problem as a perceived gap between where a person is (the current situation) and where he or she wants to be (the goal situation) and, in addition, when the person does not know immediately how to cross the gap. According to Duncker, problem solving involves "thinking," namely devising steps that will take one from the current to the desired situation; Robertson also discussed problem solving in a 2001 book.

Problems exist in many variations and in many domains. There are physics problems, algebra word problems, design problems, etc. In the literature, there are two general classifications of problems. One is on the dimension that involves the amount of domain knowledge that is necessary to solve the problem (semantically rich vs. semantically poor problems), and the second dimension has to do with the way the start state, end state, and necessary operators in the problem are defined (well-defined vs. ill-defined problems).

Semantically rich (or knowledge-rich) problems are those that require knowledge of a specific domain to be solved. Of course, these domains can be very diverse, including medicine, biology, and physics. Semantically rich problems require their own specific operations in moving from the current problem state to the goal state. In physics problems we probably have to use physics procedures, in mathematics problems we may have to apply operations such as differential equations, and in

medical problems we use medical knowledge to link symptoms with diseases. On the other side of the spectrum are semantically poor (or knowledge-lean) problems. These problems do not require prior domain knowledge; basically, all the information needed is in the problem statement. Puzzle-type problems fall into this category; examples, discussed by Simon and Hays in 1976, are the missionaries and cannibals problem and the Tower of Hanoi. The second dimension on which problems can be characterized is the level of "definedness." Well-defined problems are those for which the start state is well described, the goal state is clear, and the necessary operators are in principle known. In contrast, ill-defined problems have a start state that is not fully defined, it is not always clear when the goal state is reached precisely, and new operators may be necessary. Examples of ill-defined problems come, for example, from design. In designing a building, there may be much negotiation at the start of the project. One may debate (and one generally does) if the final design meets the requirements (and is esthetically well done), and a creative element is involved in the design process, as described, for example, by Goel and Pirolli in 1992. Ill-defined problems (which are sometimes called wicked problems) are problems that generally require collaboration of people with different expertise, as discussed by Van Bruggen *et al.* (2003).

The Problem-Solving Process

Whatever the kind of problem or the domain involved, the problem-solving process can be described in general terms. In 1957, Polya divided problem solving into the following four stages: understanding the problem, devising a plan, carrying out the plan, and looking back. Polya's seminal work has been of great importance for our first understanding of problem solving and for the design of courses on how to learn to solve problems. Later, more detailed and cognition-oriented descriptions of the problem-solving process have been presented. Figure 1 gives an overview of the problem-solving process based on a 1996 model by de Jong and Ferguson-Hessler.

In Fig. 1, a distinction is made between states and processes. The states represent (i) the problem statement, which is the description of the problem as presented to the problem solver; (ii) the problem representation, which is the internal representation of the problem as created by the problem solver; (iii) the problem solution, which is in fact the final problem representation including the "answer" to the problem; and (iv) the knowledge base, which comprises the prior knowledge that the problem solver brings to the task. The processes presented are (i) selective perception, which is the process through which the problem statement is "filtered" and in which the selection of information to be included from the problem statement in the problem representation is made; (ii) information retrieval, which is the process of selecting relevant information from the prior knowledge of the problem solver; and (iii) the problem-solving process, which transforms the problem representation from the start state to the goal state.

Figure 1 can be used to describe the problem-solving process of all kinds of problems. Differences between problems will only emerge when the states and processes are examined in more detail. For semantically poor problems, for example, no information, apart from general reasoning strategies, has to be retrieved from memory. For wicked problems, the development of an initial problem representation will be more complex than for a semantically poor problem.

Assessment of Problem Solving

Figure 1 may also form the basis for classifying techniques on how to measure problem solving and its related components. Basically, all the processes and knowledge states from Fig. 1 can be measured.

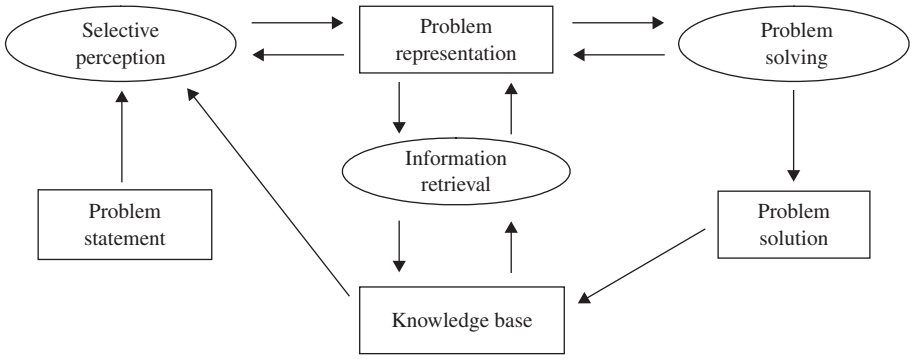


Figure 1 Schematic overview of the problem-solving process.

The problem solution is the aspect that is measured most often. In schools, students are mostly graded by the solution they give to a problem. For well-defined problems this is quite straightforward: there is only one good solution and this is straightforwardly evaluated. For ill-defined problems, solutions are less easy to assess. A rubric to judge the solution will be necessary and in some cases (e.g., architectural design) a reliable assessment will not be easily reached.

The problem representation is the evolving representation of the problem in the mind of the problem solver. Problem solvers alter the problem continuously while solving, either by adding information from the knowledge base or by applying operators that change the state of the problem. The starting state is the problem statement as given to the problem solver; the ultimate representation is the problem's solution. This solution is normally externalized, which is not necessarily the case for the intermediate states. Assessment procedures, therefore, need to include mechanisms to create access to the problem representation.

The knowledge base that the problem solver brings to the task is, of course, multifaceted. An adequate measurement of the characteristics of the knowledge base considers the knowledge base from the perspective of problem solving. In this respect, in 1996 de Jong and Ferguson-Hessler distinguished situational knowledge (knowledge of characteristic problem situations), conceptual knowledge (concepts and principles), procedural knowledge (domain-specific operations), and strategic knowledge (general approaches in the problem-solving process) and recognize that, in addition, the organization of these types of knowledge in the mind of the problem solver is especially important. Assessment methodologies for the knowledge base differ with respect to the aspect that is being measured and include techniques such as reproduction, card sorting, and concept association.

In the actual problem-solving process we can distinguish between the procedures used by the learner (procedures bring about the changes from one problem state (representation) to another) and the overall strategy used. Procedures are mostly domain specific (called strong methods). The overall strategy includes the way the problem solver finds his or her way through the problem space. These strategies or "weak" methods include approaches such as depth first, breadth first, hill climbing, and means-ends analysis. Also, overall strategies that divide the problem-solving process into distinct phases, as discussed by Polya in 1957, can be mentioned here. Measurement techniques for charting the problem-solving process include thinking aloud and cued recall.

In the information-retrieval process, the problem solver "decides" what knowledge to use from his or her knowledge base. This information can be specific concepts, procedures, etc. Together with the information that was selected from the problem statement, this information forms the basis for the problem-solving process. For this process, the situational knowledge that a problem solver brings to the task is especially important. Problem solvers can have schemata of situations that help them to add relevant information to the problem situation as presented. This is not a trivial issue since studies have shown that problem solvers often have relevant knowledge in their knowledge base that they use in solving a problem. Perfetto *et al.* in 1983 called this inert knowledge. The information-retrieval process can be measured by, for example, think-aloud techniques.

Selective perception is the process in which the problem solver decides what to use from the problem statement and, as a consequence, also what not to use. This process is strongly guided by the problems solver's situational knowledge. Selective perception can be assessed by using eye tracking techniques.

In the following sections, measurement techniques for assessing the previously mentioned aspects of problem solving are discussed. First, an analysis of an assessment technique that is most obvious and most widely used is discussed—that is, having people solve problems. Then, a number of assessment techniques that have an emphasis on measuring either processes or states are presented. The techniques discussed are empirical (and not logical) and conceptual (and not statistical) methods.

Measurement by Solving Problems

Examinations often consist of problems to be solved. As such, they measure all aspects of the knowledge base together with the actual problem-solving process. By changing the characteristics of the problems to be solved, an emphasis on different knowledge and process aspects can be achieved. A theoretical task analysis of the problems may clarify the knowledge needed to solve the problem. In the case of domain-specific problems (e.g., physics and medicine), there is, of course, an emphasis on knowledge from the specific domain. However, problems may also focus on more general skills. For example, the PISA 2003 framework includes a test for the assessment of general problem-solving skills. The test focuses on three general types of problems: decision making, system analysis, and design and troubleshooting. Analytic reasoning, quantitative reasoning, analogical reasoning, and combinatorial reasoning are required for solving these type of problems. At an intermediate level, there are problem-solving tests that focus on problem-solving abilities in specific areas (e.g., science).

A specific characteristic of problem-solving knowledge and skills is the range of applicability. In transfer test, problems are offered that differ from the problems a subject knows how to solve, and the idea is that when

a known solution procedure has general character, it will facilitate solving the transfer problems or facilitate the learning of how to solve the transfer problems. Transfer problems can take many forms, such as variations in context or problem-solving procedure.

One disadvantage of offering learners problems to solve is that normally only the problem solutions become available. Problem solvers may have achieved their result through very different routes, and, maybe even more important, problem solvers may have failed to reach the solution for many different reasons (e.g., miscalculations). One solution might be to ask problem solvers for intermediate results. Another approach could be to examine notes that problem solvers take.

A second disadvantage of placing people in a traditional problem-solving situation is that only an aggregate of performance can be measured. Snow (1989) made a strong plea for the development of instruments that measure specific aspects or qualities of knowledge and skills. The next two sections present techniques that can be used to concentrate on particular parts of the problem-solving process as distinguished in Fig. 1.

Measuring Problem-Solving Processes

Thinking Aloud

Thinking aloud (and the protocol analysis that follows it) is possibly the most widely used technique for measuring processes. The well-cited 1980 book by Ericsson and Simon marked a very strong revival of this technique after the abandoning of introspective techniques. Part of its popularity is due to the relative ease with which it can be applied. However, the analysis of the data makes the technique just as laborious as many other techniques that require more preparation time. A second advantage of this method is that it delivers qualitative data and can be applied to as few as one subject. An obvious disadvantage of thinking aloud is that it may interfere with the main task a subject has to perform and tasks may be so automated that they are not accessible for thinking aloud anymore. Also, thinking-aloud data are subject to potentially low reliability. Thinking aloud can also be used to get an impression of problem states. One can, for example, use protocol analysis for determining the students' cognitive structure of Newton's second law. In 1992, Boshuizen and Schmidt examined the encapsulation of domain terms used in protocols of experts and medical students when solving a medical diagnosis problem.

Discontinuous Thinking Aloud

In discontinuous thinking aloud, subjects are interrupted while performing a task at certain times or during specific events and asked to report on what they were doing just before being interrupted. The advantage of this method over full thinking aloud is twofold: There is no constant interference of speaking aloud with the main task and the interruptions also provide a prompt to uncover processes. As they reported in 1990, Ferguson-Hessler and de Jong used this method for charting learning processes of learners in a domain of physics. Instead of asking subjects to think aloud at certain points, they posed specific questions at several points in the problem-solving task (such as "What do you plan to do next?" and "Why would you do this?").

Stimulated Process Recall

A modification of thinking aloud is stimulated process recall. In stimulated process recall, subjects are confronted with actions that they performed while doing a task (and that were recorded in one way or another) and are asked to describe in retrospect what they did at that moment. The advantage over full thinking aloud is that the subject's main task (problem solving) is not interrupted. A clear disadvantage is that subjects may rationalize their acts and possibly will not report their factual thoughts. Stimulated recall is recommended over thinking aloud when it is expected that the main task will consume much of the subjects' cognitive resources.

Conversations

A more natural way of collecting thinking-aloud data is to record the conversation of subjects solving a task together. When the task is being done on a computer, it is common practice that subjects work together. Miyake, for example, used this method in 1986, to get a view of people's process of trying to understand the working of a sewing machine. Others have used the same technique to study mathematical problem solving.

Logfiles

When a task is performed with the use of a computer, subjects' procedures or strategies can be measured by the recording of logfiles. For example, in a 1983 study, Sweller *et al.* had subjects solve mathematics problems on a computer and inferred the strategies used from the sequence of steps employed by the subjects. The ACT-based tutors (see for example Anderson *et al.* 1992), discussed by Anderson *et al.* in 1992, are among the best known computer programs (intelligent tutoring systems) that teach problem solving and make extensive use of recording students' actions for cognitive diagnosis. A precondition for an adequate analysis here is that the domain is well structured and has a "procedural" character, as is

the case with the ACT-based tutors (that teach domains such as LISP programming and geometry proofs), because only then can a more or less exhaustive domain description be made to match the subjects' actions.

Eye Movements

Recordings of eye movements of subjects who solve problems can be used to gain information on specific parts of the problem-solving process (e.g., the selective perception). In a study in which expert and novice dentists were presented X-rays of caries problems, it was found that experts pay more attention to what they call cognitively and visually conspicuous areas.

Measuring the Knowledge Base

Reproduction

Reproduction is the main objective of most traditional tests, in which subjects are requested to reproduce information that they have previously seen and rehearsed. This type of test is very well suited for measuring knowledge of a superficial level. It normally concentrates on conceptual and procedural knowledge.

Card Sorting

We use the term card sorting as a general term for the technique in which subjects have to arrange, in some way, pieces of information ("cards") they are offered. This technique is mainly used for getting an idea of the organization of knowledge. Originally, this technique was restricted to finding the relations between concepts. One can let subjects indicate the similarity of pairs of concepts by drawing lines between them and ordering the lines, have subjects quantify the relations between concepts, or have subjects sort concepts that are printed on cards. As a modern variant of the card-sorting technique, computerized concept mapping techniques have been introduced. Here, elements of knowledge (mostly concepts) are placed on the computer screen and subjects can arrange them spatially to indicate a relation. Some of these programs include extensive algorithms for analyzing the resulting configurations.

Although most frequently used for arranging concepts, in the context of problem solving, cards have also been used with other information, such as subjects' problem situations. This approach was used by Chi *et al.* in 1981. In this well-known study, students had to sort descriptions of physics problems. Chi *et al.* found that experts sorted the problems according to domain-related features, whereas novices used surface characteristics as a basis for their sorting. From this, the authors inferred that novices and experts differ in their level of knowledge, with the experts having a deeper knowledge. In other studies, subjects had to sort problems (area-of-rectangle problems) that contained missing, sufficient, or irrelevant information into one of these three categories. In 1986, De Jong and Ferguson-Hessler offered subjects cards that contained situational, conceptual, or procedural information from the physics topic "electricity and magnetism." Subjects were instructed to sort the cards so that coherent piles would result. Comparing the sorting from the subjects to an ideal sorting according to problem schemata, it was found that good novices had an organization that was much closer to the schema-based organization than did poor novices. When the same task was performed to experts, a different type of organization emerged.

Concept Association

Word or concept association is in fact the same technique as card sorting, with the exception that subjects are offered one stimulus (frequently a concept) and they are free to name all the other concepts they can think of. This technique has been used in many studies over many years. As with the card-sorting technique, only concepts were originally used as the starting stimulus. Other types of stimuli have now also been used, such as complete problem descriptions. In their 1981 study, Chi *et al.* offered their subjects labels of categories of problems and gave their subjects 3 minutes to tell everything they could about problems involving each category label and how these might be solved. From these data, Chi *et al.* made inferences about the knowledge organization of their subjects.

Explanations

A rather open way to measure subjects' knowledge is to ask them to give a free explanation of phenomena. In a 1991 study, Andre and Ding, for example, presented subjects with diagrams of electrical circuits and asked them to tell whether the system would work and explain why or why not. Mestre *et al.*, in 1993, presented subjects with a situation in physics, the experimenter introduced a change, and the subjects had to predict the result. Then, subjects had to write down a free explanation of the phenomenon. Mestre *et al.* developed a method of analysis to measure the structured use of a key concept—in their case, the work–energy concept. In the field of medical problem solving, in a 1993 study, Schmidt and Boshuizen had subjects write down so-called "pathophysiological explanation protocols" after they had given reconstruction of clinical cases.

Reconstruction

Reconstruction techniques are frequently used for measuring or indicating the presence of schemata in the

knowledge base. Subjects are offered information (e.g., a story or a problem description) that they have to reconstruct later. The basic idea behind this technique is that what is remembered reflects the knowledge of the learner. If this knowledge is expected to comprise schemata, one might expect that information that is part of a schema is remembered better than information that is not incorporated in the schema. This latter information, for example, concerns information on details. In this technique, the experimenter takes care that the information that is offered cannot be learned by heart. To prevent learners from doing so, one can present a sequence of a number of problems and subsequently have subjects give reconstructions, or one can make the time for reading the information too short to allow for learning by heart. This technique was introduced by Jongman in 1967 in his study of memorizing chessboard positions. For a very short period of time, subjects were shown configurations of pieces on a chessboard, which they had to reproduce afterwards. In theory, the very short "exposure time" made it impossible to learn the positions by heart. The same technique in the context of problem solving was used on electronic circuits, algebra word problems, and complex devices. In 1991, de Jong and Ferguson-Hessler introduced a modification of this technique by having subjects reconstruct problem statements from physics (after a short exposure to these statements) and asking subjects in a number of cases to reconstruct a problem statement in a different modality (words or figures) from the one in which it was offered. They found that good problem solvers gave better reconstructions than poor problem solvers when they had to change modality from reading to reconstruction, but that the poor students outperformed the good students when they could stay within the same modality. They inferred that good students have a deeper understanding of problem situations. In 1992, Boshuizen and Schmidt used a reconstruction technique to assess the characteristics of conceptual knowledge of medical students. Clinical cases were presented under a controlled period of time, and subjects had to reconstruct the case later. Boshuizen and Schmidt analyzed the protocols on the type of concepts used and concluded that experts use higher level concepts than novices and intermediates. The authors varied the time of studying and found that expert's reconstructions were largely unaffected by this manipulation but that novices and intermediates produced better reconstructions when studying time was longer. They explain that this result is due to the fact that experts use so-called encapsulated knowledge while processing and reconstructing the case. Finally, in a 2002 study, Savelsbergh *et al.* let subjects construct problems on the basis of formula that could be used in a problem-solving process. In comparing novices and experts, they found that competence is related to the structure of knowledge of problem situations rather than the use of particular concepts.

See Also the Following Articles

Heuristics • Knowledge Work

References

Anderson, J. R., Corbett, A. T., Fincham, J. M., Hoffman, D., and Pelletier, R. (1992). General principles for an intelligent tutoring architecture. In *Cognitive Approaches to Automated Instruction* (J. W. Regian and V. J. Shute, eds.), pp. 81–107. Erlbaum, Hillsdale, NJ.

Andre, T., and Ding, P. (1991). Students misconceptions, declarative knowledge, stimulus conditions, and problem solving in basic electricity. *Contemporary Educational Psychol.* **16,** 303–313.

Boshuizen, H. P. A., and Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates, and novices. *Cognitive Sci.* **16,** 153–184.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Sci.* **5,** 121–152.

de Jong, T., and Ferguson-Hessler, M. G. M. (1986). Cognitive structures of good and poor novice problem solvers in physics. *J. Educational Psychol.* **78,** 279–288.

de Jong, T., and Ferguson-Hessler, M. G. M. (1991). Knowledge of problem situations in physics: A comparison of good and poor performers. *Learning Instruction* **1,** 289–302.

de Jong, T., and Ferguson-Hessler, M. G. M. (1996). Types and qualities of knowledge. *Educational Psychologist* **31,** 105–113.

Ericsson, K. A., and Simon, H. A. (1980). Verbal reports as data. *Psychol. Rev.* **87,** 215–251.

Ferguson-Hessler, M. G. M., and de Jong, T. (1990). Studying physics text; differences in study processes between good and poor performers. *Cognition Instruction* **7,** 41–54.

Goel, V., and Pirolli, P. (1992). The structure of design problem spaces. *Cognitive Sci.* **16,** 395–429.

Mayer R. E. (1991). *Thinking, Problem Solving Cognition.* Freeman and Company, New York.

Mestre, J. P., Dufresne, R. J., Gerace, W. J., and Hardiman, P. T. (1993). Promoting skilled problem-solving behavior among beginning physics students. *J. Res. Sci. Teaching* **30,** 303–317.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Sci.* **10,** 151–177.

Perfetto, G. A., Bransford, J. D., and Franks, J. J. (1983). Constraints on access in a problem solving context. *Memory Cognition* **11,** 24–31.

Robertson, S. I. (2001). *Problem Solving.* Taylor & Francis, Philadelphia.

Savelsbergh, E., de Jong, T., and Ferguson-Hessler, M. G. M. (2002). Situational knowledge in physics: The case of electrodynamics. *J. Res. Sci. Teaching* **39,** 928–952.

Schmidt, H. G., and Boshuizen, H. P. A. (1993). On the origin of intermediate effects in clinical case recall. *Memory Cognition* **21,** 338–351.

Simon, H. A., and Hayes, J. R. (1976). The understanding process: Problem isomorphs. *Cognitive Psychol.* **8,** 165–190.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher* **18,** 8–15.

Sweller, J., Mawer, R. F., and Ward, M. R. (1983). Development of expertise in mathematical problem solving. *J. Exp. Psychol. Gen.* **112,** 639–661.

van Bruggen, J. M., Boshuizen, H. P. A., and Kirschner, P. A. (2003). A cognitive framework for cooperative problem solving with argument visualization. In *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making* (P. A. Kirschner, S. J. Buckingham Shum, and C. S. Carr, eds.), pp. 25–47. Springer, London.

Process Evaluation

Sandra Vergari
University at Albany, State University of New York, Albany, New York, USA

Glossary

applied research An investigation aimed at producing findings that can be used to address a real-world problem.
evaluation The use of social science research methods to assess the effectiveness of a policy or program.
formative evaluation An assessment conducted during program implementation, aimed at providing feedback for continuous improvement.
policy An official course of conduct linked to the achievement of one or more social objectives.
program A structured social intervention or treatment in which specific resources and activities are aimed at producing one or more outcomes.
program implementation failure The outcome when program activities assumed to be necessary for program effectiveness are not performed as planned.
program monitoring The ongoing assessment of program dynamics during implementation, to determine whether actual practice reflects the assumptions and objectives of the planning phase.
program process theory A set of interrelated statements specifying the assumptions and expectations about how a program will operate.
program theory failure The outcome when a program is implemented as planned, but does not produce the expected results.
triangulation The use of multiple methods or sources to study a phenomenon, thereby increasing confidence in the validity of the findings.

Process evaluation is aimed at discovering whether a policy or program is operating as originally planned. Process evaluators examine what is actually occurring during the implementation and delivery of a program. Process evaluation assesses the extent to which program planning assumptions were accurate and program goals are being met. Whereas impact evaluation focuses on identifying effects produced by a program, process evaluation helps to explain why such results occurred.

Evolution of Process Evaluation

History and Experts

Evaluation has historical roots in the 17th century, but evaluation as currently practiced is a relatively recent phenomenon. The emergence of evaluation as a specialty field coincided with the growth of the social sciences and with increased support for research on social policies. Health and education programs were the frequent subjects of early evaluation efforts in the 20th century. Although outcomes (or "impacts") have traditionally been the focus of many evaluation efforts, experts note that process evaluation is also critically important. In contrast to early impact evaluations, which tended to assume that a program was operating as intended or as administrators claimed, process evaluation examines practice and data to determine whether the program is indeed operating as intended or claimed.

Pioneers and experts in the evaluation field include Ralph W. Tyler in the early 20th century, and Donald T. Campbell, Lee J. Cronbach, Howard E. Freeman, Peter H. Rossi, Edward A. Suchman, and Carol H. Weiss in the second half of the 20th century. By the 1960s, evaluation research was common practice and evaluation emerged as a specialty field in the 1970s. Today, there are academic journals devoted to evaluation, such as *American Journal of Evaluation* and *Evaluation Review*, and professional organizations for evaluators, such as the American Evaluation Association. Advances in social science research methods, especially developments in survey research and in applications of computer

technologies, have contributed to the rise of evaluation as a specialty field. Whereas the nascent evaluation field was shaped largely by social science researchers, contemporary evaluation has been influenced by policymakers, program administrators, and other stakeholders such as political interest groups.

Evaluation Training

Evaluators typically receive training in the social sciences or in professional schools that provide instruction in applied social research. Beyond evaluation research methodology, evaluators need to be knowledgeable about the specific program areas, such as education, crime, or welfare, that they investigate. Evaluation projects that are technically complex, longitudinal, and costly are often completed by highly skilled evaluation staffs based in evaluation centers or research firms.

Evaluation Standards

Evaluation experts have not developed a concrete evaluation theory that prescribes precise evaluation activities for particular contexts. The disciplinary diversity of the evaluation field, disagreement among experts regarding optimal methodologies, and the scope and fluidity of the settings in which evaluation occurs have precluded the formation of such a theory. However, the field has developed evaluation standards and methodological guidelines aimed at promoting ethical, feasible, and useful evaluation. One well-known set of 30 evaluation standards has been produced by The Joint Committee on Standards for Educational Evaluation. The 1981 version of the standards was revised in 1994 to include applications to sites in addition to school settings. The American Evaluation Association has also approved its own set of five "Guiding Principles for Evaluators."

Both the broader context and site of a process evaluation may be fluid and unpredictable. As a result, evaluators operate according to a carefully designed research plan and are also flexible in tailoring their efforts to particular dynamics that arise while the evaluation is underway. Accordingly, process evaluation is frequently characterized as both an art and a science.

Process Evaluation and Its Purposes

What Is Process Evaluation?

Process evaluation involves the use of social science research methods to make judgments about the operation of a policy or program. As a form of applied research, it is aimed at helping to address real-world problems or questions. Process evaluation compares a program as implemented to the program as planned. Observations are focused on what is done during the implementation or delivery of a program and how it is done. Process evaluation identifies those involved in program implementation, those affected by the program, and problems encountered by the program's administrators and clients. Process evaluation assesses the extent to which program planning assumptions were sound and program goals are being met. One of the key purposes of process evaluation is to identify changes that need to be made in order for a program to operate as originally planned, be more effective, and/or be more efficient.

Process evaluation may take place over a period of months or years and can be narrow or broad in scale. A process evaluation might assess the operation of a particular after-school program in a single school district or state, or it might assess the implementation of a multifaceted preventative health program across a range of cities.

Purposes of Process Evaluation

It is important to note that process evaluation does not assess whether a program yields the planned outcomes for its clients. When a program model has been found to be effective elsewhere and process evaluation of a given program indicates that it is well implemented, it is still possible that certain features of the particular setting have prevented the program model from producing the desired outcomes in that setting. Conversely, impact evaluation assesses outcomes, but does not scrutinize whether features and dynamics of the program delivery process are operating as planned. Whereas impact evaluation focuses on identifying effects produced by a program, process evaluation helps to explain precisely why such results occurred. The findings of process evaluation complement those of impact evaluation.

In cases in which impact evaluation discovers the absence of predicted effects, process evaluation can determine the type of program implementation failure that has occurred. Types of implementation failure include the absence of a particular program activity or insufficient activity, use of the wrong procedure or activity, and nonstandardized program delivery that varies significantly across the target population. Process evaluation is especially helpful in cases in which the program implementation or delivery process varies across several settings. This is because process evaluation can identify differences across the settings; these can be matched up with the findings of impact evaluation in order to draw conclusions about the most effective way to deliver a given program. When a process evaluation finds that a program was implemented as planned, yet the impact evaluation indicates

Table I Questions Addressed by Process Evaluation

Primary question	*Detailed assessment*
How well is the program plan being implemented?	Are administrators implementing the program according to the program plan?
	Is the target population being served and responding as planned?
	How do program personnel perceive the operation of the program?
	How do program clients perceive the operation of the program?
	How do expert observers perceive the operation of the program?
Are the theoretical assumptions underlying the program plan accurate?	Is the program operating as indicated in the program process theory?
	Is it likely that program outcomes are connected directly to particular program components and processes, or are the outcomes likely to be due to factors separate from the program?
	Which circumstances and practices appear to promote effective delivery of the program?
How can the program be improved?	What are the barriers to efficient and effective delivery of the program?
	What changes would make the program delivery process more efficient and effective?
If the program in its current form is implemented at a different site, are the desired processes and results likely to occur?	

that anticipated outcomes were not produced, the evaluators have discovered a program theory failure.

As indicated in Table I, process evaluation is aimed at addressing a number of important questions about program delivery. The primary purposes of process evaluation can be summarized as follows. First, process evaluation is a type of formative evaluation aimed at improving the design of a new program and at facilitating continuous improvement. Second, process evaluation may be focused on monitoring the operations of established programs with the aim of gauging fidelity to the original plan, effectiveness, and efficiency. Third, process evaluation may be used to assess the validity of the theoretical assumptions at the heart of the original program plan. Fourth, process evaluation may be used to demonstrate accountability to program funders and policymakers. Finally, process evaluation is undertaken in order to discover the reasons for the results observed in an impact evaluation and to enhance overall confidence in the findings of an impact evaluation.

Methods of Process Evaluation

Planning the Process Evaluation

A sound process evaluation design identifies the key program components to be assessed during the evaluation. Such components may include program strategies, target population(s), activities and opinions of program administrators and clients, and technology and media used for program delivery and recordkeeping.

In planning the process evaluation, the evaluator figures out answers to a number of key questions:

- What types of information will be included in the study?
- How often will data be collected?
- Who will collect the information?
- How will the evaluator gain access to program administrators and clients?

The process evaluator can try to prevent barriers to researcher access during the study by clarifying the research relationships between the evaluator, program personnel, and target population prior to beginning the study. In addition, the evaluator must plan data collection measures that are feasible for use within the time frame and budget for the project. Prior to beginning the process evaluation, it is wise for the evaluator and sponsor to formulate an agreement on the type of evaluation and type of report that are reasonable to expect given the resources and time frame allotted. Evaluators also take care to consider the purposes of the process evaluation as viewed by the sponsor and other stakeholders.

Many experts suggest that data collection for process evaluation is best guided by a carefully developed causal model that specifies the key components of the program, how it is intended to function, and the relationships necessary for program effectiveness. A program process theory specifies the assumptions and expectations about how a given program will operate, including the organizational plan and the types of relationships between program personnel and the target population. Prior to undertaking the process evaluation, the evaluator may develop hypotheses

about causal relationships between key program components, intervening variables, and program outcomes. The evaluator might construct diagrams that portray the functions that the program is supposed to fulfill and the specific performance goals that have been established for the program. The process evaluator then uses performance measures to gauge the extent to which the program is performing according to the planned functions and performance goals.

Data Collection Methods

Process evaluation usually employs primarily qualitative research methods. It includes observation of both informal and formal activities and patterns during program implementation. Process evaluation data may be derived from a range of sources. Program personnel and members of the target population are key sources of information for process evaluation. Structured, formal interviews with program administrators and clients can provide valuable information about program processes. Open-ended interviews and informal conversations can also provide critically important information for the process evaluation. It is important to interview both program personnel and members of the target population to obtain their respective perspectives on how the program is actually operating. Moreover, external actors with expertise on a given program, such as journalists, community members, and university professors, might be interviewed. Interviews may be face-to-face or over the telephone. Process evaluators might also administer questionnaires to program personnel and clients. For example, the process evaluator might secure the cooperation of the program administrators to ensure that each time program personnel are in contact with a client about program delivery, the client is asked to complete a questionnaire for the process evaluation.

Other data sources for process evaluation include program records (e.g., the volume of program pamphlets distributed, the venues in which they were distributed, the number of queries received in response, and the average time frame for responding to such queries), meeting schedules, meeting minutes, e-mail communications, and formal and informal direct observations of program dynamics.

Skilled process evaluators take care to engage in triangulation, whereby research findings are based on multiple perspectives. Process evaluators seek information and perspectives from knowledgeable sources located inside and outside of the program.

Process Evaluation Challenges

Process evaluators are confronted with various technical and political challenges during the research design, implementation, and dissemination phases. First, it takes time to secure approvals for evaluation procedures and access to program administrators and the target population. Program personnel may be reluctant to participate in the process evaluation and to provide information for the study. They may resent having their activities scrutinized, and may be concerned about the process evaluation findings and implications for the continuity of their program. In its most extreme form, lack of cooperation from program administrators and clients can preclude the completion of a process evaluation. Thus, prior to beginning the process evaluation, it is useful for the evaluator to meet with program administrators and personnel in order to learn about any concerns on their part, to try to ease those concerns, and to promote cooperation between the evaluator and the research subjects.

A second concern is that program circumstances and activities may change during the course of a process evaluation, requiring potentially costly changes in the evaluation research design. It may be difficult, for example, to track program dropouts, and the rate of client attrition in a program may compromise the validity of the study. Similarly, expectations for the process evaluation may not materialize. For instance, the target population may be difficult to locate and identify and may be uncooperative. In addition, the sample of program clients to which the evaluator has access may be too small to ensure confidence in the process evaluation findings.

Third, the evaluator who plans to use program records as a data source may find biased, inaccurate, and sloppy data and recordkeeping processes. For example, a school might intentionally record some forms of student misbehavior as minor disciplinary incidents rather than as serious incidences of harassment. Operators of a community renewal program may neglect to track the precise number of community members who show up for scheduled meetings. Moreover, data collected for purposes other than evaluation may not be tailored to components of the process evaluation.

Another concern is that the diversity of process evaluation perspectives and approaches may leave the evaluator with little firm guidance on the best way to proceed. Evaluators may also face pressure to complete a process evaluation within a short time frame. The process evaluator must decide on the appropriate balance between a high-quality scientific study and one that meets the immediate pragmatic demands of the evaluation sponsor. In addition, when a process evaluation produces findings that differ from what was anticipated or preferred by stakeholders, they may publicly question the credibility of the evaluator. Finally, commissioned process evaluations are not always publicized or published by the sponsors. Evaluators who have worked diligently to produce a high-quality study may be disappointed

when their work is not disseminated to a broader audience.

Process Evaluation Benefits

Process evaluation provides valuable information to program implementers. It can make them aware of problems in the implementation process and provide details that can be used to formulate and apply corrective measures. Process evaluation provides feedback on the quality of ongoing program delivery and such information can provoke efforts to make delivery consistent with the program plan. Process evaluation provides useful feedback during the developmental stages of a program as well as for more established programs by identifying program features that are operating well and those in need of improvement. Such information facilitates continuous improvement and can be used to promote the diffusion of effective programs.

Process evaluation can identify the aspects of a program that appear most responsible for the outcomes. In other words, process evaluation can illuminate precisely how and why a desired result has transpired, thus enabling administrators to focus on the most productive program activities. Process evaluation may also offer useful critical analysis of the theoretical assumptions that are the foundation of a given program. When impact evaluation discovers the absence of impact, process evaluation can reveal whether this outcome is due to implementation problems or due to errors in the theoretical assumptions underlying the program. Similarly, process evaluation can indicate whether the observed impact is actually due to key features of the program or instead the result of other factors.

Some experts recommend that process and impact evaluations be conducted as part of a coherent single research project. In this way, the findings of the outcome evaluation are understood more accurately. The findings provided by impact evaluation are more useful and complete when accompanied by the information about program dynamics provided by process evaluation.

See Also the Following Articles

Basic vs. Applied Social Science Research • Data Collection, Primary vs. Secondary

Further Reading

American Evaluation Association. (2004). *Guiding Principles for Evaluators.* Available on the Internet at http://www.eval.org

Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards,* 2nd Ed. Sage, Thousand Oaks, California.

Judd, C. M. (1987). Combining process and outcome evaluation. In *Multiple Methods in Program Evaluation* (M. Mark and R. L. Shotland, eds.), Chap. 2, pp. 23–42. Jossey-Bass, San Francisco.

King, J., Morris, L. L., and Fitz-Gibbon, C. T. (1987). *How to Assess Program Implementation.* Sage, Newbury Park, California.

Patton, M. Q. (1997). *Utilization-Focused Evaluation,* 3rd Ed. Sage, Thousand Oaks, California.

Robson, C. (2000). *Small-Scale Evaluation.* Sage, Thousand Oaks, California.

Rossi, P. H., Freeman, H. E., and Lipsey, M. W. (1999). *Evaluation: A Systematic Approach,* 6th Ed. Sage, Thousand Oaks, California.

Scheirer, M. A. (1994). Designing and using process evaluation. In *Handbook of Practical Program Evaluation* (J. S. Wholey, H. P. Hatry, and K. E. Newcomer, eds.), Chap. 3, pp. 40–68. Jossey-Bass, San Francisco.

Tyler, R. W. (1991). General statement on program evaluation. *In Evaluation and Education: At Quarter Century* (M. W. McLaughlin and D. C. Phillips, eds.), pp. 3–17. 90th Yearbook, National Society for the Study of Education, part 2 (NSSE), and University of Chicago Press, Chicago. (Original work published in *J. Edu. Resourc.* 1942.)

Weiss, C. H. (1998). *Evaluation.* 2nd Ed. Prentice Hall, Upper Saddle River, New Jersey.

Psychological Testing, Overview

Miriam W. Schustack
California State University, San Marcos, California, USA

Howard S. Friedman
University of California, Riverside, California, USA

Glossary

assessment Psychological evaluation of an individual, typically based on data from testing.
forensic psychology The application of psychology to questions and issues relating to law and the legal system, including criminal justice.
measurement The psychological assessment of human subjects, and the processes and tools used.
psychometric Relevant to the quantitative measurement of some psychological characteristic.
reliability The extent to which an instrument provides consistent measurement.
validity The extent to which an instrument measures what it purports or intends to measure.

The concept of psychological testing is very broad, in terms of the goals, content, and methodologies involved. Psychological testing is important in educational settings, in vocational settings, in legal settings, in clinical settings, and in the laboratory. In each of these domains, there are multiple methods of measurement that may be appropriately employed, each of which has benefits and limitations. All psychological testing involves some potential risk, and ethical considerations are an important part of the testing situation.

The History of Psychological Testing

The history of modern psychological testing can be traced back to the late 19th century, when interest in the measurement of intelligence coincided with the development of laboratory techniques to study human sensory and motor abilities. There were earlier antecedents, though, including the Chinese system of examinations for civil servants that began over 4000 years ago and was in active use until the early 20th century. The emperors required a grueling succession of tests over several days and nights; passing the examination was required for initial appointment as a government official, and officials were also required to complete the exam successfully every three years to remain in service. Centuries later in medieval Europe, universities began to require students to undergo oral examinations to demonstrate their knowledge. Written exams became part of the process of earning a university degree several hundred years later (when paper became more commonly available in Europe). These early examples can be seen as the backdrop to the development of modern testing, but it was not until the late 1800s that the beginnings of modern approaches emerged.

James McKeen Cattell's classic 1890 paper on "Mental Tests and Measurements" is often viewed as a foundational document for the field—it introduced the term "mental test" into general usage. In the decades immediately preceding, scientists in both Europe and the United States were involved in work that contributed to the development of psychological testing. Gustav Fechner published his *Elements of Psychophysics* in 1860, bringing a mathematical perspective from measurement in the physical sciences to the measurement of human psychological functioning. Edwin G. Boring, in his landmark 1929 book, *A History of Experimental Psychology*, credits Fechner with developing "the first

methods of mental measurement." The science of psychometrics was emerging.

Sir Francis Galton, in England, was interested in many aspects of individual differences in human abilities. Galton, inspired by the evolutionary theory of his cousin Charles Darwin, set out to measure all sorts of individual differences that might be relevant to survival. On one hand, from a modern perspective, Galton's work on "hereditary genius" appears seriously flawed by his racist preconceptions and by his lack of attention to the importance of environmental variables in influencing abilities and achievement. On the other hand, Galton pioneered the use of unobtrusive measurement, helped advance the development of techniques involving objective and repeatable measurement, and invented ingenious and reliable instrumentation for measuring human abilities.

During this same period near the end of the 19th century, Alfred Binet was working in France on the problem of predicting children's scholastic achievement. He was attempting to develop an instrument to test children's intelligence, with the goal of identifying those children who were too limited in their mental abilities to benefit from instruction in a normal classroom. He also searched for bright children who were being overlooked. With the assistance of Theodore Simon, Binet developed a test that focused on higher mental functions of judgment and reasoning. Binet and Simon published the first version of their test in 1905. The revised version of that test published just a few years later (in 1908) introduced the notion of mental level. This idea was further developed by Lewis Terman of Stanford University. Terman's revision of the Binet–Simon instrument was termed the Stanford–Binet test; it is still in use today. In their revision of the test, Terman and his colleagues significantly expanded the range of individuals for whom the test could be used, in terms of both age and ability level, but the more important modification was the introduction of the concept of the "Intelligence Quotient" or IQ, which allows chronological age to be integrated with mental ability into a single index with a mean of 100.

The next major historical influence on the development of psychological testing came with the advent of World War I. Close to four million men (draftees and volunteers) were recruited into the armed forces of the United States within a very brief period, and there was a pressing need to determine how the assignment of these recruits to different types of positions within the military could be done quickly and appropriately. Robert M. Yerkes of Harvard developed two tests, called Army Alpha and Army Beta, suitable to be administered to large groups—a departure from the prior practice of labor-intensive individual tests administered and scored by expert examiners. It is not clear that these tests were optimally designed or appropriately used (especially in the case of servicemen who were immigrants or from racial minority groups), but the enormous amount of testing that was involved seems to have set the stage for the large-scale educational testing that developed in the years that followed.

For almost 20 years before the massive testing of military recruits began, the College Entrance Examination Board (CEEB) had been in existence as a centralized college-admissions testing service, but the years just following World War I saw major changes in how the CEEB developed and administered its tests. The current Educational Testing Service (ETS), which is the successor to that organization, plays a central role in the process of admission to college and to graduate and professional school. The SAT has become a universal cultural experience among high-school students who are planning to attend college in the United States. For people who have never studied psychology, the SAT may be the most familiar form of psychological test.

For the field of psychology, however, the domain of personality testing is at least as important as that of educational or intelligence testing. Over the course of the 20th century, personality testing was a very active area of research, and the development of new tests and new methods continues to be an active and important field. Personality testing, like intelligence testing, can be traced back to Galton, but the testing domains diverged substantially at the start of the 20th century. Carl Jung developed a word-association method of exploring what he termed "complexes" (groupings of emotionally charged feelings, thoughts, and ideas). Jung's technique differed in two important ways from the earlier Freudian technique of "free association" in which a patient would verbalize thoughts and images as they reached awareness: first, Jung's "association method" was systematic and controlled, including a fixed set of word prompts to which the patient would respond, and second, it was used as an approach to assessing the patient rather than as a form of therapy in itself. At around the same time that Jung was developing the word association method, another Swiss psychiatrist slightly younger than Jung was working on a related approach. While Jung was refining a test where patients revealed their inner conflicts by their responses to a carefully selected set of words, Hermann Rorschach was developing an analogous technique that provided pictures as the stimulus instead. Rorschach created and refined a set of images formed from inkblots, and developed a methodology for interpreting a person's response in identifying the images. While there are serious reservations about the reliability and validity of Rorschach's test, it provided a foundation for the development of more modern forms of projective tests. Projective tests endeavor to assess thoughts and feelings that are too difficult to put into words or are outside of consciousness.

In more recent years, psychological testing has been transformed by the development of modern computer-based methods. New methods not only enable a more extensive analysis of the results yielded by a complex test instrument, but the development of computerized adaptive testing techniques and item response theory allow for the testing to be more responsive to the characteristics of the examinee, and thus a more accurate score or profile can be obtained with a shorter instrument.

Domains and Uses of Psychological Testing

The essential goal of virtually all forms of psychological testing is to measure or assess something about an individual person. Although many common applications of psychological testing involve administering a test to a large number of people, and although the test results from large numbers of people are often aggregated for reporting, psychological testing is nevertheless focused on individuals. There are many different domains in which psychological testing is commonly used, and a wide range of uses for the test results. Reference sources such as *Tests in Print* and the *Mental Measurements Yearbook* provide listings of thousands of tests of many different types, updated at regular intervals. One straightforward way to categorize these many domains and uses and types of tests is in terms of the different goals that the testing serves.

Intelligence Testing

One major domain of psychological testing is closely connected to its historical roots—the testing of intelligence. Intelligence testing is used in qualifying children for educational programs for those of atypically high intelligence, and is also used extensively in the classification of children in need of special education because of their atypically low intelligence. Intelligence testing is also part of the battery of tests that are used to diagnose the specific disorders of children who have some developmental abnormality or learning disability. Because of the possible effects of bias in the way IQ tests are constructed, administered, and interpreted, however, the practice of using IQ tests in the placement of low-IQ children in special education is in question. Federal court decisions (in 1979 and 1986) barred school districts in the state of California from using IQ test results to place African-American children in classes for the "educable mentally retarded" or equivalent categories, based on the tests being biased against this group and thus not providing valid results for these children.

Intelligence tests are much less commonly used in adults at present—among the most common current uses are as part of a general ability and skill battery, or to qualify for membership in the MENSA society. They are also used in court-ordered tests of mental capacity (see section below on forensic and clinical testing).

Educational Testing

Everyday classroom testing, such as a spelling test, a quiz on the state capitals, a geometry chapter examination, or a test on the names of the bones in the human body can be seen as outside the realm of psychological testing. There is no well-defined boundary, though, that separates such testing from the forms of educational testing that are more obviously within the realm of psychological testing, such as aptitude tests, or tests of learning styles.

The very prominent testing programs offered through the Educational Testing Service (including the SAT, PSAT/NMSQT, GRE, AP, MCAT, LSAT, CLEP, and many others) are central to current practices for admission to college and to graduate and professional schools, and for awarding educational credit and scholarships. The educational fates of students at all levels are heavily influenced by how well they score on these standardized tests, and there has been substantial controversy over the years about how these scores should be used. Scores on these standardized tests do have some correlation with college success, but the consensus of studies of the relationship is that the scores predict freshman year grades only to a modest degree, and predict overall college success (such as cumulative GPA and graduation) to an even smaller extent. The scores are statistically valid as predictors, but not quite as good as high school grades at predicting academic success in college.

In recent years, the use of standardized tests in educational settings has become even more prominent with the advent of what is often called "high-stakes" testing. The stakes are high because scores on a standardized test of academic achievement are used to make consequential decisions both about individual students, such as whether the student will be promoted to the next grade or granted a high-school diploma, and also about schools, which may lose some funding or even be shut down based on the scores received by students. This trend is worrisome to those in education who are sensitive to the limitations of this type of testing. The American Educational Research Association, concerned about the increasing prevalence of high-stakes testing, recently developed a position statement setting forth conditions that are necessary for sound and appropriate implementation of this kind of testing, intended both to guide and to caution those involved in mandating, implementing, or using these testing programs.

Personality Testing

Another very active area of psychological testing is in personality. Hundreds of personality tests are available,

arising from a variety of theoretical orientations, and using a broad range of methods. Some tests are focused on careful measurement of a single construct or variable, while others attempt to assess the personality as a whole, including many dimensions or variables. The methodologies of these tests run the gamut as well, including virtually all the methods discussed in the methodology section below. Among the most heavily used general personality inventories is the MMPI (Minnesota Multiphasic Personality Inventory), and the NEO-PI (based on the current five-factor or "Big 5" model of personality structure). Also common is the Myers–Briggs Type Indicator, which is based on the typological personality theory of Carl Jung. The goals in the use of personality tests are quite varied as well—people who are well adjusted may simply be interested in learning more about themselves, or those who are having psychological problems may want to use the testing to find areas of weakness or dissatisfaction. Those who have problems in a relationship may find that personality testing can improve their understanding of differences or incompatibilities with their partners.

Forensic Testing

Forensic psychology is the application of psychology to questions and issues relating to law and the legal system, including criminal justice. Psychological tests are used at many points in the investigation, prosecution, and punishment of a crime. One common application of psychological testing in the legal system is in the determination of whether an accused person is psychologically fit to stand trial. In the United States, the criminal justice system requires that a defendant be able to understand the charges that he or she is facing, and requires that the defendant be able to understand that the criminal behavior being charged was wrong. In cases where the accused person suffers from severe mental illness or extreme mental retardation, he or she may be unable to meet that standard; the determination would be made through psychological testing. A less common application of psychological testing is in determining the status of a defendant who pleads not guilty by reason of insanity. It is the task of psychological examiners (working either for the defense or for the government) to report on the status of the defendant's sanity. Although the public hears quite frequently about defendants claiming "the insanity defense," it is much more common for a defendant to be unable to stand trial than to plead insanity.

Another use of psychological testing within the legal system is for the determination of custody in contested cases. This may involve testing of one or both parents to determine fitness to have custody of a child or children, and may also involve testing the child to better understand the child's needs and the child's relationship with each parent.

In civil lawsuits, psychological testing may be necessary to determine the extent of harm from the actions of a person or corporation. For example, if a lawsuit involves compensation for brain damage, psychological testing may be involved in determining the extent of the harm.

Within the legal system, psychological testing is also called upon to determine the risks posed by those who have been convicted. Tests are used to help assess the risk of recidivism (re-offending), the suitability of a convict for parole, and the level of risk posed by a sex offender who has completed a sentence. However, such tests have limited validity; that is, they are far from perfect predictors.

Some forms of psychological testing that do not meet current standards continue to be used in the legal system. The polygraph or "lie detector" test is still commonly used in criminal investigations (although its use is restricted in formal court testimony), despite decades of data documenting its serious limitations. Polygraphy works by recording blood pressure, breathing (rate and depth of each breath), and the electrical conductance of the skin (which increases in the presence of sweat) as a person responds to a series of questions. Changes in these bodily functions (which are largely under the control of the autonomic nervous system) are interpreted as reflecting the increased anxiety associated with lying. Law enforcement also continues to use a device called a voice stress analyzer (VSA) as a way of determining if people under interrogation are telling the truth. This test relies on a machine that detects the occurrence of tiny tremors in the voice that are not audible to the listener, and can measure changes in the prevalence of these subaudible tremors. Under stress, these subaudible tremors decrease. The logic (and fallacy) of both these techniques is similar—they rely on a chain with several weak links. These tests are useful only to the extent that telling a lie necessarily causes stress, and only to the extent that this stress necessarily causes physiological changes that can be detected by the machines. Some people lie but are not detected (a false negative), while some people are "detected" as liars even when they are telling the truth (a false positive).

Clinical and Psychopathology Testing

In the clinical setting, psychological testing plays an important role in diagnosis of mental illness. In addition to allowing for the planning of proper treatment, psychological testing is also an important component in determining if a patient needs to be treated on an in-patient versus out-patient basis, or even if the patient may need to be involuntarily committed to a hospital to prevent further harm to the patient or to others. Many tests have been developed that can be used to diagnose specific forms of psychopathology, using the categories of the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM-IV).

Testing of Attitudes, Values, and Personal Orientations

In many situations, both in everyday life and in the psychology laboratory, it would be useful to have reliable and valid measures of a person's values, attitudes, and personal orientation. Tests of attitudes can use either direct or indirect methodologies. One direct method is based on simply asking a person to evaluate how much he or she agrees with a statement expressing an attitude. For example, an item about capital punishment on an attitude measure might take the form of asking a person how strongly they agree with a certain statement about capital punishment. An alternative method is to offer a pair of statements on the topic and require the test-taker to choose the statement that more closely matches his or her opinion. In many cases, test-takers may not be willing (or able) to disclose their true attitudes if those attitudes are controversial or socially unacceptable. For example, many people would not disclose an overtly racist attitude, even if they were aware that they held such an attitude. Indirect measures can sometimes provide an alternative mechanism for measuring attitudes.

Using a similar variety of techniques, a person's values (what the person thinks is important and worthwhile) can also be determined. Personal orientation (sometimes also called world-view) is amenable to similar techniques. Notable test approaches in this domain include the Personal Orientation Inventory (POI), which focuses on Maslow's construct of self-actualization. There are also multiple commonly used measures of gender roles, and masculinity/femininity such as Sandra Bem's scale, called the Bem Sex-Role Inventory. Sexual orientation can also be assessed indirectly by measuring physiological arousal to different stimuli.

Vocational/Industrial Testing

Psychological testing is used in many ways in employment settings. Students who are about to embark on their first career as well as people who are interested in career change can be helped by the use of tests that have been developed specifically for career guidance. These tests take many forms: they can help people to understand their own preferences for types of work environments and tasks, or they can help construct a profile of a person's skills and abilities. One such test that is given every year on a massive scale is the Armed Services Vocational Aptitude Battery (ASVAB), which is taken by all new military enlistees as well as by about one million high school students in the United States every year. As many such career guidance tests do, it combines an aptitude component with an interest inventory, with the goal of guiding the examinee into a career track that will be interesting as well as suitable for the person's skills.

Another vocational application of psychological testing is as an employment-screening tool. Many employers require job applicants to take psychological tests as a condition of employment, although they are often euphemistically referred to as "pre-employment profiles" rather than the psychological tests they really are. In addition to specific skills tests that have clear relevance to the position of interest (such as a typing test for a secretarial applicant, or a computer programming test for a programming applicant), employers sometimes require applicants to undergo personality assessments. While there have been many legal challenges to this practice over the years, it is still commonly used. Employers are now prohibited under federal law from requiring pre-employment polygraph testing, except for certain occupations.

Also in the vocational arena, testing is sometimes used as a condition of continued employment (not unlike the situation of the Chinese civil servants thousands of years ago), or of advancement. This is most common for government employees (such as security screeners or police officers), but also occurs within the private sector.

The vocational uses described above are all examples of the use of psychological testing to determine if the candidate is well-suited to the job. Psychological testing methods are also used to ensure that the demands of the job are suitable for the people who will be doing the job. Appropriate use of psychological testing can ensure that equipment and processes to be used in a job setting are suitable for the skills and preferences of the individuals who will be using them; "human error" can be minimized by tailoring the demands of the job to the characteristics of the person.

Methods of Psychological Testing

Objective vs Subjective Tests

One central distinction that differentiates among psychological tests is whether they are essentially objective or subjective measures. If a test requires judgment or interpretation to turn the responses into a score or result, then it is considered to be subjective. When an examinee provides a description of a Rorschach inkblot, for example, it takes the judgment of a trained scorer to translate that description into a characterization of the examinee. In the case of an objective test, the responses of the examinee feed directly into the score via a mathematical algorithm. For example, a person taking the MMPI fills out an answer sheet that is run through a computer, which then provides the scores. Objective measurement is most often associated with tests where respondents select from a predetermined set of choices (as in a multiple-choice

test, or rating on a 7-point scale), while subjective measurement is normally required when the respondents construct their own responses.

Projective Tests

Projective tests are commonly used in the measurement of personality. In a projective test, respondents must interpret or describe an ambiguous stimulus (as in a Rorschach inkblot or a Thematic Apperception Test photo), or come up with a drawing in response to a minimal prompt ("Draw a person"), or say a word in response to a stimulus word (as in Jung's Word Association Test). Because of the substantial interpretation required on the part of the scorer, this type of test tends to be fairly low in reliability.

Self-Report vs Report by Others

Another important dimension of psychological testing is whether the information about the subject of the test is coming from the subject him- or herself, or from another person. Clearly, a report from another person provides information about how the subject appears to an observer, and only indirectly provides information about the subject's own thoughts, feelings, and self-perceptions. In many cases, though, information provided by an external observer is more appropriate for the purposes of the test than a self-report would be. For example, the diagnosis of ADHD (Attention Deficit Hyperactivity Disorder) in children is often based on a test in the form of a behavior checklist given to a child's parents and/or teachers. Because the goal in this case is the diagnosis of a deviation from normal, age-appropriate behavior, external observers such as parents and teachers can be the most suitable informants. Another example of the value of data from knowledgeable others comes from the decades-long longitudinal study begun by Lewis Terman in the 1920s. Terman recorded personality characteristics of school-age children as rated by their parents and teachers during the 1921–1922 school year. These ratings turned out to be a reliable predictor of the children's longevity across the life-span.

Most psychological tests, though, use data provided by the subject him- or herself, except in the case of testing that is focused on the question of how self-image differs from external perceptions by others.

Interview Approaches

Some psychological tests rely on an interview technique. A trained interviewer conducts an interview with the subject of the test. Such interview approaches are often characterized as either structured interviews or open-ended interviews. While an open-ended interview is not completely free-form (as a social conversation might be), the structured interview requires that a specific set of questions be asked using fixed wording, and the responses are scored according to a predetermined rating system. For example, there is a structured interview instrument that was developed for assessing the Type A Behavior Pattern.

Physical Indicators

There are many forms of psychological tests in current use that rely on directly measuring some aspect of a person's neuro-psychological functioning, or some other biologically based characteristic. Table I includes descriptions of a variety of such tests. While we now discredit the technique of phrenology, an approach developed in the early 1800s in which a person's characteristics were established by examination of the shape of the skull and the location of its protuberances, over the past decades we have developed a host of new techniques that probe the mind by measurement of various physical characteristics. Some of these techniques have been in use for decades or more, such as the measurement of galvanic skin response (electrodermal conductivity) as an index of anxiety, or the measurement of heart rate as an index of arousal. These are the kind of measures that are used in polygraph analyses. Other techniques are on the cutting edge of current technology: for example, the use of functional magnetic resonance imagery (fMRI), where brain activity is recorded as a person performs some mental task.

Risks and Dangers of Psychological Testing

Given the power of psychological tests, there are many dangers inherent in their use, and great potential for abuse. There are potential problems that can arise from poorly constructed tests, from bias on the part of those involved with the testing, and other potential problems that arise from insufficient attention to ethical requirements.

Reliability and validity are critical to the usefulness of any test, and the absence of these attributes poses risk to any user of the test. A reliable test will yield consistent outcomes—it will provide its results with relatively small error of measurement. There are many statistical techniques that allow for the determination of a test's reliability, and techniques of test construction have been developed that can ensure good reliability. The use of an unreliable psychological test does pose risks, since the result obtained may not be correct, but it is not difficult to avoid unreliable tests given that reliability can be easily assessed. Validity is both more complex and more

Table I Biological/Neuroscience Measures of Individual Differences

Measure	*Description*
Skin conductance, heart rate, blood pressure	Reflect activity of the autonomic nervous system; often are too broad to be useful measures of personality.
PET (Positron Emission Tomography)	Uses radioactively tagged molecules to probe brain function, such as radioactive glucose to examine changes in energy metabolism which are associated with activity. Other compounds can also be tagged radioactively and used to examine other brain processes. But metabolic changes may be slow and delayed.
fMRI (functional Magnetic Resonance Imaging)	Uses very large magnetic fields to probe the movements of molecules. fMRI takes advantage of differences in the properties of oxygenated and deoxygenated hemoglobin, thus yielding a signal that is related to brain activity (since neural activity uses oxygen).
EEG (Electroencephalography)	Measures electrical potentials at the scalp that are caused by large populations of neurons becoming active simultaneously. The P300 wave occurs in response to novel stimuli and may be prove useful in differences in reactions to novelty.
MEG (MagnetoEncephalography)	Similar to EEG, but instead of recording electrical potentials, it records the magnetic fields that result from the electric currents in the brain.
Neurochemical assays	Chemical analyses for the presence of certain neurotransmitters, transmitter metabolites, or hormones. The sites of assay can be in the cerebrospinal fluid or in the blood. In animal studies, assays can be directly in the brain (by microdialysis).
Postmortem analysis	For examining individual differences in anatomy (both gross and cellular) and in the numbers and locations of neuroreceptors (after death).
Candidate gene studies	Search for specific genes that correlate with specific characteristics, although multiple genes likely contribute to any psychological trait. With the succcesful unravelling of the human genome, this biological approach is likely to gain prominence in the years ahead.

difficult to ensure, so its potential absence poses greater risk. Validity is commonly defined as the extent to which an instrument measures what it purports to measure. There are several different categories of validity, and multiple approaches to determining the degree of validity of a given test.

Bias, which interferes with validity, can arise from the actions of the test developer, the test taker, the test administrator, or the test interpreter. There are many such influences that are generally recognized, and well-designed tests are constructed in a manner that minimizes the extent to which the outcome is affected by extraneous factors.

Bias on the part of the test taker can be defined as the effect of any influences that make a response by the test-taker not fully accurate or honest. Some of these factors are intuitively obvious: for example, people tend to present themselves in a positive light, by favoring those responses that are more socially desirable. Another straightforward example is when the test-taker misinterprets the meaning of an item (due to language or cultural differences) and responds accordingly. To preserve validity, these biases should be anticipated by the developers and users of the test. There are more subtle forms of bias as well, that are less blatant and require appropriate modification. For example, there is a pervasive phenomenon of an acquiescence response set, where examinees are somewhat more likely to respond "yes" or "I agree" or "that describes me" than to respond "no" or "I disagree" or "that does not describe me." Careful wording of the items, and including equal numbers of positively coded and negatively coded ("reverse-coded") items, can normally neutralize this bias. A modern technique for test construction called Item Response Theory uses a mathematical analysis of item scores, focusing on the probability that a person with the trait or ability being assessed will give a particular answer to a given item.

Bias can be inadvertently introduced by the examiner, especially if the examiner and the test-taker differ on characteristics of ethnicity and gender. The examiner may communicate hostility or low expectations to the test taker. But even if the examiner behaves in a perfectly appropriate way in administering the test, the demographic or interpersonal characteristics of the examiner may influence the performance of the test-taker. These effects can be subtle and are not necessarily easily remedied or overcome.

A potentially subtle form of bias on the part of the interpreter can occur when instruments that have been developed and standardized on one population group are then applied to a different cultural, racial, gender, or

ethnic group. Clear examples of this type of bias occur when tests are focused on a characteristic where community norms differ between cultural groups. For example, if a test assesses an individual's level of extroversion by including a measure of how far the examinee chooses to stand from the examiner, the measure will be flawed if it fails to adjust for appropriate norms that are part of the individual's cultural background.

Any professional who administers a test, scores it, or receives its results incurs multiple ethical obligations. Both the American Psychological Association and the American Educational Research Association have formal ethical standards documents, providing guidelines that cover many aspects of the testing situation. These ethical standards concern issues such as the privacy of test scores, the requirement that test takers give informed consent before testing, the right of test-takers (and their legal representatives in some cases) to be provided with their results in language they can understand, the appropriate treatment of people who are undergoing testing, the restricted distribution of test instruments, the required training and credentials for those administering tests, and the special ethical considerations that arise with court-ordered testing.

In addition to the set of unique obligations that the testing situation imposes, broader ethical guidelines require that test data used in research be as prudently safeguarded and as carefully analyzed as any other forms of data. Psychological testing can provide a wealth of information about individuals, but like any other powerful tool, it needs to be used carefully and appropriately so that it is beneficial rather than destructive.

See Also the Following Articles

Classical Test Theory • Computer-Based Testing • Education, Tests and Measures in • Intelligence Testing • Measurement Theory • Psychometrics of Intelligence • Reliability Assessment • Validity, Data Sources

Further Reading

American Educational Research Association (1999). *Standards for Educational and Psychological Testing 1999.* American Educational Research Association, Washington, DC.

American Psychological Association (2002). *Ethical Principles of Psychologists and Code of Conduct* (in effect as of June 1, 2003). www.apa.org/ethics/code2002.html. American Psychological Association, Washington, DC.

Friedman, H. S., and Schustack, M. W. (2003). *Personality: Classic Theories and Modern Research,* 2nd Ed. Allyn & Bacon, Boston.

Hersen, M. (ed.) (2004). *Comprehensive Handbook of Psychological Assessment* (4 volumes). Jossey–Bass/Wiley, New York.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: Sinking shafts at a few critical points. *Ann. Rev. Psychol.* **51,** 405–444.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., and Moreland, K. L. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *Am. Psychologist* **56**(2), 128–165.

Murphy, L. L., Plake, B. S., Impara, J. C., and Spies, R. A. (eds.) (2002). *Tests in Print VI.* University of Nebraska Press, Lincoln, NE.

Plake, B. S., Impara, J. C., and Spies, R. A. (eds.) (2003). *The Fifteenth Mental Measurements Yearbook*. University of Nebraska Press, Lincoln, NE.

Psychometrics of Intelligence

Peter H. Schonemann
Purdue University, West Lafayette, Indiana, USA

Glossary

common factors Latent variables implied by Spearman's (and later Thurstone's) factor analysis model. On partialling out all common factors of the observed variables, only uncorrelated specific factors are left.
congruence coefficient Cosine.
correlation Measure of linear relationship between two variables varying between −1 and 1 (0, no linear relationship; 1, perfect linear relationship).
criterion A variable of practical interest that a test is intended to predict.
dominant eigenvector The eigenvector associated with the largest (dominant) eigenvalue of a symmetric matrix.
eigenvector A vector mapped into a scalar multiple of itself by a square matrix. The scalar is called (the associated) eigenvalue.
general ability (*g*) An unobserved variable implied by a mathematical model proposed by Spearman to account for positive correlations among intelligence tests. In contrast to a PC1, *g* is not a linear combination of the observed tests.
intelligence Technically undefined, the term refers to a broad spectrum of "cognitive" skills presumed relevant for educational and economic success.
IQ total test score on an IQ test, normed to have a mean of 100 and a standard deviation of 15 or 16.
IQ test A test presumed to measure intelligence.
item A subtest of a test, usually scored 1/0 (pass/fail, binary item). The total test score is the sum of the item scores.
latent Implied but not observed overtly.
linear combination (linear composite) Weighted average.
matrix A rectangular array of real numbers.
partial correlation A correlation that remains after the influence of one or more other variables has been statistically removed (partialled out).
PC1, first principal component A linear combination of the observed tests that has largest variance among all possible linear combinations (subject to the constraint that the defining weight vector be of unit length).
reliability A measure of stability of test scores under repeated application to the same subjects; usually expressed as a correlation.
scalar A real or complex number.
validity A measure of the extent to which a test measures what it is designed to measure; often expressed as a correlation.
vector A linear array of real numbers set out either as a row or as a column.

Psychometrics, which literally means "measurement of the soul," is a subdiscipline of psychology devoted to the development, evaluation, and application of mental tests. It is useful to distinguish between two branches of psychometrics, a theoretical and an applied branch. They do not interact as much as might be expected. Here, both branches will be tracked side by side since it is impossible to gauge the merits of a theory without knowing what benefits it produced in practice.

This article discusses developments in test theory and factor analysis, with emphasis on applications to intelligence. This topic originally spawned interest in psychometrics, spurred on its early growth, inspired most of its lasting achievements, and, in the end, also revealed its limitations. It also has had the most profound social impact.

Introduction

The history of psychometrics spans approximately a century, beginning in earnest in approximately 1904/1905, when Binet and Spearman laid the foundations for future developments. This era was followed by a period of consolidation and growing acceptance of IQ tests and college admission tests in the United States.

Under Thurstone in the 1940s, psychometrics became academically respectable but also progressively more dogmatic and mechanical. The field reached its creative apogee under Cronbach and Guttman in the 1950s. Then, it began to stagnate and eventually regressed back to its racist roots in the 1920s.

In this article, matrices are denoted by capital letters, and scalars are denoted by lowercase letters. Column vectors are denoted by boldface lowercase letters, and row vectors are denoted by boldface lowercase letters followed by a prime denoting transposition.

Auspicious Beginnings: Charles Spearman and Alfred Binet

Galton, Wundt, and Wissler

The applied branch of psychometrics goes back at least to the days of Wundt and Galton (1880s). Inspired by the example of the physical sciences, both scholars attempted to measure basic sensory and physiological aspects of human behavior first, such as reaction time, visual acuity, and the like, to lay the grounds for the investigation of more complex variables later.

Galton, together with some of his younger colleagues, especially Karl Pearson and Udny Yule, also contributed to the foundation of the theoretical branch by developing basic correlational techniques that soon were to play a central role in both branches of psychometrics.

At the turn of the 20th century, Wissler applied these new techniques to school grades and some of the mental tests available at the time. The results proved disappointing. The sensory and physiological measures correlated moderately with each other, as did school grades, as indicators of more complex forms of mental ability. However, the correlations between both sets of variables were virtually zero. At this critical juncture, and almost simultaneously, two events breathed new life into the seemingly moribund new science.

Alfred Binet

On the applied side, in 1905 Alfred Binet (with Simon) published a new mental test that violated all canons of the Galton/Wundt school. Instead of trying to synthesize measures of more complex behaviors from simpler, more easily measured sensory variables, Binet set out to measure a highly complex mental trait of undisputed practical importance—intelligence—directly. The test he devised comprised a large number of items designed to sample various aspects of the implied target variable. This variable soon acquired the technical-sounding acronym IQ. It was defined, eventually, in terms of the total item score (which skipping some of the intervening mental age arithmetic that never applied to adults anyway). Binet's test became the prototype of most IQ tests and also of the scholastic aptitude tests still in use today. The problems with this approach begin after one constructs a second IQ test that does not correlate perfectly with the first, since then it is no longer clear which one is the true measure of intelligence.

Charles Spearman's General Ability (*g*)

To cope with this problem, in 1904 Spearman developed an entirely new theory aimed at supplanting the widely used but murky term intelligence with a clear-cut operational definition. Starting with the observation that most measures thought to relate to intelligence, such as school grades, tended to correlate positively, he reasoned this means they all measure the same latent variable, namely intelligence. To account for the fact that the correlations, although generally positive, were not perfect, he appealed to the recently developed machinery of partial correlations: If all positive correlations in a test battery are due to a single underlying variable—which he called *g*, short for general ability, to steer clear of the term intelligence—then its statistical removal should produce a matrix of partial correlations that are all zero. If this should turn out to be true for all batteries of "intelligence tests" worthy of the name, then this latent variable *g* could serve as an unequivocal definition of intelligence, which thus would have been "objectively determined and measured" (the title of his 1904 paper). The globality clause is usually omitted in modern accounts of Spearman's work, although it is absolutely critical for the stringency of this argument. After testing his theory on 12 previously published small data sets, he found it consistently confirmed: Dodd noted that "it seemed to be the most striking quantitative fact in the history of psychology."

Spearman left no doubt about how he felt about the social relevance of his presumed discovery: "Citizens, instead of choosing their careers at almost blind hazard, will undertake just the professions suited to their capacities. One can even conceive the establishment of a minimum index to qualify for parliamentary vote, and above all, for the right to have offspring" (Hart and Spearman, 1912).

Early Criticisms

This initial phase of exuberance soon gave way to sobering reappraisals. In 1916, Thomson noticed that the same type of correlation matrix that Spearman's theory predicts on postulating one common factor is also implied by a diametrically opposed theory that postulates infinitely many common factors (Thomson's sampling theory). To make matters worse, in 1928 E. B. Wilson, a polymath whose stature Spearman instantly acknowledged, wryly observed in an otherwise favorable review of Spearman's

"Ability of Man" (1927) that Spearman's theory did not suffice to define *g* uniquely because it postulated more independent factors than observed tests. As a result, many widely different "intelligence scores" can be assigned to the same subject on the basis of his or her observed test scores. This "factor indeterminacy" issue further undermined Spearman's claim of having defined intelligence objectively as *g*. For a while, it became the focus of lively debate, which soon died down after Thurstone entered the stage.

Eventually, it also became apparent empirically that the early accolades bestowed on Spearman's presumed discovery had been premature. As it turned out, the number of common factors needed to account for the observed test correlations did not stop at one for larger batteries, as the theory required; rather, it increased with the number of subtests in the battery. Typically, approximately one-third as many common factors as tests are needed to reduce the partial correlations to zero.

The fact that his claim of having empirically defined and measured intelligence turned out to be untenable in no way diminishes Spearman's stature for having been first to recognize and address this fundamental challenge and for dealing with it in the spirit of an empirical science.

1910–1930: Consolidation: World War I, Louis Terman, and Carl Brigham

Classical True Score Theory

Spearman's closely knit theory contained a theory of test scores as a special case that, for a considerable period of time, provided the needed definitions for constructing and evaluating new tests. This so-called classical true score theory (CTT) results on applying Spearman's factor model to only two tests and assuming that both contain the underlying common factor—now termed "true score"—in equal amounts (perfectly parallel tests). Under these assumptions, it is not difficult to derive plausible definitions of test reliability and test validity in correlational terms. Since both concepts are derived from the same underlying model, they are closely related. For example, CTT permits to predict how test reliability increases as one increases the number of items comprising the test (Spearman–Brown prophecy formula) and how test validity varies with the reliability of a predictor test, the criterion measure, or both (correction for attenuation). Kuder and Richardson used CTT to derive one of the most popular reliability estimates still widely used today. Cronbach later renamed it coefficient alpha.

Louis Terman

On the applied side, Terman promoted Binet's version of intelligence that is tied to a particular test. Availing himself of the concepts and techniques of CTT, he adapted it to an English-speaking clientele. He shared with many others of his generation, but in marked contrast to Binet, the eugenic prejudices of the early English pioneers (Galton, Pearson, Spearman, and others). This thought collective took for granted that intelligence—whatever it might be—(i) exists and can be measured and (ii) is primarily genetically predetermined (the figure usually given, until very recently, was 80%). Therefore, Terman and his disciples strove to purge their "intelligence tests" as much as possible of environmentally induced impurities (such as educational background) that might mask the underlying, presumed immutable genetic contribution. Note that the first premise conflicts with the outcome of Spearman's efforts.

World War I Army Tests

World War I provided an opportunity to put psychometric theories into practice when the need arose to assess, on short notice, the military qualifications of millions of inductees with widely different social and educational backgrounds. Under the stewardship of a former American Psychological Association president, Robert Yerkes, a committee was charged with the development of intelligence tests suited to the particular needs of the U.S. Army. These efforts resulted in two different tests, the Army Alpha and Army Beta tests. By necessity, both were conceived as group tests (in contrast to the classical Binet test and Terman's American adaptation that had to be administered individually). The Army Alpha was essentially a group version of the Stanford–Binet that relied on the assumption that the testee was fluent in English. The Army Beta was intended as a "nonverbal" substitute for linguistically handicapped inductees whose innate intellectual potential might be masked by traditional verbal tests.

Charles Brigham

World War I offered an opportunity for both branches of psychometrics to promote the fruits of their labors. It also produced a huge database waiting to be mined after the war for new findings to further improve the new technology. This task fell to a young assistant professor, Charles Brigham, who published his findings in a book titled "A Study of American Intelligence," with a foreword by Yerkes.

Brigham's conclusions were in tune with the zeitgeist and seemed perfectly timed to answer steadily mounting concerns over untoward consequences of unrestricted

immigration. He found that the average intelligence of American soldiers was frightfully low. On stratifying his data by country of origin, he further found that "according to all evidence available . . . the American intelligence is declining, and will proceed with an accelerating rate as the racial admixture [having shifted, as he observed, from 'Nordic' to 'Alpine'] becomes more and more extensive" (Brigham, 1923, p. 210). Partly in response to these ominous tidings, the U.S. Congress enacted more restrictive immigration laws in 1924.

These laws were not revoked when Brigham later, to his credit, recanted his earlier prophecies (Brigham, 1930):

> *This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies—the writer's own—was without foundation.* (p. 165)

The Scholastic Aptitude Test

Brigham went on to develop the first standardized college admissions test [the Scholastic Aptitude Test (SAT)], which in essence was an IQ test tailored to the needs of the decentralized education system in the United States, again stressing the need to tap into innate potential uncontaminated by educational experience. Just as Spearman would have wished, the SAT soon became virtually mandatory for access to higher education in the United States.

Predictive Validities of College Admission Tests

Over the decades, these tests have been refined by infusing ever more sophisticated theoretical and computational advances. However, this did not help improve them in terms of traditional measures of test efficiency. For example, it has been well-known virtually since its inception that the SAT has validities for First Year College GPA (GPA1) that hover around 0.4, approximately 10 correlation points below those of High School Rank (HSR). A tabulation published by the College Board shows that over the time span from 1964 through 1982, the HSR validities varied little (between 0.46 and 0.54, thus explaining approximately 25% of the criterion variance). For the SAT, they range between 0.37 and 0.46 (17%). To make matters worse, Humphreys showed in 1968 that the ACT validities drop into the 0.20s (4%) once the prediction interval is extended to the eighth semester GPA (GPA8). Similarly, it has repeatedly been shown (e.g., by Horn and Hofer and by Sternberg and Williams) that for long-range criteria of practical interest, such as graduation, the validities of the GRE are virtually zero. These findings cannot be dismissed as being due to random error. In contrast to the dubious figures reported for commercial IQ tests, such as the Wechsler and the Stanford–Binet, the sample sizes for the SAT, ACT, and GRE often range in the millions.

From Model to Method: 1930s–1940s—Louis Thurstone

Multiple Factor Analysis

Thurstone's reign of psychometrics extends back into the late 1920's when he published a number of papers devoted to applications of testing models to psychophysics and social psychology, especially attitude measurement. However, he scored his greatest triumph when he extended Spearman's failed factor theory of intelligence from one common factor to more than one common factor (multiple factor analysis). In the process, he transformed factor analysis from a substantive theory into a "general scientific method" susceptible to widespread abuse.

The underlying idea seems straightforward. If, after partialling out one common factor, one finds the resulting partial correlations are still not zero, one might consider partialling out a second common factor, and perhaps a third, until all remaining partial correlations are deemed close enough to zero (multiple factor analysis). In hindsight, this idea seems so obvious that it may not surprise to learn that it had already occurred to others (e.g., Garnett in 1919) long before Thurstone took credit for it in 1947.

Simple Structure

Some additional technical work was required before this simple idea could become useful in practice. Virtually single-handedly, and with considerable ingenuity, Thurstone developed both the theoretical and the technical refinements needed to transform what Spearman had originally intended as a substantive theory of intelligence into a general scientific method.

His most important theoretical contribution was his concept of simple structure intended to resolve the so-called rotation problem. This problem does not arise with Spearman's model because it invokes only one common factor. If there is more than one common factor (e.g., two), then the same observed correlations can be described in infinitely many different ways so that a choice has to be made.

This problem is most easily understood geometrically. If there are two common factors, both can be viewed as the orthogonal axes of a two-dimensional coordinate system used to describe a swarm of test points in a plane. This

swarm of points can be described by many other coordinate systems, even if the origin is kept fixed and the two axes are kept orthogonal to each other. If one chooses a new pair of orthogonal axes by rotating the old system, then the coordinates of the test points—that is, their numerical representation relative to the chosen coordinate system—will change but not the relations among the points, and it is these that constitute the empirical observations. Thus, the question arises which coordinate system to select so as to maximize the scientific utility of the resulting numerical representation of the empirically observed relationships.

Thurstone solved this problem by appealing to the principle of parsimony that is usually attributed to Occam: Explanatory causes should not be multiplied beyond necessity. This means, in the present context, that each test should require as few nonzero coordinates ("loadings") as possible so as to explain it with the smallest possible number of common factors. Hence, after starting with an arbitrary coordinate system, an attempt should be made to rotate it in such a way that each test has as many near zero coordinates as possible. Thurstone called this ideal position "simple structure". If there are only two common factors, such a coordinate system is easily found by visual inspection. However, as the number of common factors increases, this task becomes more difficult. Only with the help of modern computers was this "orthogonal rotation problem" eventually solved to everyone's satisfaction. Thurstone later generalized this rotation problem to correlated factors. The between-factor correlations could then be analyzed for "second-order factors" (hierarchical factor analysis).

Using this methodology, Thurstone arrived at a comprehensive theory of mental tests that dominated American psychometrics during the 1940s and 1950s. Most psychometricians agreed with Thurstone that Spearman had been on the wrong track when he postulated a single common factor of intelligence. Of course intelligence is multidimensional, they argued. All one had to do to see this was apply Thurstone's general scientific method to any battery of intelligence tests. Proceeding in this way, Thurstone derived between 5 and 7 primary mental abilities (PMAs), Eysenck derived 4, R. B. Cattell derived 2 (crystallized and fluid ability), and Guilford derived no less than 120 intelligence factors. However, in practice, Thurstone's test, the PMA, did not perform markedly better than other tests that had been constructed without the benefit of factor analysis.

Consequences

In retrospect, it seems clear that Thurstone's approach was actually a step backward compared to that of Spearman. One reason why Thurstone, Cattell, Eysenck, Guilford, and numerous other psychometricians gave widely differing answers to Spearman's question—what do we mean when we say we "measure intelligence"?—was that they could not even agree on the number of intelligence factors. Yet, whatever the phrase meant, both Spearman and elementary logic tell us that it cannot possibly refer to a multidimensional concept of intelligence. Only unidimensional variables can be "measured" in the sense that exactly one real number is assigned to each subject so that, at a minimum, the empirical order relations (implied by statements of the type "A is more intelligent than B") are preserved. If there were two different intelligences, then all such statements would have to be qualified with a reference to the particular intelligence that is being measured.

In hindsight, it is not surprising that the Thurstone era of exploratory factor analysis, imposing as it seemed at the time, left few empirical imprints still remembered today. Where Spearman had set himself a clearly defined substantive problem, Thurstone promised a research technique uncommitted to any particular subject area, a technique that, moreover, could never fail. No correlation matrix can ever falsify what amounted to a tautological claim that a given number of observed variables can be reasonably well approximated by a smaller number of common factors. Statistical tests of the simple structure hypothesis, although available, were carefully avoided.

Although Thurstone's empirical results are virtually forgotten today, his gospel of a research automaton capable of dispensing scientific progress without requiring any ingenuity, technical skill, or even familiarity with a substantive area proved hugely popular and subsequently resurfaced in various statistical disguises.

1950s–1960s: Apogee—Louis Guttman and Lee Cronbach

Psychometrics reached its apogee in the 1950s under the leadership of Guttman and Cronbach. Both had strong commitments to both branches of psychometrics and were heavily engaged in empirical research.

Louis Guttman

Guttman was without question technically the most accomplished and creative psychometrician in the history of psychometrics. Early in his career he addressed fundamental problems in (unidimensional) scaling. He is most widely known for the joint scale that bears his name. In test theory, he subjected the seemingly innocuous notion of reliability inherited from CTT to searing logical criticism, anticipating later developments by Cronbach and others.

The Importance of Falsifiability

In marked contrast to Thurstone, Guttman never tired of stressing the need to test strong assumptions (e.g., simple structure). Turning his attention to factor analysis, he emphasized that neither parsimony pillar supporting Thurstone's edifice, small rank and simple structure, can be taken for granted. These hypotheses are not just strong but most likely false. A recurrent theme in Guttman's papers is the need to replicate empirical findings instead of simply relying on facile significance tests ("star gazers").

With his radex theory, he proposed a family of alternative structural models unconstrained by these assumptions. In 1955, he also revived interest in the dormant factor indeterminacy issue, recognizing it as a fundamental problem that undermines the very purpose of factor analysis. Thurstone, in contrast, had never faced up to it.

Facet Theory

Later in his career, Guttman returned to Spearman's question, What is intelligence?, which he sought to answer with his facet theory. This research strategy starts with a definition of the domain of the intended variable. Only after this has been done does it become feasible to determine empirically, with a minimum of untested auxiliary assumptions (such as linearity, normality, and the like), whether the staked out domain is indeed unidimensional as Spearman had claimed. If it is not, then one has to lower one's aim and concede that intelligence, whatever else it may be, cannot be measured.

Lee Cronbach

Cronbach also went his own way. As author of a popular text on testing, he was familiar with the problems practicing psychometricians had to face and not easily dazzled by mathematical pyrotechnics devoid of empirical substance. Just as Guttman before him, he also questioned the simplistic assumptions of CTT. Searching for alternatives, he developed a reliability theory that recognized several different types of measurement error—thus yielding several different types of reliability—to be analyzed within the framework of analysis of variance (generalizability theory).

Mental Tests and Personnel Decisions

Most important, in a short book he wrote with Goldine Gleser in 1957, the authors radically departed from the traditional correlational approach for gauging the merits of a test. Instead of asking the conventional questions in correlational terms, they asked a different question in probabilistic terms: How well does the test perform in terms of misclassification rates?

It is surprising that this elementary question had not received more attention earlier. In hindsight, it seems rather obvious that, since use of a test always entails a decision problem, its merit cannot be judged solely in terms of its validity. How useful a test will be in practice also depends on prior knowledge, including the percentage of qualified applicants in the total pool of testees (the base rate) and the stringency of the admission criterion used (the admission quota).

Base Rate Problem

That knowledge of the correlation between the test and the criterion alone cannot possibly suffice to judge the worth of a test is most easily seen in the context of clinical decisions, which often involve very skewed base rates, with the preponderance of subjects being assessed as "normal."

For example, assume the actual incidence (base rate) of normal is 0.90, and that for "pathological" it is 0.10. Suppose further that the test cutoff is adjusted so that 90% of the testees are classified as normal on the basis of their test scores and 10% as pathological. If the joint probability of actually being normal and of being correctly classified as such is 0.85, then one finds that the so-called phi coefficient (as an index of validity) is 0.44. This is quite respectable for a mental test. However, on using it, one finds the following for the probability of total correct classifications (i.e., the joint probability of actually being normal and also being so diagnosed plus the joint probability of actually being pathological and being so diagnosed): $0.85 + 0.05 = 0.90$. This value exactly equals the assumed base rate. Thus, the proportion of total correct classifications achieved on using the test could also have been achieved without it by simply classifying all testees as normal.

The moral of his tale is that the practical utility of a test is a joint function of, among other things, validity, base rate, and admission quota. Validity by itself tells us nothing about the worth of a test. Meehl and Rosen had made much the same point a few years earlier. Cronbach and Gleser expanded on it systematically, tracing out the consequences such a decision-theoretic perspective implies.

To my knowledge, the only currently available complete tables for hit rates (the conditional probability that a qualified testee passes the test) and total percentage correct classifications, as joint functions of validity, base rate, and admission quota, are those published in Schonemann (1997b), in which it is also shown that no test with realistic validity (<0.5) improves over random admission in terms of total percentage correct if either base rate exceeds 0.7.

Notwithstanding the elementary nature of this basic problem and its transparent social relevance, the Social Sciences Citation Index records few, if any, references to it in *Psychometrika*. This is surprising considering that much of modern test theory, with its narrow focus on item analysis, may well become obsolete once one adopts Cronbach's broader perspective. In contrast, some more applied journals did pay attention to the work of Meehl, Rosen, Cronbach, and Gleser.

1970s–1980s: Eugenics Revisited—Arthur Jensen

Some purists may wonder why Jensen is included in this review, since the psychometric elite tends to ignore his work. On the other hand, he also has many admirers. Brandt (1985) calls him "the world's most impressive psychometrician" (p. 222). Whatever one may think of his work, Jensen has clearly had a profound impact on the recent course of psychometrics.

How Much Can We Boost Intelligence and Scholastic Achievement?

Jensen achieved instant notoriety when he challenged the received view that intelligence is primarily a function of environment, not genes. This position had gained ground during World War II, gradually replacing the earlier eugenic thesis to the contrary. To appreciate the progress that the field had made since Spearman, note that logically neither stance makes much sense as long as intelligence still remains undefined.

In his *Harvard Educational Review* paper, Jensen proclaimed that previous attempts to narrow the black/white gap on IQ tests were doomed to failure because, according to him, blacks are deficient in the particular genes required for complex information processing. Although this message initially met with some skepticism, it slowly mutated into the new received view, virtually uncontested by psychometricians and statisticians and deemed worthy of a full-page ad in the *Wall Street Journal* that several prominent psychometricians signed.

Undeterred by his detractors, Jensen set out to back up his unorthodox thesis with extensive empirical evidence. He tabulated IQ scores for various ethnic groups, designed an apparatus for measuring complex reaction times, and contributed suggestions on how best to measure the genetic portion of the IQ test variance. In the end, all these efforts converged on the same conclusion: Spearman had been right all along—g does exist and intelligence can be measured. In addition, it is ubiquitous and of utmost importance in virtually all spheres of life. Just as the early pioneers had claimed, it is primarily genetically predetermined. True, blacks are disadvantaged in this respect. However, this is no cause for alarm, as long as costly but futile efforts to bring them up to the level of whites (such as Head Start) are replaced with more realistic alternatives to guard against "dysgenic trends" toward "genetic enslavement" (Jensen, 1969, p. 91f).

The psychometric and statistical establishment initially reacted with quiet scorn to these heresies, fearing perhaps unwanted public scrutiny of psychometrics more generally. What it did not do, in any case, was to squarely face up to Jensen's challenge and simply show what was wrong with his reasoning.

Jensen's *g* Ersatz versus Spearman's *g*

What was wrong with it was that Jensen, who greatly admires Spearman, had quietly abandoned Spearman's original theory by replacing his g factor with the first principal component (PC1): "For example, the g [the first principal component] extracted from the six verbal subtests of the Wechsler Adult Intelligence Scale has a correlation of 0.80 with the g extracted from the five Performance subtests, in the standardization sample" (Jensen, 1979, p. 17).

To understand why this seemingly innocuous change actually amounts to a grotesque perversion of Spearman's theory, one has to know how principal components differ from common factors. The technical difference is most easily demonstrated with recourse to elementary matrix algebra. Since a correlation matrix of p tests is symmetric, its eigenvalues are always real and can be ordered. The dominant eigenvector of R, if used as a weight vector, results in a linear combination, PC1, that has largest variance among all possible linear combinations of the observed tests (subject to the constraint that the weight vector be of unit length). Note that this is a definition of a PC1, not an empirical discovery. The variance of the resulting PC1 is given by the dominant eigenvalue c.

Artifacts

Now let us see what happens when the correlation matrix R is "positive" throughout—that is, it has only positive elements, which was the point of departure for Spearman. For the sake of argument, let us assume all correlations are equal to $r\,(>0)$. This greatly simplifies the algebra without unduly violating reality. In this case, R can be written as a sum, $R = r\mathbf{e}\mathbf{e}' + (1-r)I$, where r is the correlation, $\mathbf{e}$ is a column vector of p ones, and I is the identity matrix of order p. The left-hand summand is a matrix of rank 1. It has dominant eigenvalue $r\mathbf{e}'\mathbf{e} = pr$ [since $R\mathbf{e} = (r\mathbf{e}\mathbf{e}')\mathbf{e} = r(\mathbf{e}'\mathbf{e})\mathbf{e} = pr\mathbf{e}$], while all other eigenvalues are zero. The right-hand summand, $(1-r)I$, leaves all eigenvectors intact and simply adds $1-r$ to all eigenvalues. Hence, the dominant eigenvalue of R is

$pr + (1 - r)$, which equals the variance of the PC1. The remaining $p - 1$ eigenvalues are all equal to $1 - r$.

As a result, the percentage of variance of the observed tests accounted for by the PC1 of R is $100[r + (1 - r)/p]$, which tends to $100r$ as p increases. The ratio of the largest eigenvalue to the next largest eigenvalue is $pr/(1 - r) + 1$. This ratio quickly increases as the number of tests, p, increases. Concretely, if $p = 10$ and $r = 0.5$, then this ratio is $p + 1 = 11$. If $p = 20$ and $r = 0.33$, it also is 11. This means that a PC1 always explains much more variance in a positive data matrix R than any of the remaining $p - 1$ PCs, as soon as p is reasonably large.

This is the reason why Jensen was so successful in convincing the uninitiated that "g" (as he calls the PC1 in gross abuse of language) is a powerful variable wherever he looks. The problem is that a PC1 is not g as Spearman had defined it in 1904. Partialling out the PC1 does not leave zero partial correlations. Rather, every one of Jensen's PC1s is a local description of the data at hand. Spearman, in contrast, was searching for a g that underlies all intelligence test batteries, not just a particular one. Jensen obtains a different g every time he analyzes a new R.

Congruence Coefficients

Jensen finesses the question whether all his g's are the same by pointing to so-called congruence coefficients as "a measure of similarity" between the dominant eigenvectors extracted from different batteries: "*Congruence coefficients (a measure of factor similarity)* are typically above 0.95, indicating virtually identical factors, usually with the highest congruence for the g factor" (Jensen, 1998, p. 363, italics added).

Usually these cosines are high. The problem is that they tell nothing whatsoever about "factor similarity": Two sets of variables can have identical within-set correlations while all between-set correlations are zero. In this case, the congruence coefficient will be 1 and the correlation between both PC1s will be zero.

This is Jensen's second sleight of hand: It is not at all difficult to tabulate the cosines between dominant eigenvectors extracted from randomly generated positive R's and a vector of 1's (**e**) of the same order. When this was done, the cosines varied between 0.995 and 0.999 when the parent distribution was uniform and between 0.949 and 0.998 when it was chi-square 1 (Schonemann, 1997a, p. 806). This means that Jensen's g ersatz is simply the average test score of whatever tests he analyzes. It is not g, but a travesty of Spearman's g.

Unfortunately, there are few examples in the psychometric and, more significantly, the statistical literature of anyone seriously challenging Jensen on methodological grounds (rare exceptions are Kempthorne and Schonemann). Typically, he is challenged on ideological grounds because critics do not like his conclusions.

Retrospective

Looking back, it is difficult not to notice that psychometrics did not live up to the promise of its auspicious beginnings. After Thurstone had shifted the focus from substantive theory to general scientific method, the field progressively lost its moorings and began to drift toward "ideational stagnation" (Sternberg and Williams, 1998, p. 577). On the applied side, steadily declining technical standards eventually empowered prophets of dysgenic doom to revive the eugenic myths of the 1920s, virtually unopposed by psychometricians and statisticians alike. On the theoretical side, Thurstone's implied message "that knowledge is an almost automatic result of gimmickry, an assembly line, a methodology" (Koch, 1969, p. 64) easily won out against Guttman's stern admonition that the essence of science lies in painstaking replication, not in facile significance tests. Numerous "general scientific methods" came and went, ranging from multidimensional scaling ("a picture is worth more than a thousand words") to latent trait theory, item response theory, linear structural models (LISREL), and meta-analysis and log-linear models. None of them has markedly improved the practical utility of mental tests, let alone helped answer any of the basic theoretical questions, such as "What is intelligence?" Just as Spearman noted years ago, "Test results and numerical tables are further accumulated; consequent action affecting the welfare of persons are proposed, and even taken, on the grounds of—nobody knows what!" (Spearman, 1927, p. 15).

See Also the Following Articles

Binet, Alfred • Education, Tests and Measures in • Human Growth and Development • Intelligence Testing • Psychological Testing, Overview • Thurstone's Scales of Primary Abilities • Thurstone, L.L.

Further Reading

Brandt, C. (1985). Jensen's compromise with componentialism. *Behav. Brain Sci.* **8,** 222–223.

Brigham, C. C. (1923). *A Study of American Intelligence.* Princeton University Press, Princeton, NJ.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychol. Rev.* **37,** 158–165.

Cronbach, L. J., and Gleser, G. C. (1957). *Psychological Tests and Personnel Decisions.* University of Illinois Press, Urbana.

Dodd, S. C. (1928). The theory of factors. *Psychol. Rev.* **35,** 211–234, 261–279.

Donlon, T. F. (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests.* College Entrance Examination Board, New York.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common factor theory. *Br. J. Statistical Psychol.* **8,** 65–81.

Hart, B., and Spearman, C. (1912). General ability. Its existence and nature. *Br. J. Psychol.* **5,** 51–84.

Humphreys, L. G. (1968). The fleeting nature of the prediction of college academic success. *J. Educational Psychol.* **59,** 375–380.

Jensen, A. (1969). How much can we boost IQ and academic achievement. *Harvard Educational Rev.* **39,** 1–123.

Jensen, A. (1979). g: Outmoded theory or unconquered frontier? *Creative Sci. Technol.* **3,** 16–29.

Jensen, A. (1998). *The g Factor. The Science of Mental Ability.* Praeger, Westport, CT.

Kempthorne, O. (1978). Logical, epistomological and statistical aspects of nature–nurture. *Biometrics* **34,** 1–23.

Koch, S. (1969). Psychology cannot be a coherent science. *Psychol. Today* **14**, 64.

Schonemann, P. H. (1997a). The rise and fall of Spearman's hypothesis. *Cahier Psychol. Cognitive/Curr. Psychol. Cognition* **16,** 788–812.

Schonemann, P. H. (1997b). Some new results on hit-rates and base-rates in mental testing. *Chinese J. Psychol.* **39,** 173–192.

Schonemann, P. H. (1997c). Models and muddles of heritability. *Genetica* **99,** 97–108.

Sternberg, R. J., and Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in graduate training of psychologists? A case study. *Am. Psychologist* **52,** 630–641.

Sternberg, R. J., and Williams, W. M. (1998). You proved our point better than we did. A reply to our critics. *Am. Psychologist* **53,** 576–577.

Thurstone, L. L. (1947). *Multiple Factor Analysis.* University of Chicago Press, Chicago.

Public Sector Performance

Tony Bovaird
University of the West of England, Bristol, United Kingdom

Glossary

economy Provision of inputs of a specified level of quality at the lowest cost.
effectiveness Usually conceived to be the extent to which objectives have been achieved, but sometimes considered to be the ratio of outputs to outcomes. (Cost-effectiveness is usually conceived of as the level of achievement of objectives per dollar spent.)
efficiency The relationship between inputs (of a given quality) and outputs (of a given quality), or the rate at which inputs are converted into outputs. Technical efficiency relates physical inputs to physical outputs (e.g., staff productivity, which relates staff inputs to organizational outputs).
equity The achievement of the desired level of fairness or social justice, sometimes expressed in terms of the level of equality achieved (e.g., in relation to equality of opportunity, access, cost, use, or outcomes). It is often compared between specific groups in society, such as low-income groups, women, and ethnic minorities (horizontal equity), or between people in the same group but in different circumstances (vertical equity).
input A resource used to execute a policy or provide a service.
outcome The impact of the organization's activities on the welfare of stakeholders (which may be intended or unintended).
output The result of the activities of an organization; in the public sector, this is often the level of service provided. Where the level of service is difficult to measure, the level of activity is often used as a proxy.
performance indicator (PI) A variable whose value suggests the level of achievement of inputs, outputs, outcomes, equity, or sustainability or the level of achievement of the ratios between these concepts (such as economy, efficiency, or effectiveness). Such indicators are often imprecise, particularly because they may be jointly produced or subjective in nature.
performance measure A performance indicator that meets stringent tests of clarity, relevance, validity, reliability, causality, and ability to be aggregated.
performance management system A set of structures and processes for making decisions on the basis of the information from the performance measurement system.
performance measurement system A set of structures and processes for the production of performance information about a public sector organization or policy or project.
public sector The set of organizations whose activities are financed from public money and controlled, directly or indirectly, by elected politicians.
quality The level of achievement of a set of characteristics desired by stakeholders. An overall summary of the achievement of these separate characteristics is sometimes made (e.g., conformance to specification, fitness for purpose, or meeting of user expectations).
sustainability The extent to which current levels of performance are likely to be feasible into the future, given known constraints in terms of resources (physical and financial) and expected economic, social, environmental, and political conditions.
value for money An overall index of the level of achievement of economy, efficiency, and effectiveness (sometimes weighted mainly toward economy and efficiency).

The issue of public sector performance continues to be topical, but the concept is used in many different ways. Here, it is taken to mean how well the public sector meets the expectations of its different stakeholders (i.e., those people who significantly affect or are affected by a policy or the actions of an organization). The increasing interest in performance measurement, particularly during the second half of the 20th century, has been linked to the growth of public sector expenditure. Interest has especially focused on systems of performance measurement and management that will support performance improvement in the public sector. There is no single definition of public sector performance or of the dimensions of performance measurement. However, most stakeholders give some

importance to measures of economy, efficiency, and effectiveness, and many stakeholders add elements such as equity, quality, and sustainability. The characteristics of an appropriate system of performance measurement are contested, but the main thread in most approaches is the need to support appropriate performance management processes. Contractualization of public sector activities in many countries in the 1980s gave particular momentum to performance measurement, often aping private sector practices and usually giving heavy weight to efficiency. In the 1990s, public sector reforms became more oriented to improvements in public governance, but performance measurement has been slow to catch up with these developments.

Introduction

The concept of public sector performance has become common in both everyday and academic discussion, but it is used in many different ways, partly because there are no official definitions. Politicians often seem to view it as meaning the reinforcement of their ideological preferences or keeping the electorate happy. Service users are usually more concerned about service availability, quality, or price, whereas citizens seek reassurance that services will be available when needed but costs will be low in the meantime.

Here, it is taken to mean how well the public sector meets the expectations of its different stakeholders. It is therefore inherently a subjective term, although many stakeholders may use objectively based performance measures to assess it. It is also a multifaceted concept, rather than singular, since many stakeholders will have different perspectives.

The increasing interest in performance measurement, particularly during the second half of the 20th century, has been linked to the growth of public sector expenditure. At first, the main interest was from economists, keen to find ways of comparing the productivity of activities in the public sector with those in the marketized sector. Recently, there has been increased interest in systems of performance measurement and management that will support performance improvement initiatives in the public sector.

Although there is no single definition of public sector performance or its dimensions, most stakeholders give some importance to measures of economy, efficiency and effectiveness, albeit with varying weights. In addition, many stakeholders wish to add elements such as equity (sometimes differently expressed in terms of social justice, fairness, equality, etc.), quality (sometimes emphasizing particular dimensions of quality, such as reliability, accessibility, responsiveness, and due process), and sustainability (sometimes narrowly defined, e.g., as environmental sustainability, but often widely defined, e.g., as resource sustainability or electoral sustainability).

The economics literature considers the economic performance of the public sector as a whole (public finance theory), resource allocation between activities (welfare economics), and the design of optimal performance incentives between the individual and the organization (principal–agent and contract theory). Political science is particularly interested in the potential conflicts between social values and economic efficiency and in the political processes by which these conflicts are managed. Organizational theorists tend to consider public sector performance measurement as a set of socially constructed processes by which different stakeholders seek to increase their power within the governance systems to which they belong. On the other hand, policymakers (including politicians) and practitioners tend to view performance measurement simply as the process of producing performance information for use in their decision making.

This article approaches the subject from a management perspective, drawing on aspects of the theoretical perspectives but focusing primarily on how they throw light on practical issues surrounding the use of performance measures in developed countries.

Background

The roots of performance measurement grew from two very different traditions. The performance of individuals was first measured systematically in the "scientific management" approach associated with Frederick W. Taylor. The performance of organizational units was originally measured in the Soviet Five Year Industrial Plans, which started in 1928. This latter tradition was part of the more general turn to "planning," which was a key element of the modernist movement in the 20th century. Planning in Western Europe has its roots in spatial planning (Garden Cities and the industrial philanthropist colonies before World War I; town and country planning and mass housing programs after World War I; and regional planning, transportation planning, and health and education planning after World War II). An intrinsic part of planning involves comparing plan outcomes to plan targets, although not always with high degrees of sophistication.

In the 1950s, the monitoring of financial performance became more systematic in both private and public sector organizations, sometimes extending beyond budgetary control to unit costing. In the mid-1950s, Peter Drucker was influential in arguing for "managing by objectives" and "managing for results," giving a key role to performance measures. This paved the way in the 1960s for the development of the planning, programming, and

budgeting systems in the United States that emphasized nonfinancial measures of performance.

In the 1970s, the concept of value for money (VFM), incorporating "economy, efficiency, and effectiveness," came into vogue, particularly in the United States and United Kingdom. During the 1980s, VFM was increasingly incorporated into external audit regimes in many areas of the world, often with a particular emphasis on efficiency.

In the 1980s, performance measurement in the public sector became seen as critical to the success of the reform movements of the time, which became known collectively as the new public management (NPM). In the NPM, it was suggested that responsibility for budgets and for performance should be delegated down the organization to the managers whose decisions were most likely to impact performance. The corollary was that performance indicators would be agreed with these managers so that they could be held accountable for their decisions. Another strand of the NPM was a move to contract management (e.g., compulsory competitive tendering in the United Kingdom), whether internally in the organization or through externalization of services, which gave particular momentum to performance measurement, often aping private sector practices and usually giving heavy weight to measuring efficiency and unit costs.

In the 1990s, there was increasing interest in measuring quality of services (partially in order to defend public services against the cost-cutting pressures evident in externalization initiatives). In many European public sector organizations, this led to the utilization of the Excellence Model of the European Foundation for Quality Management (EFQM). This model has five "enablers" to excellence [for each of which performance indicator (PIs) could be developed] and four areas of results (people results, customer results, society results, and key performance results, with the latter category referring to internal indicators of success). At the same time, the Balanced Scorecard spread quickly from its roots in private sector performance reporting to many public sector agencies throughout the world. Its key innovation was the insistence that organizations should report not one bottom line (no surprise in the public sector) but at least four results areas (evidencing clear overlaps with the EFQM results areas)—financial results, customer results, internal business processes, and learning and growth. In each of these areas, organizations were encouraged to state objectives and matching performance indicators.

Recently, public governance has begun to rival the NPM as a paradigm for public sector management. It emphasizes the understanding of public decision making as a multistakeholder activity and not just a government activity, as a "fuzzy" negotiative process rather than a set of clear and firm events, and as a social and political process rather than simply managerial. Consequently, it suggests much more focus on measuring outcomes (particularly quality of life changes for key stakeholders) and measuring the quality of the processes by which decisions are made through the interactions of stakeholders. In this way, it gives more importance to the measurement of "value added" in economic, social, political, and environmental terms, and not only value added to direct service users. Performance measurement has been rather slow to catch up with these developments.

Types and Examples of Measures

Many different categorizations of performance indicators have been developed for use in public administration. One powerful set that has driven the relationships between central and local government in the United Kingdom was developed by the Audit Commission, which divided PIs into those that illustrated achievement in terms of strategic objectives, cost/efficiency, service delivery outcomes, quality, and fair access. Most such approaches are essentially based on finding PIs for all stages in the production of welfare model. This model suggests a simple causative chain as follows:

$$\text{Inputs} \rightarrow \text{outputs} \rightarrow \text{outcomes}.$$

Moreover, PIs are typically developed for the ratios between these categories: for example, the ratio between inputs and outputs (efficiency and productivity), the ratio between outputs and outcomes (effectiveness), and the ratio between inputs and outcomes (cost-effectiveness). Some of these categories are often broken down further; for example, process indicators are often developed to assess how well activities contribute to outputs, whereas quality and equity are often seen as particular elements of effectiveness. Because the production of welfare model is very general, the specific PIs that different stakeholders derive on the basis of it may vary significantly. For example, under effectiveness a service user group may emphasize the achievement of service reliability or speed, whereas professional staff may focus on conformity to the procedural standards established for the service.

It is widely accepted that some types of performance measurement are particularly difficult; for example, PIs for quality of service and for outcomes are generally seen as more problematic than PIs for quantity of service and for output levels. However, there have been major advances in recent years in developing and using PIs even in these areas. For example, the EFQM Excellence Model has helped in the measurement of quality and the Audit Commission has developed and piloted approximately 40 quality of life PIs for local government in England and Wales.

One category of performance has been rarely measured and used; that is, PIs for the achievement of equity goals. Nevertheless, PIs are indeed possible in this area. For example, in New Zealand there is legislation compelling local authorities to report who uses and who pays for each local authority service, and in the United Kingdom many local authorities have an annual equalities report, which reports achievements against an equalities plan. This suggests that the problem here is more political than technical.

What Is a Good Performance Measurement System?

Performance measurement systems can serve a range of purposes so that judgments of what is a good system must depend on its purpose. For example, some systems are designed essentially to give strategic direction to staff (through targets based on measures of strategic performance), others seek to enhance staff and organizational capabilities (through "stretch targets" based on performance measures), and yet others seek to support evidence-based management (through providing PIs for benchmarking exercises) and organizational learning (through providing measures of "what works"). Performance measurement can support each of these purposes.

However, there is another purpose for performance measurement in organizations that creates real problems—control of organizational behavior. Since staff will usually be alert to (and resent) this purpose, it typically gives rise to undesirable distortions in performance measurement, reporting, and behavior. Indeed, where the control purpose is paramount and the sanctions for poor performance are significant, this can result in the perverse control syndrome. In this syndrome, when top management specifies very precise PIs and clearly gives high priority to them, staff are likely to report successful performance against these PIs, whatever the real state of performance and whatever damage it does to the underlying mission and values of the organization. In practice, there are many ways in which staff can distort the values of PIs that get reported in any performance measurement system. It is therefore important that control is only one of the purposes of a performance measurement and reporting system, if such a system is not to be dysfunctional.

A good performance measurement system must also be able to help in all three phases of evaluation of the organization's activities: appraisal of options, monitoring of current activities, and review of past experience. However, the level of detail in monitoring is typically less than in appraising options, whereas review often uses the most detailed performance information, including data collected on a one-off rather than a regular basis.

These three phases of evaluation occur at all organizational levels, from top strategic decisions to front-line decisions on priorities in treating individual service users. It is therefore important to find performance measures that can be cascaded downward (and upward) in the organization so that there is consistency between the way in which strategies are assessed by top management (before, during, and after their implementation) and how operational decisions are evaluated elsewhere in the organization.

A further critical decision in designing performance measurement systems is the balance between self-assessment and external assessment. Although external assessment has the benefit of seeming more independent, internal assessment is usually better informed about the context of the service or activity, more sensitive to its important parameters, and more likely to convince those who will later need to implement its lessons. However, it may lack credibility to an external audience. Since external assessment is usually expensive, most performance measurement is likely to be done by self-assessment. In these circumstances, it is essential that proper auditing of the data is undertaken within the organization; otherwise, all managers will be under pressure to ensure that the data they report show their own organizational unit in a favorable light, which will contaminate the whole organization's information system and systematically disable its ability to identify the key problems and learn what are the appropriate solutions.

There is a major debate about whether a performance measurement system should attempt to provide a balanced and comprehensive picture of the organization's performance or whether it should focus on assessing performance in relation to the organization's priority objectives and make reports only when performance levels drop below predetermined critical values. The former approach is believed to allow a better informed strategic overview to be formed. It has the disadvantages, however, of being expensive, cumbersome, and often rather slow. The latter approach is quicker, cheaper, and more focused but may result in some important developments not being noticed in time, and some organizational units may "hide," being too small to merit review.

Finally, PIs within a performance measurement system are usually assessed against a number of criteria. In particular, they are expected to be clear (readily understandable to all users), relevant (related to the desired objectives and outcomes of the organization), valid (measuring what they purport to measure), reliable (likely to produce similar results when measured again in similar circumstances), able to be unambiguously interpreted (when they change in value), additive (across stakeholders and organizational units), not able to be manipulated (by the behavioral decisions of those whose performance is measured), and reasonably cheap to collect. Unfortunately, very few PIs ever pass all these tests, so all PIs in practice have conceptual limitations.

Issues in Performance Measurement

There are a number of major issues in performance measurement that are the subject of ongoing debate and research. Perhaps the most critical issue is the extent to which performance measurement and reporting are inevitably biased by the interests of the stakeholders concerned—the principal–agent dilemma. Economic analysis has not derived a conceptual solution to this issue, and it seems unlikely that it ever will. The most promising avenue for solving this dilemma appears to be "political"—the finding of appropriate checks and balances by means of which the undertaking of performance measurement and reporting is less likely to be biased. However, it should be clear that evaluation is inextricably bound with assessment against values, which must include the values of all stakeholders: Evaluation can never be value free. The approaches to performance measurement inherent in the EFQM and Balanced Scorecard approaches can be seen as attempts to give other stakeholders a clearer opportunity to put forward their interpretation of how performance in an organization can be viewed.

Another major debate concerns whether it is important to have a single overall measure of performance for a public sector organization. For many decades, the private sector was believed to have a major advantage because the measure of profit gives it a bottom line—a single performance measure. However, overall measures of performance have been undermined recently, in both private and public sectors, by Balanced Scorecard approaches. Moreover, there have been interesting new developments in England and Wales, in which all the public sector inspectorates have been pressed by government in recent years to score public agencies against the achievement of "corporate objectives," which typically include objectives for social inclusion, diversity, environmental protection and enhancement, antipoverty, antidiscrimination, and (less often) fair employment. Consequently, all corporate strategies and business plans in public agencies have been assessed against multiple "top line" objectives. Nevertheless, central government in the United Kingdom has also developed aggregate scores for local authorities (the so-called comprehensive performance assessment), health trusts (the star ratings system), and universities (albeit two-dimensional, with separate scores for research and teaching). However, few other countries seem interested in this one-dimensional approach.

The final issue concerns the amount of resources to be devoted to performance measurement. For many decades, it has been argued, particularly by academics, that the public sector has devoted too little effort to understanding its performance levels and that a learning organization would put more resources into performance measurement. However, the NPM reforms have changed this situation. Indeed, the United Kingdom has seen the emergence of an "audit society." Hood *et al*. have estimated that there are now more regulators, auditors, and inspectors in the United Kingdom than there are taxi drivers. Many now argue that the pendulum has swung too far in the opposite direction and it would be better to put more resources into improving performance rather than simply measuring it.

The Future of Public Sector Performance Measurement

Finally, the literature suggests the following future directions for performance measurement:

- Support for evidence-based policy and management
- Measuring outcomes and quality of life rather than outputs
- Measuring the capacity for innovation and dynamic change
- Measuring achievement of public governance principles and processes
- Using public management for problem solving rather than just dial watching
- More self-assessment (e.g., in EFQM), with appropriate auditing of data.

See Also the Following Articles

Critical Views of Performance Measurement • History of Business Performance Measurement

Further Reading

Audit Commission (2002). *Quality of Life: Using Quality of Life Indicators.* Audit Commission, London.

Bovaird, T., and Loeffler, E. (2003). *Public Management and Governance.* Routledge, London.

Drucker, P. F. (1954). *The Practice of Management.* Harper & Row, New York.

Hood, C. (1991). A public management for all seasons? *Public Administration* **69,** 3–19.

Hood, C., Scott, C., James, O., Jones, G., and Travers, T. (1998). *Regulation Inside Government: Waste-Watchers, Quality Police and Sleaze Busters.* Oxford University Press, Oxford, UK.

Kaplan, R., and Norton, D. (1996). *The Balanced Scorecard—Translating Strategy into Action.* Harvard Business School Press, Boston.

Paton, R. (2003). *Managing and Measuring Social Enterprises.* Sage, London.

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *Int. J. Public Administration* **18**(2/3), 277–310.

Q Methodology

Paul Robbins
The Ohio State University, Columbus, Ohio, USA

Glossary

concourse The possible range of opinions and subject positions on a specific topic.
condition of instruction The contextual statement against which the Q-set is sorted by respondents; for example, "Most agree/Most disagree" or "Most like myself/Most unlike myself."
factor array A model Q-sort, used for interpretation, that has been abstracted from the Q-sorts that significantly load on each factor.
factor rotation Judgmental or theory-driven rotation of factor axes in analysis. This contrasts with "objective" rotational procedures, including varimax and quartimax.
operant subjectivity A model of subjectivity that assumes individual viewpoints are self-referent and are expressed and behaved contextually.
Q factor Structured products of Q factor analysis from clusters of statistically similar Q-sorts.
Q-set A subsample of the concourse, individual stimuli for ordering/ranking by study respondents; usually quotes or statements but also may be a set of photographs, pictures, or other objects.
Q-sort The ordered ranking of the Q-set by an individual participant usually using a quasi-normal distribution, expressing the individual's ranking of individual statements/items relative to the condition of instruction (e.g. "most agree").

Q methodology uses a controlled technique to elicit subjective viewpoints in order to make them explicit and comparable. Unlike more common "R" methods, such as traditional ranking opinion surveys that are based on variance analysis and sample averages, Q is less concerned with comparing patterns of opinion between groups than it is with determining what these patterns are to begin with and determining their structure within individual people. It is therefore used to understand the relationship between subjective opinions/claims/understandings as they vary throughout populations. Based on a by-person factor analysis technique and a somewhat standardized protocol for data collection, Q reveals common patterns (factors) of subjectivity, allowing comparison of individual opinions based on their relationship to these idealized patterns. Since it elicits the patterns of opinion directly from the sampled individuals, rather than using a preassumed or a priori set of opinions formed by the researcher, Q method has been effectively employed not only by social scientists but also by practitioners in fields ranging from health science to public policy. Despite its long history, some unsettled issues remain in Q method, including the determination of the numbers of factors in a population, the use of normalized distribution in Q-sorts, and the question of the minimum number of loaders/study subjects required for factor stability.

History

Q method is a technique for the study of subjectivity pioneered in the 1930s, emerging in recent years throughout the human sciences and increasingly common across a range of disciplines. William Stephenson, a psychologist and physicist, first introduced Q in the 1930s in a letter to the editor of *Nature* in which he first described person versus trait correlation. Stephenson explained that traditional factor analysis technique, in which a population of n individuals is measured with m tests/stimuli, might be inverted so that n different tests/stimuli are measured by m individuals. In this way, factor analysis was altered so that individual people, rather than test items, become variates. The factors that emerged would suggest common and empirically grounded structures/relationships between individuals.

Stephenson's Q methodology extended this insight beyond by-person factor analysis, however, specifically using this technique to query individual people and have them rank matters of opinion in relation to one another. Stephenson published his definitive work on the subject, *The Study of Behavior: Q-Technique and Its Methodology*, in 1953, 20 years after this initial contribution, stressing the technique as part of a larger, more comprehensive philosophy. Q method allows a respondent to assemble a model of her or his own subjectivity, preserving its internal and self-referent characteristics during analysis. This work remains a seminal contribution to Q methodology as the science of subjectivity.

The technique has been adopted by researchers in political science, sociology, and psychology, and it has enjoyed popularity in professional practices ranging from marketing to the health sciences. The technique, moreover, has its own professional society and its own journal, *Operant Subjectivity*, which explores applications of the technique and discussion of statistical issues in the method.

Defining Subjectivity in Q Research

The systematic study of subjectivity outlined by Stephenson and later Q methodologists is undergirded by a few key assumptions about the nature of subjectivity. Subjectivity is assumed to be (i) communicable and (ii) operant.

In the first case, the subjectivity of individuals is understood to be self-reflexive, where individuals have a discursive awareness that enables coherent explanation of beliefs and motivations. In this sense, subjectivity refers simply to the distinction between "your" point of view and "mine," as articulated in communication. This assumption is not necessarily compatible with some understandings of the subject articulated in critical theory and psychoanalytic approaches to behavior but does provide a sound basis for the practical exploration of people's points of view.

In the second case, subjectivity is viewed as operant: behaved, self-referent, and contextual. It is behaved and self-referent in the sense that it is performed anytime someone articulates his or her point of view or agrees or disagrees with others. It is contextual in the sense that people's opinions on individual matters are interrelated and realized/articulated together as a coherent whole.

This last assumption suggests a departure from methodologies that inquire into the opinions of subjects based on isolated and apparently unconnected questions, as most opinion research and pollemetrics do. To understand someone's self-referent subjectivity requires propositions and queries that are internally related and whose interconnections are as important as their individual characteristics. Q methodology is specifically designed to query subjectivity in this way.

Measures of Subjectivity: Q versus R

Q method is most commonly compared with so-called R methods, the title of the latter derived from Pearson's product–moment correlation *r*. For Q methodologists, R method refers to any form of standard opinion or subjectivity research in which individuals are queried along a number of individual tests ("Too much money is spent on public education," "The Labor Party is good for local business," etc.), with which they assert scaling agreement or disagreement to varying degrees ("strongly agree," "slightly agree," slightly disagree," etc.). R analysis might then descriptively assess the proportion or character of the population agreeing or disagreeing with such a test: "45% of women strongly agree with statement *X*, while only 10% of men do." Alternately, R analysis can determine significant statistical factors of the population that assemble along particular test/question clusters or latent variates through correlation and factor analysis.

Q versus R Factor Analysis

Q method is not simply an inversion of traditional R factor analysis to group test stimuli or traits, although it does entail such an inversion. Demonstrated simply, factor analysis of a basic data matrix (Fig. 1) in social science typically involves the correlation down columns of *N* traits across a population of *n* persons. This procedure produces

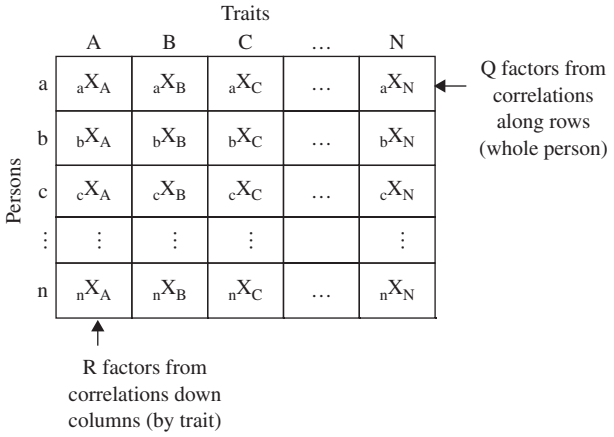


Figure 1 Basic data matrix.

a correlation matrix and a set of factors, against which specific traits can be evaluated based on their clustering or loading. Q analysis (as a factor technique and not a method) is the by-person factoring of correlations of traits, thereby identifying statistically correlated groups of people (i.e., whole individuals, not traits).

Such statistical analysis, in and of itself, may not be useful or meaningful insofar as the tests may be of differing orders or scales. For example, a survey might determine the heights, weights, and ages of a population. Traditional factor analysis might be performed on the population, determining patterns across the population for traits. Inverted "Q" factor analysis, determining correlations within individuals across the many traits, would be meaningless since the measures for height, weight, and age are incomparable. If, however, a survey determined the responses of an individual or a sample of people to opinion questions, and where each trait was a claim or opinion ranked or rated for level of agreement or importance to oneself, then such an inversion becomes useful and powerful. The internal relationships of opinions/issues to one another within people's personal frames of reference become measurable as groups of individuals (factors) of shared and common subjectivity.

Q Method versus R Method Assumptions

Q method was designed to elicit such factors of subjectivity directly from the population rather than a priori measures determined by researchers. If a researcher were interested in environmental opinions in a population, for example, traditionally he or she might produce a questionnaire designed to elicit scores on a predetermined scale. By answering questions one way on an environmental survey, for example, a given subject might score as a "visionary green," a "maybe green," or a "hard brown." The proportion and characteristics of these populations might then be described or examined. These designations (green, brown, etc.) and the threshold scores that designate them are determined, using traditional R approaches, by the researcher before the fact. The assumption is that there is an existing, coherent, subjective pattern out there in the world that scales from "green" to "brown."

Q method begins from a significantly different perspective and seeks as its goal to determine what structure of subjectivity is "out there" in the world, the very structure that R research assumes to exist before the fact (some people are greens and some are browns in the previous example). A Q method study of the same issue would, in contrast, assemble questions on environmental issues and related concerns, allow subjects to rank them, and, through the kind of analysis described later, determine what patterns exist, only later considering a given individual's similarity to those factors. Such a study, continuing from the previous example, may confirm the traditionally assumed pattern of "green" and "brown" viewpoints. It may determine, conversely, that such an axis of differentiation is meaningless, and that real differences lie between those concerned about health and those concerned about wildlife, for example, or some other variation that otherwise would be rendered invisible through the a priori categorization of opinions.

Structures of Subjectivity versus Populations of Subjects

Fundamentally, Q and R differ in the kinds of questions that are asked and the assumptions made in analysis. The two techniques are not contradictory methods but can in fact be highly compatible. Any issue in social science may be explored, for example, through exploratory Q method research, seeking to find empirically the variations of opinion that might inhere in key issues. These may later be explored through sampled survey techniques following an R format.

The main difference is that Q seeks to determine the structures of subjectivity and their variance, whereas R methods seek to characterize populations of subjects. R method can reliably ask and answer the question, "What proportion of women support gun control?" Q method, on the other hand, can reliably answer the question, "What are the variations of opinions about guns, and what are their internal logics?" The differences between each method suits a different phase of research, a different research agenda, and a different political and social project.

Steps in Q Method

Q method is carried out in a series of reasonably consistent steps: (i) The domain of subjectivity is determined; (ii) a concourse of statements is obtained or recorded; (iii) all the representative ideas of the concourse are sampled and included in a much smaller Q-set of statements or stimuli (pictures, words, etc.); (iv) Q-sorts of the statements are arranged by subjects, who rank the statements based on a condition of instruction (e.g., "most agree/most disagree"); (v) subjects are asked to comment in open-ended interview after sorting to explain their reasons and logic for the sorting of items; (vi) the Q-sorts are intercorrelated, two by two, and subjected to a by-person factor analysis; (vii) the factor structure is simplified by axis rotation and interpreted; and (viii) wherever possible, the resulting accounts of study subjectivities are returned to study subjects for review, comment, and reconsideration. Q method software programs have been developed to perform many of the statistical functions described here; software programs include PQ Method and PCQ, which are available from their designers through the Q Web site (www.qmethod.org).

Defining the Domain of Subjectivity

The domain of subjectivity is simply the broad area or topic of concern to the researcher, such as school funding, water quality science, lawn chemicals, or mad cow disease. In selecting the domain of subjectivity, it is useful to consider the breadth and openness of the area and the interest of the researcher. Selecting the domain of attitudes about forests might be more productive than specifying attitudes about deforestation, for example, especially for initial exploration of the range of subject positions in the area.

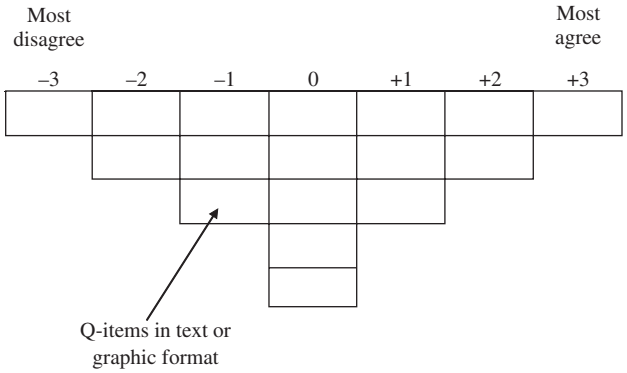


Figure 2 Example of typical normalized Q-sort structure.

Identifying/Sampling a Q-Set

Around the domain of subjectivity, a series of statements is next assembled. The selection of the sample (also known as a Q-set—items to be sorted by the subject), in terms of both composition and quantity, depends on the broader purpose and study design. In some cases, a sample of statements might be drawn purposefully from transcripts of intensive interviews or other techniques. Such a "naturalistic" sample is selected to represent a range of convergent and divergent views on varying facets of a topic of importance to all those interviewed.

Conversely, the statements may be drawn at random from a larger set of secondary quotes. Primary and secondary source materials may also be mixed to produce a comprehensive concourse. Similarly, a standard "deck" of quotes and statements, drawn from previous research, may be used for comparative purposes between groups or over time. The well-known California Q-set is one such standardized tool used in a range of personality studies. This option strays somewhat from the principle of drawing from context, however, which is important to many Q researchers. Moreover, the use of contextual or indigenous items in no way invalidates the generalizability of Q study results since these arguably represent localized reflections of more general patterns of subjectivity.

The number of statements in the concourse is also variable. Mathematically, the factor analysis can be performed with very few statements and some successful studies have utilized as few as 10 items. The maximum number of sorted items is reported to be more than 1000, although as a practical matter, a smaller sample of statements is equally effective.

The use of nonnarrative/text stimuli is also possible, although less prevalent in practice. Colors, photos, music, and even odors have been used in Q-sorts to good effect, especially in the examination of aesthetics, environmental perception, and landscape preference. This is also of particular use when working with groups or populations with highly divergent literacy skills.

Eliciting Q-Sorts from Respondents

The items in the deck are then presented to respondents, who rank order them based on a condition of instruction. The sample of respondents may be drawn randomly from a larger population or may be purposively selected to represent key positions, decision makers, or groups, depending on research goals and hypotheses.

The condition of instruction might be simple (e.g., "Which statements most reflect your feelings?") or more specific (e.g., "Which approaches to development provide more benefit to the local community?"). In a robust study, multiple sorts may be performed using the same items under varying conditions of instruction.

The items are then arrayed, usually along a normal distribution, as shown in Fig. 2, in which items placed in the middle are viewed as "indifferent" or less characteristic. This assumes that items at the extremes are the most salient or significant.

The actual shape and structure of the distribution curve of the Q-sort arguably matter very little since the factors of subjectivity tend to be broad and robust enough to be reproduced under a variety of configurations. Whether the distribution is +5/−5 or +3/−3, for example, has little effect on final results.

Correlation and Factor Analysis

As outlined by Brown, the steps that follow are fairly standardized. Correlation coefficients are computed based on the sorts, and for each pair of Q-sorts a Pearson product−moment correlation coefficient (r) is calculated. The matrix of these intercorrelations is usually not used by itself in analysis but is a step toward the factor analysis that follows. Although a range of newer and more sophisticated techniques are available for extracting factors, factor analysis is typically performed using the centroid or simple summation method. From this, a number of factors will be produced, each with its

own eigenvalues and percentage of explained variance. The factors in the unrotated matrix represent abstracted trends/clusters of similar Q-sorts—in other words, common or idealized subjectivities.

The number of factors prior to rotation is a matter of theoretical as well as practical concern. Statistically, a rule of thumb may be employed to retain all factors with eigenvalues greater than 1.00. This may force the analyst to overlook theoretically and substantively important factors, however, and should be employed with caution. Generally, each factor is examined in its structure and its relevance in the sample population. The factor may, for example, reflect strongly the views of a single, important individual and therefore be retained for a full and robust examination.

Factor Rotation

The factors are next rotated in multidimensional space to align the individual Q-sorts along idealized and abstracted subjective patterns and commonalties. Axis rotation is performed to produce greater clarity and perspective and in no way changes the positional relationships among the factors in multidimensional factor space. Unlike correlation and factor analysis, factor rotation may be executed by a variety of statistical algorithms but may also be performed entirely by manual or judgmental rotation in a less standardized and more theoretically driven fashion. This allows the testing of "hunches" and the exploration of "interesting possibilities."

Typically, the method of rotation is an "objective," statistically derived scheme, such as varimax rotation, in which the purpose, following McKeown and Thomas, is to "maximize the purity of saturation of as many variates (Q-sorts) as possible" along the fewest number of factors derived earlier. In this case, factors are aligned in an orthogonal fashion along perpendicular axes so Q-sorts that load high on one factor will load low on another, maximizing the distinction and differentiation of subject positions while minimizing the correlation among factors.

Such an approach may be less than satisfactory, however, for theoretical reasons, for testing hypotheses, or for revealing meaningful patterns. In many cases, the Q-sort of a key individual or group of individuals may be understood by the researcher as an ideal, important, or definitive type. By rotating the factors to align with those ideal sorts, it is possible to examine the degree of deviation from these opinions among other individuals and groups, for example.

The case of theoretical rotation is shown in Fig. 3, in which a theoretical varimax rotation aligns centroids using Q-sorts from locals and state officials concerning issues in forest management. In this case, the varimax rotation was discarded in favor of a theoretical rotation,

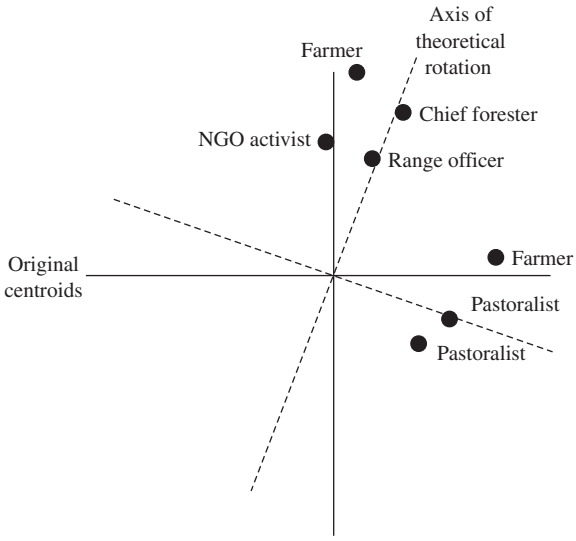


Figure 3 Example of theoretical rotation diagrammed in two-factor space.

wherein the subjectivity of specific specimens (the sorts of foresters and pastoralists) provides "ideal types" and can be examined both in contradistinction to one another and relative to that of other individuals and groups.

Factor Interpretation

Interpretation involves both an examination of the constituent items that make up each factor and the relationship of individual Q-sorts to each factor. In the first case, the factor arrays, those combined elements that represent abstract or idealized subject positions, are of the greatest interest to the Q researcher since they reflect, as Brown (1980, p. 247) notes, "attitudes as attitudes, quite independently of whoever may have provided them." These will take the form of a coherent story, with significant sorted elements making up strong negative and positive components in each factor/story. The factors can be understood as significant and empirically derived viewpoints that exist in the population.

Similarly, each individual Q-sort is scored with a loading on each factor. Some individuals, with high loadings on a single factor, will represent relatively "pure" articulations of a subject position, and their viewpoint is often used to define this factor. Others, however, may straddle multiple positions. This is both methodologically and theoretically desirable insofar as the definition of subjectivity underlying Q method assumes the possibility that predefined and conceptually simple descriptors of subjectivity (e.g., left/right, green/brown, and authoritarian/egalitarian) do not exist in real populations and that the complexity of subjectivity allows for multiple positions and subjectivities.

Iteration with Respondents

The final step of a Q study should involve a presentation to the study subjects of the derived factors in textual or graphic form. This not only allows assessment of the results (e.g., "How close does this factor describe your point of view?") but also opens the possibility for ongoing interpretation of what factors mean. This is especially the case when respondents' Q-sorts load significantly on more than one factor. It is also particularly useful if the technique is being used in social or therapeutic projects in which comparison and ongoing dialogue are research goals (e.g., participatory development projects).

Other Approaches to Q

Q method can also be applied "after the fact" to treat data not derived in the fashion outlined previously. R method questionnaires, for example, that incorporate Likert scale ratings or, more easily, ranking questions can be recoded as individual Q-sorts and subjected to Q factor analysis without violating the assumptions of Q methodology in any fundamental way. Such *ex post facto* uses of Q, however, should open the door to study designs that more explicitly allow survey respondents to model their own subjectivity in the ranking process. Nor is this approach universally accepted in the field since many researchers view Likert ratings as inherently incapable of capturing relevant meaning.

Example: Views of Water Quality Regulation

Maddock conducted a Q-based research project to determine the variation of subjective viewpoints among a commission of diverse stakeholders seeking to create total maximum daily load (TMDL) regulatory structures to manage non-point source pollution. The topic was ideal for Q method since the range of interests (industry, farming, regulatory, scientific, and environmental) was wide, and the traditionally assumed differences between the individuals and groups involved (e.g. "green" versus "brown" views of the environment or "left" versus "right" politics) did not seem to hold.

The study used 23 statements as items for sorting. These statements were drawn directly from in-depth interviews among the various groups and placed on large index cards for easy management and handling. The deck included a range of statements from a variety of perspectives, including the following:

> *The majority of decisions made on this TMDL were not technically based, they were politically based, they were modified to meet the politics.*
>
> *Nobody causes more erosion than farmers. So for how long can they justify regulating the construction industry for example, which is about 5% of the load, and leave farming which is 90% of the load?*
>
> *Non-point source modeling is uncertain. We shouldn't get caught up in the numbers but move ahead to the water quality goals.*
>
> *The knowledge, quantitative tools and models are accurate, but what we need is intense data collection to support calibration and verification.*

The sorts were administered over several weeks and then correlated and subjected to factor analysis. The results of analysis produced four meaningful and significant factors. A fifth factor, although statistically feasible, was determined to add little to the explanatory power of the model and, by reducing the number of defining variables, increased error. Following a varimax rotation, the characteristics of the factors were explored.

The factors reflected four general points of view among commission members: (i) a "technocratic" view, which held that science was accurate and reliable, that participatory structures were troublesome, and that agriculture and construction were the central water quality problems; (ii) a "scientific uncertainty" view, which held that although science was not fully reliable, participatory decision making was ineffectual, and it also held agriculture to blame for current pollution; (iii) a "participatory science" view, which held that science is unreliable and value laden, that participatory decision making is prerequisite to a solution, and that point source polluters and industry were central causes of degradation; and (iv) a "science is political" view, which held no faith in quantitative models and which held that participation was necessary, with little emphasis on pollution causes. A wide range of subject positions was evident, as was the absence of clear or simple binary divisions between views of science, participation, and causes of pollution.

Equally revealing, however, were the surprising constituencies and divergences of these factors across and between interest groups typically viewed as monolithic. As shown in Table I, stakeholder group affiliation did not necessarily dictate or define the specific opinion factor, especially within groups typically or popularly associated with monolithic views, as in the case of environmentalists. What accounts for variation? How are participatory processes and the uses of science implicated in opinion formation within and between groups? These results lead to further hypotheses and motivate further research. Such results are the hallmark of rigorous Q-based research.

Debates and Unsettled Issues in Q Methodology

Despite, and perhaps because of, the method's long history of ongoing development, several methodological

Table I Factor Loading by Individual Q-Sort in a Water Quality Study

		Factor			
Participant ID no.	*Interest group affiliation*	*A ("Technocratic")*	*B ("Scientific uncertainty")*	*C ("Participatory science")*	*D ("Science is political")*
1	Industry	0.87**	0.05	−0.13	0.01
2	Industry	0.73**	0.29	0.06	0.37
3	Government	0.69**	0.38	0.37	−0.22
4	Environmental	0.59**	0.02	0.00	−0.30
5	Agriculture	0.55**	−0.23	0.21	0.37
6	Government	0.41*	−0.57**	−0.35	−0.06
7	Forestry	0.01	0.88**	−0.07	0.21
8	Government	0.15	0.81**	−0.17	−0.01
9	Environmental	0.15	0.66**	0.46	0.11
10	Research	0.03	0.63**	0.28	−0.15
11	Environmental	0.48	0.56**	0.19	−0.15
12	Environmental	0.30	0.39**	0.34	−0.68**
13	Government	−0.03	0.08	0.89**	0.28
14	Environmental	0.36	0.16	0.74**	−0.29
15	Government	0.40	−0.05	0.47*	0.42
16	Industry	0.07	−0.06	−0.78**	0.06
17	Government	−0.24	−0.20	−0.12	0.56**
18	Government	−0.02	0.11	0.25	0.56**
19	Agriculture	0.20	0.27	−0.05	0.52**
20	Construction	−0.04	−0.04	−0.02	−0.50

*$p = 0.05$.
**$p = 0.01$.

issues remain under discussion. The appropriate number of Q-sorts to support robust factors is a matter of debate, as is the necessity of normalized distribution of Q-sorts. Optimal methods for the formation and rotation of factors is also a point of methodological development. Much discussion also focuses on the question of representativeness among factors and populations. In this last case, some observers mistakenly equate the small sample populations with poor representation and generalizability in Q study results. This is far from the case, however, and reflects, according to adherents, a misunderstanding of what is being generalized in Q study—characteristics of subjectivity rather than characteristics of populations. More productively, many discussions center on the role of Q method in helping to establish free and just communication and decision making more generally, leading to the possibility of "discursive democracy." For a current tracking of such discussions, the Q method Web site is a key source, as is the journal *Operant Subjectivity*, both published and maintained by the International Society for the Scientific Study of Subjectivity.

See Also the Following Articles

Correlations • Factor Analysis

Further Reading

Addams, H., and Proops, J. (eds.) (2001). *Social Discourse and Environmental Policy: An Application of Q Methodology*. Elgar, Cheltenham, UK.

Brown, S. (1980). *Political Subjectivity: Applications of Q Methodology in Political Science*. Yale University Press, New Haven, CT.

Dryzek, J. S. (1990). *Discursive Democracy: Politics, Policy and Political Science*. Cambridge University Press, Cambridge, UK.

Fairweather, J. R., and Swaffield, S. (1996). Preferences for scenarios of land use change in the Mackenzie/Waitaki Basin. *N. Z. Forestry* **41**(1), 17–26.

Maddock, T. (2001). *The Science and Politics of Water Quality Regulation: Ohio's TMDL Policy*. The Ohio State University, Department of Geography, Columbus.

McKeown, B., and Thomas, D. (1988). *Q Methodology*. Sage, Newbury Park, CA.

Robbins, P., and Kreuger, R. (2000). Beyond bias? The promise and limits of Q-methodology in human geography. *Professional Geographer* **52**(4), 636–649.

Stainton-Rogers, W. (1998). Q methodology, textuality, and tectonics. *Operant Subjectivity* **21**(1/2), 1–18.

Stephenson, W. (1935). Correlating persons instead of tests. *Character Personality* **4,** 17–24.

Stephenson, W. (1953). *The Study of Behavior*. Chicago University Press, Chicago.

Qualitative Analysis, Anthropology

D. Jean Clandinin
University of Alberta, Edmonton, Alberta, Canada

Glossary

context Refers both to the situations (places, times, relationships) in which participants live and to the situations in which data is collected.
data (field texts) The material or information collected from research or study participants.
method The techniques or procedures a researcher uses to collect data. Methods are consistent with the philosophic assumptions of a particular methodology.
methodology The fundamental epistemological and ontological view embodied in the research stance.
participant (informant) The people from whom data are collected; the term "subject" is used in quantitative research.
research reports (research texts) The reports written to convey to an audience research findings and their interpretations.
researcher signature The construction of a writer's identity and the development of an authorial presence within a text; a term attributed to Clifford Geertz. In research texts, a researcher signature points to the researcher's participation in producing the text.
voice The ways participants' experiences are expressed in both field texts and research texts.

Anthropologists have begun to question the kinds of methodologies appropriate to studying human experience; in this regard, it is of interest to explore the turn to qualitative methodologies in anthropology. Under the broad heading of qualitative research, there are a number of forms of research methodologies, each with a central purpose of enabling researchers to attend to human experiences. However, within each methodology, the research purpose may be grounded within different theoretical and epistemological conceptions of human experience. These methodologies are being used and developed in anthropology as well as in other disciplines.

Introduction

Broadly defined, the discipline of anthropology is the study of humans, their origins, and their religious beliefs and social relationships. Although there is agreement about such a general definition of anthropology, most anthropologists agree with Clifford Geertz, who has spoken of a crisis in anthropology. Anthropology, of all the human sciences, is the most self-questioning discipline. Within anthropology, there is a fragmentation; a period of instability has emerged, brought on by the growing realization of the complexity of trying to study human experience. Increasingly, anthropology is being defined relative to another discipline or field of study. For example, cultural anthropology, medical anthropology, economic anthropology, and paleoanthropology have emerged as unique areas as anthropologists have recognized and attempted to study human experience systematically. As researchers within anthropology have moved away from a belief in what Geertz described as "a single and sovereign scientific method and the associated notion that truth is to be had by radically objectivizing the procedures of inquiry," there has been a turn toward qualitative research methodologies.

The term "qualitative" turns the attention of researchers toward trying to understand the qualities of something. Usually, qualitative, in a research context, is a term defined to contrast with "quantitative." This distinction between quantitative and qualitative research is frequently problematic and only partially helpful. Whereas qualitative means a researcher is attending to the qualities of an experience, sometimes attending to qualities can involve quantifying aspects of an

experience. In general, as stated by Norman Denzin and Yvonna Lincoln, "qualitative researchers stress the socially constructed nature of reality, the intimate relationship between the researcher and what is studied, and the situational constraints that shape inquiry." Qualitative research, as a set of research practices, is composed of many methodologies. That said, however, in general, as framed by Denzin and Lincoln, "qualitative research is multimethod in focus, involving an interpretive, naturalistic approach to its subject matter. This means that qualitative researchers study things in their natural settings, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them." In the following discussions, the term "qualitative research" points to the centrality of attending to peoples' experiences and, more particularly, to the ways researchers understand and compose meanings about those experiences.

History of Qualitative Research

Denzin and Lincoln offered one helpful way to understand the history of the development of qualitative research—that is, as a series of moments. Each temporal period, or moment, marks a shift in the ways qualitative research is defined. The five moments are defined as traditional (1900–1950), modernist (1950–1970), blurred genres (1970–1986), crisis of representation (1986–1990), and postmodern (1990–present). Each succeeding moment or time period includes its successors, with the result that, in present time, traces of all of the ways qualitative research have been defined are present.

Kinds of Qualitative Analysis

An exhaustive review of all qualitative methodologies being used in anthropology is beyond the scope here, but the most common qualitative methodologies are grounded theory, ethnography, phenomenology, narrative inquiry, and case study. Each qualitative methodology, with its attendant analysis, focuses on the qualities of experience, although in different ways (for example, descriptive analysis, categorical analysis, thematic analysis, or narrative analysis).

Grounded Theory Methodology

Grounded theory methodology is a research methodology with a central purpose to study the experience of participants in order to develop a theory grounded in the data gathered from participants. The qualitative analysis draws mainly on interview data from numerous participants in order to construct a grounded theory. Based on that grounded theory, a researcher is able to construct hypotheses and make predictions about other experiences.

Ethnographic Methodology

Ethnographic methodology is a research methodology with a central purpose to study a group of individuals within the setting in which they live and/or work and to construct a portrayal of those individuals that describes the shared patterns of group behavior, beliefs, language, and so on. An ethnographic analysis draws on a range of data including field notes, interview transcripts, documents, and artifacts in order to delineate themes, issues, and group behaviors that have developed over time in the local setting. There are unique types of ethnographies, including realist ethnography (objective, scientifically written), confessional ethnography (report of an ethnographer's fieldwork experience), autoethnography (reflective examination of an ethnographer's experience), microethnography (focused on a specific aspect of a group), critical ethnography (focused on the shared patterns of a marginalized group with the aim of advocacy), and feminist ethnography (focused on women and cultural practices that serve to disempower and oppress) (Cresswell, 2002).

Phenomenological Methodology

Phenomenological methodology has as its central purpose to study a phenomenon that a number of individuals might share and to discern the core or essence of the experience of the phenomenon. Phenomenology is a methodology grounded in lived experience that attempts to transcend lived experience in order to situate and comprehend a particular lived experience. A phenomenological analysis draws primarily on interview data.

Narrative Inquiry Methodology

Narrative inquiry methodology has as its central purpose to study the storied experience of one person or a number of individuals. Narrative inquirers describe the lives of individuals, collect and tell stories about the lives of individuals situated within cultural, social, and institutional narratives, and write narratives of the experiences of those individuals. A narrative inquiry draws on a range of data (field texts), including conversation transcripts, interview transcripts, artifacts, photographs, field notes, documents, memory box items, autobiographical writing, and journal writing.

Case Study Methodology

Case study methodology has as a central purpose to study a bounded system, an individual, whether that individual is a person, an institution, or a group, such as a school

class. The purpose is to provide an in-depth understanding of a case. There can be different kinds of case studies, including intrinsic, instrumental, or multiple cases. There can be multiple forms of data, including artifacts, documents, and interview transcripts. There are at least seven presentation styles: realistic, impressionistic, confessional, critical, formal, literary, and jointly told.

Key Considerations in Qualitative Design

These are central considerations in all forms of qualitative analysis.

Starting Points in Qualitative Analysis

In its most general sense, the starting point for all qualitative methodologies in anthropology is human experience. Although each kind of qualitative methodology adopts a particular purpose and allows a researcher to attend in a particular way, the overall research purpose is to understand the meanings of human experiences.

Role of the Researcher

In qualitative research, considerations regarding the role of the researcher are central throughout a study. Because the researcher is the central instrument in all phases of the research process, from framing the question, to sampling, to gathering data, to analyzing and interpreting data, and to preparing the research reports, it is crucial that researcher knowledge is considered. Central to the process of qualitative research is the researcher living out his/her autobiography and speaking from the perspective of a particular background. In some qualitative methodologies, researchers are advised to undergo a process of self-examination and analysis in order to bracket out their subjectivity and to attempt to set aside their biases. In other more relational qualitative methodologies, such as narrative inquiry, researchers are advised to give an account of how they have shaped the research process. In addition, the struggle to share the experiences of others from within (that is, alongside the research participants), as well as to be able to observe participants with some detachment, is present for all qualitative researchers.

Participant and Site Selection

Selection of participants and research sites is crucial to any form of qualitative analysis. People and sites are selected not on the basis of random sampling but as instances that might best help a researcher understand the particular experience under study. Such sampling is called purposeful sampling. Nine purposeful sampling strategies are used in most forms of qualitative analysis: maximal variation sampling, when a range of participants with the most variation in experience is chosen; extreme case sampling, when the extremes of participants are chosen; typical sampling; theory or concept sampling, when a clear understanding of the concept or theory is in place; homogeneous sampling, based on membership with defining characteristics; critical sampling of exceptional cases; opportunistic sampling that permits a researcher to take advantage of unfolding events; snowball sampling; and confirming/disconfirming sampling.

Participant Voice

Another consideration in qualitative analysis is the place of the voices of the participants. This is a particularly compelling issue in anthropology because issues of power and speaking for others are particularly important. Dependent on the kind of methodology, participants are more or less involved in framing the research puzzles and/or questions, in collecting and analyzing the data, and in writing the final research texts. For example, in case study methodology, participants often help frame issues or questions around the case but then act as sources of data. In narrative inquiry, participants frequently play a collaborative role in actively shaping and reshaping the research puzzle and carry through to working on final research texts. In grounded theory, participants are sources of data (that is, informants) and then serve to validate data through member checks. However, in all forms of qualitative analysis, the way participants' voices will be included is a key consideration. When researchers are working in cross-cultural settings in anthropology, issues of language and culture are especially sensitive. When researchers rely on interpreters, they become vulnerable to added layers of interpretive complexity.

Kinds of Data

The most common data format in qualitative analysis is interview data. A relatively new form of data is conversational data. In interviews, the researcher poses the questions and the participant answers by framing responses in the researcher's terms. In conversation, researcher and participants co-construct data by sharing their experiences. Conversation or dialogue shifts the power differential between researcher and participants to a more mutual, equally vulnerable relationship.

In some forms of qualitative analysis, other kinds of data are also collected, such as field notes based on participation and/or observation (along a continuum from observer to observant participant to participant observer), documents, individual artifacts (such as photographs and memory items), journal writing, and community artifacts.

Depending on the research methodology and the methods by which data are gathered, the kinds and range of data will vary.

Data Saturation

A related issue is what constitutes enough data. This is most frequently called data saturation. In most qualitative research, the amount and variety of data collected are usually almost overwhelming before researchers decide they have enough. Data saturation is reached in different ways, depending on the qualitative methodology used. Sometimes saturation is said to be reached when no new information is being gathered, as in grounded theory and phenomenology. However, in relational forms of qualitative research such as narrative inquiry, the length of time and the depth of the researcher–participant relationship are better guides to knowing when enough data have been collected. As Barbara Myerhoff writes of ethnographic fieldwork, "There is no definite or correct solution to the problem of what to include, how to cut up the pie of social reality, when precisely to leave or stop." The question of "enough" data relies in many ways on researcher judgment or time constraints imposed by participants, researcher, or an outside constraining force such as a funding agency.

Key Considerations in Qualitative Analysis

Data Preparation and Organization

When data have been collected within the parameters of the selected methodology and using methods congruent with the methodology, data must be prepared for analysis. Interview tapes must be transcribed, dated, and labeled as to context and participant identities. Artifacts must be dated and labeled as to context. The data can then be organized in four ways: by type (for example, by sorting interview transcripts as distinct from journal records), by site (for example, by the place of the data collection), by participants, and by time intervals. In a narrative inquiry, a researcher would tend to organize by participant. However, if the inquiry stretches over a long time frame, a researcher may also organize by temporal period—that is, by the data collected with participant A during the first 6 months and subsequent time intervals, by the data collected with participant B during the first 6 months and subsequent time intervals, and so on.

A related issue is deciding whether the qualitative analysis will be completed by hand or whether a computer program will be used in the analysis. This decision will partially affect how the data are prepared and organized. Some forms of qualitative analysis are better suited to computer analysis; examples include grounded theory methodology (using such computer programs as NUD*IST or NVIVO) or some ethnographic methodologies (using a computer program such as Ethnograph). More relational methodologies, such as narrative inquiry, do not easily lend themselves to computer data analysis, although having the data entered into text processing programs allows word searches that are helpful in locating connected themes.

Exploring the Data

When the data have been prepared and organized, the next consideration surrounds the ways researchers immerse themselves in the data. Given that the overall purpose of qualitative research is to understand human experience, it is essential that the researcher read and reread the data, studying the participants' words and silences, gaps, and hesitations in the transcripts, studying photographs and artifacts for multiple meanings and possible interpretations, and reading the field notes in order to try to remember the experience of being in the setting with the participants. Frequently, researchers listen and relisten to the audiotaped materials as they read the transcripts. This process allows a researcher to attend closely to the speech, sounds, and emotions being conveyed. As researchers engage in this immersion process, they begin to interpret the data as they write notes about what they are attending to and/or write theoretical memos about what they are beginning to theorize from the data. As they begin the reflexive and recursive process of analyzing and interpreting the data, they may begin to code the data. In this process, they may begin to foreshadow emerging issues, categories, themes, and/or threads in the data.

Coding the Data

Depending on the research purpose and the methodology being used, data can be coded descriptively, categorically, thematically, or narratively. For example, in most forms of ethnography, data are coded in order to provide thick descriptions of the culture under study. Though data can be coded in many ways in ethnography, in the broad view, ethnographic analysis is the exploration of cultural elements and their relationships as conceptualized by individuals. Data coding can begin when the researcher begins collecting data. Within ethnography, there can be at least four kinds of ethnographic analysis: domain analysis involves a search for the larger units of cultural knowledge (domains), taxonomic analysis involves a search for the internal structure of domains, componential analysis involves a search for the attributes that signal differences among symbols in a domain, and theme analysis involves a search for relationships among domains and how they lead to culture as a whole. The coding occurs

on two levels: at the level of small details of a culture and at the level of the broader features of the cultural landscape.

In grounded theory methodology, there is a continuous interplay between analysis and data collection using a constant comparative method whereby particular data points are constantly compared to other data points in order to form categories and concepts. In case study methodology, the process of coding data occurs over the entire research process. In intrinsic case studies, the researcher codes data to document the importance of a case within the context of the situation, and issues, contexts, and interpretations are developed. The coding is undertaken with an eye to seeking patterns of data to develop issues.

In phenomenology, the process of coding data begins with researcher immersion in the data and continues through incubation, illumination, and explication phases. The overall purpose of phenomenological analysis is to delineate the major essences of a particular phenomenon. In order to do this, the data are coded into categories. These coded categories are then read and reread recursively into the data. In an interpretive process of delineating categories and themes through a process of condensation, essences of a particular experience are illuminated and explicated.

In narrative inquiry, coding is completed with an eye to developing narrative accounts of participants' experiences. Coding occurs within a metaphoric three-dimensional narrative inquiry space, with attention to the personal/social dimension, the temporal dimension, and the dimension of place. Coding of events or actions that lead to narrative themes within a participant's life, narrative plotlines of a participant's life, and narrative threads are all possibilities. As threads, themes, or plotlines within an individual life become evident through coding of individual actions or events, institutional, cultural, family, and/or social narratives that shape a life become apparent and are also coded.

Coding data in all qualitative methodologies is an interpretive process. Although there are particular methods for seeking consistency in the process of coding the data, in qualitative analysis, a great deal depends on the meaning the researcher sees in particular pieces of data.

Interpreting the Data

When data are tentatively coded, they are then sorted into themes (in grounded theory), thick descriptions (in many kinds of ethnography and case study), essences (in phenomenology), and narrative threads or plotlines (in narrative inquiry). The process is a highly interpretive one, guided by what the researcher sees as emerging from the data. As noted earlier, the subjective role of the researcher is central to the process. There are various ways that researcher subjectivity can be accounted for, including a process of bracketing, creating an audit trail that other researchers can follow to verify interpretations, and having other researchers code and interpret some of the data to create interrater reliability. Other ways to verify the coding and interpretive processes include member checking, whereby codes and interpretations are checked with the involved participants for accuracy. Still other processes involve triangulation, whereby multiple kinds of data or data from multiple sources are cross-verified for adequacy of coding and interpretation. However, in narrative inquiry, which is a relational form of qualitative analysis, interim interpretations are also negotiated with participants in order to ensure that an interpretation resonant with participants is being created. Interpretation of data is frequently a messy process. Sometimes as data are interpreted, researchers realize that more data are needed before an adequate interpretation can be made. In these cases, researchers need to return to the field to gather more data from participants. The process is rarely linear; researchers move from living in the field and collecting field texts to coding and interpreting field texts and writing research texts. Issues of participant voice are central to data interpretation and to all attempts to compose an understanding of human experience.

The Place of the Research Literature

What is already published in the research literature is salient at every step of the research process. Although it is important to try to understand a particular aspect of human experience on its own terms in qualitative research, it is also important to read as widely as possible about that aspect of human experience, prior to beginning the research. Situating the research question within the literature allows a more informed framing of the question. Some qualitative researchers frame research questions using particular formalistic categories such as race, class, or gender, whereas other researchers begin with research puzzles or enigmas situated within the participants' and the researchers' experiences. However, for a researcher at any point along this continuum, it is important to have read broadly in the area of research interest. At the point of coding and interpretation, it is also important to situate the data alongside existing research. During coding and interpretation, other research literature may become relevant because of the emerging interpretations of the experiences of participants.

Key Considerations in Writing Qualitative Research Texts

As with all kinds of social science inquiry, research texts drawing on qualitative analysis require evidence,

interpretive plausibility, logical constructions, and disciplined thought. However, within each particular kind of qualitative analysis, there are some unique features concerning forms of representation, the audience, ethical issues, and criteria for assessing qualitative research texts.

Forms of Representation

Forms of representation (data displays) vary across qualitative methodologies. However, issues of form, voice, and researcher signature are central to each qualitative analysis. Forms of representation include thick descriptions and portraits for case study and ethnography; narrative accounts for narrative inquiry; midlevel theories for grounded theory, and descriptions of essences for phenomenology. Some qualitative researchers use arts-based forms of representation or forms called "bricolage" (from the French, *bricolere*, to putter about), a term meaning using anything that is convenient and available, now used in anthropology to denote the more improvisatory way that data can be represented in order to offer multiple meanings. Within each form of representation, unique features vary, depending on the research purpose. Although research reports have fairly standard formats when quantitative methodologies are used, research reports produced when qualitative methodologies are used are often quite different in form, depending on the researchers' interpretive frame and purpose. Researchers frequently experiment with form in order to find ways that best represent the participants' experiences. Qualitative methodologies in anthropology lead researchers to work *ad hoc* and *ad interim* as they compose their research texts.

Questions of Audience

The audience for whom research reports are prepared also influences the forms of representation. Funding agency requirements and university requirements for theses and dissertations influence the forms of representation. Journal and book formats determine, to some extent, the ways that visual and textual materials can be included. Although some journals allow flexibility in form, others do not. As electronic publishing becomes more common, qualitative researchers may continue to push audiences to take seriously more experimental forms of representation, including arts-based forms, bricolage, and a range of narrative forms.

Ethical Issues

Ethical issues and tensions are apparent throughout research employing qualitative methodologies. Qualitative researchers are governed by institutional research ethics boards, but issues of confidentiality and anonymity take on added significance when human experience is being studied. These issues need to be carefully negotiated with participants at the outset of the study. They may also need to be renegotiated at the request of participants and/or researchers at different times throughout the study. Furthermore, many qualitative researchers negotiate interim research texts with participants to ensure that the participants' experiences are represented in a resonant way. Although the guiding principle of all research ethics is to do no harm, the intensity of participant/researcher relationships during qualitative research frequently requires an attentiveness to the participants' experiences that is not necessary when quantitative methodologies are used. In qualitative research with some cultural groups and age groups, such as aboriginal cultural groups and children, additional issues around informed consent also need to be considered.

Criteria for Assessing Qualitative Research Texts

Although it is generally agreed among qualitative researchers that the criteria for judging qualitative research are not validity, reliability, and generalizability in the ways those terms are understood in quantitative methodologies, the criteria for judging qualitative research are still under development. Triangulation, member checking, and audit trails that allow external researchers to reconstruct the research process are used in some qualitative methodologies. Criteria such as plausibility, persuasiveness, authenticity, and verisimilitude are under consideration. Resonance with the experience of readers is another criterion currently in use as a way to judge the quality of research.

Features of Qualitative Analysis

In addition to variability across qualitative research methodologies, there are some features of qualitative analysis that cut across all qualitative methodologies in anthropology. Results of qualitative analysis are always bound by a particular time frame, particular contexts, and particular participants' experiences. There is no one generalizable truth that cuts across experience in general. That said, however, there is a continuum across qualitative methodologies that speaks to how bound by individuals' experiences the results are. For example, at one end of the continuum, a grounded theory methodology does generate a middle-level theory that speaks to others' experiences within a limited range. At the other end of the continuum, a narrative inquiry methodology gives an account of the experience of only those participants

in the study, although there may be resonance with other readers' experiences.

Qualitative analysis does not yield one ultimate truth but offers multiple possible interpretations of human experience. The purpose of qualitative research is to offer insights into the nature of human experience, and results are always open to other interpretations and other understandings. Qualitative analysis is a recursive process that moves back and forth across data gathering, coding, and interpretation. Because of the unfolding recursive nature of qualitative research, it is not always apparent what will emerge. Researchers working with qualitative methodologies need a tolerance for ambiguity and the ability to follow promising avenues in data collection and interpretation that may not have been evident when research questions were initially formulated. This requires tenacity and a willingness to stay with the research, often for extended periods of time.

Given the reliance on language to represent lived human experience, the writing of research texts is also a recursive reflexive process. Researchers need to stay open to imagining alternate forms of representation that may emerge from the interpretations of the data. Clifford Geertz wrote that the research texts needed in anthropology are "tableaus, anecdotes, parables, tales: mini-narratives with the narrator in them." Many qualitative researchers establish sustained response communities to allow them continual feedback as they code and interpret data and write research reports. The complex, multilayered, ambiguous, and unfolding process of research undertaken using qualitative methodologies is one that requires researchers to stay engaged and thoughtfully aware of the idea of trying to understand and represent the complexity of human experience.

See Also the Following Articles

Ethnography • Phenomenology

Further Reading

Bogdan, R. C., and Biklin, S. K. (1998). *Qualitative Research for Education: An Introduction to Theory and Methods.* 3rd Ed. Allyn & Bacon, Boston, MA.

Clandinin, D. J., and Connelly, F. M. (2000). *Narrative Inquiry: Experience and Story in Qualitative Research.* Jossey Bass, San Francisco, CA.

Cresswell, J. W. (2002). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research.* Merrill Prentice Hall Publ., Upper Saddle River, NJ.

Denzin, N. K., and Lincoln, Y. S. (eds.) (1994). *Handbook of Qualitative Research*. Sage, Thousand Oaks, California.

Eisner, E. W. (1991). *The Enlightened Eye. Qualitative Inquiry and the Enhancement of Educational Practice.* Macmillan, New York.

Geertz, C. (1988). *Works and Lives: The Anthropologist as Author.* Stanford University Press, Stanford, CA.

Geertz, C. (1995). *After the Fact: Two Countries, Four Decades, One Anthropologist.* Harvard University Press, Cambridge, MA.

Glaser, B., and Strauss, A. (1967). *The Discovery of Grounded Theory.* Aldine, Chicago, IL.

Lévi-Strauss, C. (1966). *The Savage Mind.* (G. Weidenfeld and Nicolson Ltd. transl.). University of Chicago Press, Chicago, IL.

Lincoln, Y. S., and Guba, E. G. (1985). *Naturalistic Inquiry.* Sage, Newbury Park, CA.

Myerhoff, B. (1978). *Number Our Days.* Simon and Schuster, New York.

Spradley, J. P. (1979). *The Ethnographic Interview.* Harcourt, Brace, Jovanovich, Ft. Worth, TX.

Spradley, J. P. (1980). *Participant Observation.* Holt, Rinehart and Winston, New York.

Stake, R. (1994). Case studies. In *Handbook of Qualitative Research* (N. Denzin and Y. Lincoln, eds.), pp. 236–247. Sage, Thousand Oaks, CA.

Stake, R. E. (1995). *The Art of Case Study Research.* Sage, Thousand Oaks, CA.

Strauss, A., and Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* Sage, Newbury Park, CA.

Van Maanen, J. (1988). *Tales of the Field: On Writing Ethnography.* University of Chicago Press, Chicago, IL.

Qualitative Analysis, Political Science

Kevin G. Barnhurst
University of Illinois at Chicago, Chicago, Illinois, USA

Glossary

archival techniques The use of source materials stored in any media format (books, manuscripts, paper, electronic files, or cultural artifacts) to gather information about human activity or to pursue humanistic ends.
case study A particular example of a phenomenon, ranging from a single individual to a larger institution or from a single event to a longer process, separated for close, multidimensional study.
documentary The collection or creation during fieldwork of informational items or artifacts in any medium, including written, audio, or visual records.
fieldwork Any human-centered method of study involving direct interaction with individuals, groups, or entities (past or present) in their natural surroundings.
participant observation A technique of fieldwork in which the researcher joins in the activities of those being studied, while simultaneously documenting the experience.
unstructured (ethnographic) interviewing A technique of fieldwork in which the researcher asks for and documents information from a research participant, who is allowed to respond at will.

Qualitative research methods include any techniques, except those focused primarily on counting, measuring, and analyzing statistical data, to study any social phenomenon. Qualitative social research grew out of and retains the human-centered and literary focus of the humanities. A qualitative study typically involves fieldwork, in which the researcher, rather than remaining in the office or laboratory, goes instead to the settings where the people being studied live, work, play, and so forth. The techniques employed range from passive, such as observation and gathering existing artifacts or records with minimal interaction with others, through active, including creating new documentation, interviewing people, and participating in their lives. Such methods have application in all of the social sciences, but have had a particular trajectory in the research conducted by political scientists.

Current Status of Qualitative Methods

Discussions of qualitative methods often divide the scientific world into two camps, those who accept a natural-science approach and insist on quantification as central to the development of positive facts and those who prefer a qualitative approach to human understanding of social phenomena. At one level, these two camps are imaginary, based on the stereotypes researchers assign to each other. At this level, the argument is often correctly characterized as superficial and unwarranted, and further discussion of qualitative methods could end here. The truth, however, is more complicated than that. If long-standing conflict is conceived as a symptom that tends to arise in the presence of larger philosophical dissension in the social sciences, then qualitative methods deserve closer examination. This is especially the case for political science, which has been unusually slow to find more than token accommodations for qualitative methods in its repertoire of research techniques.

New Knowledge in a Scholarly Venue

The leading edge of research for any discipline tends to appear in current papers from colloquia, symposia, and scholarly conferences. The American Political Science Association (APSA) produces the largest body of such research, with members concentrated in North America but also present in many other countries. A close examination of the electronic databases for the APSA annual meetings since 1999 (available online and through the

association) reveals several obvious qualities. The first is how difficult it is to find qualitative work. A casual browser of the conference program is unlikely to run across any. Electronic keyword searches, however, do yield results. Where social science research predominates in the APSA, papers with titles or abstracts containing qualitative terms cluster in just a few of the better established divisions (for example, the sections on presidency research, on race, ethnicity, and politics, and on foreign policy). Another quality of the work found through the APSA database is the wide variety of techniques reported in the scattering of papers scheduled for delivery, ranging from participant observation, ethnographic interviews, and other field research, to case studies, documentaries, and archival reports, to narrative and rhetorical analyses, to critical/cultural readings of texts. Finally, social scientists using these techniques may adopt in their reports an apologetic or defensive tone regarding their methods, acknowledging that any conclusions must be tentative and generalizations must be postponed. The reports imply that only other means (quantification and statistical analysis) can create positive knowledge.

The emergence of a new section of the APSA dedicated to qualitative methods may seem to present evidence of an advancing interest in the area, perhaps following the classic pattern of subdiscipline formation. However, the definition of "qualitative" in research papers sponsored by the APSA section bearing that name remains primarily quantitative in its background assumptions and in its foreground logic. It is elsewhere, such as in the relatively new APSA sections on comparative democracy and on human rights, that a humanistic epistemology of qualitative work seems not only accepted but also strongly encouraged. In sum, as an example of one of the leading edges of knowledge creation in political science, the recent APSA conferences include qualitative methods as a minor, although perhaps advancing, element.

Established "Facts" in Textbooks

The trailing edge of academic research is usually found in reference books and textbooks, where scholars summarize the widely accepted or little disputed facts of the discipline. Here the story of qualitative methods might be seen as a slow progress toward inclusion, if not acceptance. The seven-volume *Handbook of Political Science*, published in the mid-1970s by Addison-Wesley for a scholarly readership, did include a volume on case studies. But the volume on methods, *Strategies of Inquiry*, treats the stuff of qualitative work, such as manuscript and other archival material, as a source of statistical data, rather than as an appropriate starting point for other modes of understanding political life.

The progress of qualitative inquiry since the mid-1970s can be traced in textbooks for U.S. political science methods courses, where qualitative approaches have played a minor role. They were hardly present before the mid-1970s. First published in 1974, W. Phillips Shively's primer on methods designed for undergraduate students focuses entirely on quantitative methods, after introductory chapters on theory and research questions. Qualitative methods did get mentioned a decade later. The same author's edited textbook from the mid-1980s is a good example: of seven chapters illustrating political science research processes, five, including two labeled "field research," are quantitative. Of the two remaining studies, one concerns John Stuart Mills' political theory and the other uses historical archives.

By the mid-1990s, references to qualitative methods might remain absent from some textbooks, but became a chapter in, or an addendum to, other, new editions of university political science methods textbooks. For example, the third edition of *Political Science Research Methods* by Jean Johnson and Richard Joslyn recites the received history of political science origins, from the time when scholars in the humanities began documenting history and describing government, an approach that held sway until roughly the 1950s. By this account, modern political science then grew from behaviorism and came to rely on statistical methods (becoming "behavioralism," which, unlike behaviorism, attends to unseen attitudes and affinities to quantify political behavior). Such political science textbooks today imagine research following the same story line, with nominally "qualitative" methods as preliminary steps preparing the way for the arrival of quantitative work. Despite recent changes, the coverage of qualitative methods has remained marginal, at least in textbooks used to teach the discipline in the United States.

Qualitative Methods as "Traditional" Political Science

Qualitative methods in political science did develop from philosophy and history, which provided the earliest scholarly venues for studying politics. The methods and assumptions of these humanistic disciplines remain closely associated with qualitative methods. Political theorists continue to analyze documents, and theoretical approaches continue to be central not only to political philosophy but also to the study of institutional and legal settings for politics.

Document Analysis in History

Political historians continue to value archival methods as more than data mining. One way that historical work manages to hold its own is by paying close attention to

social statistics from other sources. Frances Fox Piven and Richard Cloward, in their 1971 book, *Regulating the Poor*, discuss critical history in the presence of data, citing the quantitative studies of others to make a larger historical argument of their own. Although documentary and archival techniques persist, along with purely theoretical work, textbooks describe such scholarship as "traditional" in political science, as a backward or static set of research practices. In its methods and assumptions, political science has been consistently qualitative primarily when verging on other disciplines, such as philosophy, law, history, and sociology (Theda Skocpol has been especially important in bridging between political science and the latter two, and Piven and Cloward bridged from social work and sociology). And even within such work, controversy has emerged about its scientific justification.

Comparative Case Studies

Another early genre of qualitative political science techniques, comparative case studies, developed along with historical and theoretical approaches and came into its own by the middle 20th century, dedicated primarily to comparing the governments either of different countries, or especially in the United States, of different states. Although debates went on between those who preferred a more theoretical political science and those who preferred a more professional, government- and citizen-focused political science, the central methodological assumption of both parties emphasized measurement, with natural science providing the ideal model. This quantitative current advanced among those conducting case studies by the 1970s, such as Arend Lijphart (and later Ada Finifter). Quantification spread so extensively that recent case-study-methods books focus on quantitative procedures rather than on qualitative processes or interpretation.

The 1970s Critical Movement

The social unrest emerging from the late 1960s through the 1970s had a deep and broad influence on other social sciences, including sociology, anthropology, and communication, but less so on political science. The antibehaviorist movement beginning in the late 1960s proposed a politically responsible science based on values and engaged in the life of society, and Marvin Surkin and Alan Wolfe documented the emergence from the movement of an APSA caucus that did critique the constricting effects of behavioral methods. The 1971 history of the poor by Piven and Cloward illustrates the relation of activism to scholarship of the period. But polemics aside, the early results of the movement, found in its journal, *Politics & Society*, were not primarily empirical. Instead, the journal published philosophical discussions, alternative policy studies, and the like, with the shared aim of providing thoughtful interventions from an engaged political science. Only later did the journal emphasize more empirical work, welcoming a wide mix of methods. In recent conference papers, the APSA section on new political science retained an interest in the social consequences and policy impacts of research, as against a commitment to any particular methods.

Social Scientific Techniques of Qualitative Inquiry

In the social sciences, interpretive and similar modes of qualitative inquiry depend principally on field research, which attempts to enter into group life self-reflexively. Political scientists conduct qualitative fieldwork using two main techniques, participant observation and unstructured interviewing.

Participant Observation

In a well-known example of participant observation, Richard Fenno for several years made trips with U.S. congressional representatives to their home districts and watched over their shoulders to understand how they related to their constituents. The resulting 1978 book, *Home Style*, contains information largely beyond the reach of other techniques—about the context and sequence (time and process), as well as the personal dimension, of politics. However, Fenno describes participant observation as exploratory, and he issues several caveats about the limits of its potential contribution to knowledge. Given such a tentative assertion of fact creation, it is not surprising that, despite the wide respect for Fenno and for his book, participant observation has remained at the margins. In the introduction to his collected essays more than a decade after *Home Style* came out, he laments that such work is not quantifiable and is "hard to discipline."

Interviewing

Another field technique, unstructured interviews, usually involves open-ended conversations documented either by tape recording and transcription or by note taking, followed by writing up field notes after each session. Scholars then do multiple, close readings of the texts to look for patterns, sometimes aided by qualitative analysis software, which allows the researcher to tag ideas, relate them to other places where they occur or to other ideas, and develop a map of group understanding. Because ideas are expressed in many ways, not relying on the specific terms or phrases used in quantitative content analysis,

such software is only an aid, like index cards. The scholar must still accomplish the qualitative tasks of exploring, thinking, and writing about a (usually extensive) body of text.

Early on, researchers such as Harriet Zuckerman used the unstructured interview in political science to focus on the study of elites. Doug McAdam, however, published *Freedom Summer*, his study of young people involved in the U.S. civil rights movement, and Bill Gamson's *Talking Politics* explores the experiences of citizens. McAdam began with a questionnaire and followed up with interviews, but other research reverses the process, beginning with qualitative interviews, and then taking a quantitative direction once the analysis begins. A typical study, such as one John Kornacki included in his anthology, uses interviews to discover how members of a legislative body develop patterns of informal leadership, but reports the results in tables that reduce the texts to a few key questions. (The reduction would typify a different technique, i.e., structured interviewing, which occupies a middle ground between surveys and interviews, something like a questionnaire administered face to face). The study also describes its results as general, with little self-reflection, and uses a footnote to mention its methods. The footnotes also reveal that the interviews were part of a larger, funded project based on a systematic (quantitative) survey. Unlike field techniques as practiced elsewhere in social inquiry, their use in political science often springs from or leads into this sort of data processing and analysis.

Philosophical Controversies

Impact of Critical Theory and Cultural Studies

Interpretive methods of empirical research have their philosophical underpinnings in schools of thought that reject the attitude of detachment and the operationalizing strategies that characterize quantitative social science of the 20th century. Critical theory, which developed in the 1940s among Max Horkheimer, Theodor Adorno, and others of the Frankfurt School, and cultural studies, which grew in the 1960s from the British literary and social analyses of Raymond Williams, Richard Hoggart, and the Birmingham School, have wielded a large influence over intellectual life of the past quarter century. These critical/cultural approaches, however, have made only some headway in political science. One anthology, for example, illustrates the difficulty that political scientists face when incorporating cultural approaches into the mainstream of the discipline. Compiled for scholars of comparative politics, a field the authors say has lost its bearings, the collection characterizes three competing traditions: structuralist, rationalist, and culturalist. The latter receives the least extensive treatment in the volume, and in the key theoretical chapter on culture, Marc Ross observes that "cultural contributions to political analysis are relatively rare and far less developed" and that "few graduate students take culture very seriously." Other chapters mention the cultural turn in political science, but the consensus is that such work is lacking rigor.

There are pockets in political science that foster overt constructivism, the notion that the key to understanding politics is the meanings people (groups, institutions, and governments) assign to events and other aspects of political life. Scholars in international relations, for example, have found that norms play an important part in political transformations around the world. An influential example of constructivist thinking is Alexander Wendt's 1999 book, *Social Theory of International Politics*, which attempts to span the divide between constructivist and positivist epistemologies by arguing, for example, that whether nations view each other as friends, rivals, or enemies determines the fundamental ways they act toward each other. Such work, however, is the exception. This state of affairs is due largely to the theoretical status quo in the discipline, especially in the United States. In the intellectual hierarchy, as encountered in the rankings of journals, departments, and other status-granting institutions of the discipline, pride of place has been reserved for a positive orientation. When Kornacki asked political experts to set an agenda for studying the U.S. Congress, for example, the specific ideas they came up with were mainly quantitative: surveys of the members, data on resources available to them, and similar information regarding personnel working for them. An interpretive orientation, despite long roots, has not created a major challenge to this way of thinking. The earlier protest movement within the discipline, despite its forceful introduction in the ferment of the early 1970s, and the critical and cultural orientations emerging in the 1980s have made little headway. Even in works of political history, a logical home to humanistic thinking, the current approach is marked by efforts to find metrics for temporal and sequential analysis.

The Quantitative Reaction and Ensuing Controversies

The discussion of qualitative methods within political science has continued despite the limited approaches. A key event of the 1990s was the publication of a textbook, *Designing Social Inquiry: Scientific Inference in Qualitative Research*. The authors, routinely reduced in reviews and course syllabi to their surname initials

KKV, propose that qualitative and quantitative methods are fundamentally alike. Any differences are stylistic only, because both aspire to do systematic science by following valid rules of causal and descriptive inference. The authors, Gary King, Robert Keohane, and Sidney Verba, then propose ways for qualitative researchers to apply quantitative tools to that end. Tested extensively before publication, the textbook circulated widely thereafter, going through several printings. From the time it appeared in Samizdat copies and then went into print, it has been viewed with admiration. For instance, partisans of rational choice theory, which builds historical explanations and political predictions of human action based on the notion that people maximize utility, point to KKV as a source of inspiration. The textbook also met with some antagonism, which appeared prominently in the bulletin *PS: Political Science & Politics*, and later played in the background of other controversies.

In 2000, anger against the exclusion of qualitative methods flared up online with the circulation of a letter from "Mr. Perestroika," representing what many assumed to be one or more younger political scientists or graduate students. The debate, which continued in the *Voices* column of the APSA bulletin from 2000 on, focused on the structure of the APSA, such as the presentation of a single slate of officers for election, as well as on the editorial practices of its flagship journal, the *American Political Science Review*. Although financed by dues from all APSA members, the journal favors statistical research designs, its detractors say, while ignoring not only their qualitative work but also the normative and practical world of politics. The critique was met with some openness in the APSA, which expanded the discussion through its publications and conferences. As part of these arguments, recent examinations of political science methods employed in journals and taught in graduate programs, which also appeared in the APSA bulletin, seemed to suggest that qualitative methods are losing ground. In the top journals, qualitative articles have been in steep decline since 1975 and are increasingly segregated from and rarely cited in quantitative studies. At the same time, qualitative methods courses play a very small role in doctoral curricula, according to a study by Peregrine Schwartz-Shea, who concludes that the message to graduate students is clear: political scientists value quantitative over other methods.

The controversies over qualitative research are not new. Gabriel Almond documented their emergence in the 1970s and earlier, tracing back to the origins of the discipline, when political science detached itself from the humanities and relocated to the social sciences during the early 20th century. These debates may be central to the disciplinary identity of political science, as scholars attempt to define themselves in contrast to their forebears. If this is so, then conflict over qualitative methods will likely continue to wax and wane as it has done in the past. Whether or not qualitative approaches to knowledge have advanced in political science during the past quarter century, they remain at the margins. This appears to be due to the positive philosophical orientation of the discipline. Although they have provided a staging ground for complaints against that orientation in intermittent debates, qualitative methods of research, analysis, and writing, in current practice, serve principally as a tool for exploratory work and as a foil to police the boundaries of mainstream political science.

See Also the Following Articles

Case Study • Ethnography • Field Experimentation • Interviews • Participant Observation

Further Reading

Almond, G. A. (1990). *A Discipline Divided: Schools and Sects in Political Science.* Sage, Newbury Park, CA.

American Political Science Review. (1995). The qualitative–quantitative disputation. *Am. Politic. Sci. Rev.* **89,** 454–481.

Burgess, R. G. (1982). *Field Research: A Sourcebook and Field Manual.* Allen & Unwin, London.

Fenno, R. F., Jr. (1978). *Home Style: House Members in Their Districts.* Little, Brown, Boston, MA.

Fenno, R. F., Jr. (1990). *Watching Politicians: Essays on Participant Observation.* Institute of Governmental Studies Press, Berkeley, CA.

Finifter, A. (ed.) (1993). *Political Science: The State of the Discipline II.* APSA, Washington, D.C.

Gamson, W. (1992). *Talking Politics.* Cambridge University Press, Cambridge.

Johnson, J. B., and Joslyn, R. A. (1995). *Political Science Research Methods.* 3d Ed. Congressional Quarterly Press, Washington, D.C.

Kornacki, J. J. (ed.) (1990). *Leading Congress: New Styles, New Strategies*. CQ Press, Washington, D.C.

Lijphart, A. (1971). Comparative politics and the comparative method. *Am. Politic. Sci. Rev.* **65,** 682–693.

McAdam, D. (1988). *Freedom Summer.* Oxford University Press, New York.

Piven, F. F., and Cloward, R. A. (1993). *Regulating the Poor: The Functions of Public Welfare.* Vintage, New York.

Political Science & Politics (2002). Shaking things up? Thoughts about the future of political science. Symposium. *Politic. Sci. Pol.* **35,** 181–205.

Political Science & Politics (2003). Methodological pluralism in journals and graduate education? Commentaries on new evidence. Symposium. *Politic. Sci. Pol.* **36,** 371–399.

Ross, M. H. (1997). Culture and identity in comparative political analysis. In *Comparative Politics: Rationality, Culture, and Structure* (M. Lichbach and A. Zuckerman, eds.), pp. 42–80. Cambridge University Press, New York.

Rubin, H. J., and Rubin, I. S. (1995). *Qualitative Interviewing: The Art of Hearing Data.* Sage, Thousand Oaks, CA.

Surkin, M., and Wolfe, A. (eds.) (1970). *An End to Political Science: The Caucus Papers*. Basic Books, New York.

Waldo, D. (1975). Political science: Tradition, discipline, profession, science, enterprise. In *Handbook of Political Science* (F. I. Greenstein and N. W. Polsby, eds.), pp. 1–130. Addison-Wesley, Reading, MA.

Wendt, A. (1999). *Social Theory of International Politics.* Cambridge University Press, Cambridge.

Zuckerman, H. (1972). Interviewing an ultra-elite. *Public Opin. Q.* **36,** 159–175.

Qualitative Analysis, Sociology

Dorothy Pawluch
McMaster University Hamilton, Ontario, Canada

Glossary

analytic induction A procedure for verifying theories and propositions based on the assumption that the research should formulate explanations that apply to all cases of the phenomenon under study.
ethnography The process of intensively studying a social group by immersing oneself in the day-to-day lives of people in the group; also the product or outcome of such research.
naturalistic inquiry An approach to empirical research that emphasizes the need to study social actors *in situ*—that is, in their natural environment rather than in settings that are manipulated, contrived, or artificially arranged.
sympathetic introspection A term coined by Charles Horton Cooley to describe a methodology for gaining access to the meanings and interpretations of those social actors one is studying.
theoretical saturation The point in a qualitative study at which the researcher is learning from an examination of new cases nothing new that will add to the understanding of a particular phenomenon.
triangulation A technique for checking different types of data (e.g., field notes, interview transcripts, and documentary evidence) against each other in order to assess or refine a particular interpretation or inference drawn from the evidence.

Qualitative analysis in sociology is rooted in a particular view of the social world and how best to study it. The goal in qualitative analysis is not to count or measure but to capture the subjective meanings of situations that undergird human group life and to understand the social processes by which those meanings are constructed.

The Interpretive Tradition in Sociology

The Chicago School

Qualitative analysis in sociology can be traced back to the Chicago school of the 1920s and 1930s. Influenced by the sociology of Max Weber and Georg Simmel, a number of sociologists at the University of Chicago—including W. I. Thomas, Robert Park, and Ernest Burgess—encouraged students to learn about social life by going out into their own city (the field) and observing it. Chicago became a laboratory for the direct observation of social phenomenon. A series of classic ethnographic studies came out of the Chicago school, including W. I. Thomas and Florian Znaniecki's "The Polish Peasant in Europe and America," Louis Wirths's "The Ghetto," Harvey Zorbaugh's "The Gold Coast and the Slum," Clifford Shaw's "The Jack Roller," Paul Cressey's "The Taxi-Dance Hall," and Nels Anderson's "The Hobo."

Symbolic Interactionism

The methodological tradition of ethnographic fieldwork that the first generation of Chicago school sociologists established and that a second generation, notably Everett Hughes, continued was buttressed theoretically through the 1940s and 1950s by the work of Herbert Blumer. Most of the key elements in Blumer's writing originated in the ideas of George Herbert Mead, a University of Chicago pragmatist philosopher with whom Blumer and others in the Chicago school studied. However, it was Blumer who took Meadian notions about the relationship between the individual and society and fashioned them into a clear theoretical statement (symbolic interactionism) that continues to inform much of the qualitative analysis conducted today. The three basic premises of symbolic

interactionism, as Blumer laid them out, are as follows: (i) Human beings act toward things (objects, situations, people, and themselves) on the basis of the meanings that these things have for them, (ii) the meaning of things arises out of interaction, and (iii) the meanings of things are handled and modified through a process of interpretation that individuals engage in as they deal with the things they encounter.

Apart from his articulation of symbolic interactionism, Blumer was pivotal in laying out a methodological approach that was consistent with symbolic interactionist premises, an approach that respected the processual nature of human group life and treated human beings as active, interpreting, and interacting agents. Arguing that to try to catch the interpretive process by remaining aloof as an "objective" observer and refusing to take the role of the acting unit is to risk the worst kind of subjectivism, Blumer urged naturalistic observation of the empirical world. The object of social research, as he saw it, was to get close to the social world and dig deeply into it, achieving an intimate familiarity with the perspective of social actors.

Theoretical Divergence and Convergence

Since the 1960s, a number of other interpretive perspectives have emerged, including phenomenology, ethnomethodology, dramaturgy, and, recently, social constructionism, postmodernism, poststructuralism, feminism, and critical and standpoint theory. These perspectives differ over a range of ontological, epistemological, political, and ethical issues, perhaps most fundamentally over the degree to which there is even an external "objective reality" accessible to social researchers and, relatedly, what the accounts that researchers produce actually represent. There is a view that interpretive theories have become so eclectic and the differences between them so profound that the qualitative analysis they tend to favor has become unmoored from the distinct theoretical assumptions to which it was once so firmly linked. In this view, qualitative analysis has become nothing more than a rallying call for a group of increasingly disparate scholars who find themselves under the interpretive umbrella.

There is another view, however, that despite the diverse forms that interpretive theories now take and the profound differences between them, at their core they continue to share a common concern with human action, the construction of meaning and the agency of social actors. All share an interest in knowledge about the social world from the inside. Some sociologists, such as David Maines, have argued that the discipline of sociology more generally has moved toward an integration of perspectives, with the premises of symbolic interactionism and its preference for a more naturalistic, qualitative methodology at the center of the convergence. The importance of situations and context for understanding human conduct, the necessity to consider the meanings or definitions that social actors attach to things and situations around them, and the value in understanding social structures as an expression of social processes are being recognized in virtually every area of sociology, although they may not be recognized as distinctly symbolic interactionist insights. This may explain why, since the 1980s, there has been a significant shift within the discipline toward greater use of qualitative analysis. This shift has paralleled trends in other social sciences, the humanities, and fields such as education, nursing, public health, medicine, marketing, and even accounting. A "qualitative revolution" is unfolding.

The Research Process

Analytic Induction

Qualitative analysis in sociology, then, cannot be understood apart from the fundamental assumption that the social world is constituted by the meaning-making practices of social actors. Nor can it be understood apart from the way interpretive sociologists conceive of the research process. In contrast to positivism, which is based on a deductive logic of collecting data to assess preconceived models, hypotheses, or theories, qualitative analysis proceeds largely on the basis of analytic induction. Using inductive reasoning, explanations for social phenomena arise from the data rather than from preconceived categories that force the empirical social world into the operational definitions that researchers construct. Rather than formulating theories that are tested against the social world, qualitative researchers use a "grounded theory" approach, relying on their observations of the social world to generate theory. The theories they construct are consistent with what they see.

Emergent Design

Moreover, in the grounded theory approach, there is a more or less constant interplay between planning, data gathering, analysis, and even writing. These activities are seen not as discrete and sequential steps but as related and intertwined aspects of the research act. The analysis or process of searching for themes or answers to research questions begins as soon as the researcher begins the study. Decisions are made about a primary focus and a series of generalizations are formulated to describe what is happening, generalizations that are then taken into the next observation, interview, or document

examined. Indeed, in a process referred to as theoretical sampling, the researcher may seek out additional cases on the basis of the potential for those cases to test, refine, or extend the generalizations. When cases that do not fit are encountered, either generalizations are revised so that they do or the phenomenon is redefined so as to exclude those cases that do not fit the generalization. This process continues until a point of theoretical saturation is reached. Theoretical saturation is the point at which additional cases no longer contribute to learning anything new about the phenomenon in question. In the end, the process of analytic induction yields a proposition or statement that applies to all the cases examined. Howard Becker describes the process as being in continuous dialogue with the data. The research design remains emergent, fluid, and flexible throughout.

Generic Social Processes

The same constant comparative method that produces grounded generalizations about a particular group or social setting can be extended to produce theory at higher levels of abstraction. Qualitative researchers can make conceptual comparisons across substantive contexts to produce more formal theory. For example, involvement in different subcultures—religious, criminal, professional, and sport—may be compared in a way that allows one to theorize more generally about involvement in any subculture. Prus has argued for the need to use the rich, textured data that qualitative research yields to formulate concepts and to describe processes that transcend the particular settings in which the data were gathered. Generic social processes may be delineated that have transsituational and cross-contextual relevance, tying together a great deal of research that would otherwise remain disconnected or scattered across a range of substantive contexts. In this way, insight is gleaned into the most foundational question of sociology—how human group life is accomplished.

Types of Data

An understanding of qualitative analysis requires an appreciation for the kinds of data that sociologists work with and how they collect these data. Since the goal in qualitative analysis is to grasp as faithfully as possible the lived experience of social actors and the processes by which they construct meanings, there is a preference for data-gathering techniques that bridge the gap between the analyst and the empirical world.

Participant Observation

Central among these techniques has been participant observation, which involves direct observation of ongoing group life. Participant observation, or ethnography, encompasses a broad spectrum of possible roles for the researcher, from passive observer to active participant. Most sociologists strive for a level of contact and involvement that allows them to achieve an intimate familiarity with, or sympathetic understanding of, the group they are studying. Usually this involves not only watching but also talking to the people one is studying either through casual conversation or in the context of a more formal interview. Participant observation is sometimes described as "hanging around." Although accurate on one level, the phrase belies the range of often thorny issues that need to be negotiated—gaining access to the site or group, building trust with its members, ascertaining on an ongoing basis how to present oneself, developing and maintaining relationships with informants, dealing with attachments and emotions, confronting ethical dilemmas, and deciding when it is time to leave the field and how to do so. Throughout the process, the researcher is engaged in continuous stock taking, trying to figure out what is going on, eventually focusing on a particular aspect of what is happening and making decisions about what to do next.

Interviews

In-depth interviews are another common way to collect qualitative data. Interviews may be standardized or structured so that the same list of open- and close-ended questions is posed to each respondent. When the researcher is really interested in the perspective of respondents, however, interviews are more likely to be either semistandardized or unstandardized and informal. In the case of semistandardized interviews, there may be an interview guide (a list of topics to address and possible wording and follow-up questions), but researchers give themselves the latitude to digress as they see fit and to probe into areas that appear to be analytically promising. Unstructured interviews tend to be the most conversational in style. The researcher may have only the roughest idea of where the interview will go, allowing the talk of respondents to drive the course the conversation takes. This requires adept listening skills and a preparedness to improvise and generate questions on the spot as the situation dictates. Apart from interviewing styles, there are decisions to make about who to interview; how to reach them; how to establish rapport; how to word, order, and pose questions; how much personal information, if any, to divulge; whether to play the role of a naive or informed listener; how to record what is being said (tape or notes); and when to stop. If the interview has touched on sensitive subjects, there are concerns about what happens to respondents when the researcher has walked away. Many of these decisions are guided by the researcher's ongoing analysis of the data.

Documents and Other Data-Gathering Techniques

Another technique involves analysis of personal, public, or historical documents—diaries, letters, biographies, autobiographies, photographs, newspapers, archival material, organizational records, actuarial records, official reports, manuscripts, contracts, court proceedings, films, TV programs, graffiti, song lyrics, Web sites, Internet chat rooms, and so on. These sources of data have the advantage of being unobtrusive. They cannot interrupt the flow of interaction and there are fewer concerns about how the researcher's presence may be affecting interaction. During the past two decades, qualitative sociologists have developed yet other techniques, including focus group interviews, life histories (using one person's first-hand account of his or her life), the study of material artifacts, and autoethnographies (mining one's own thoughts, experiences, and biography as a way of shedding light on the lives of others).

Triangulation

In many instances, qualitative researchers work with not one but several sources of data within the context of the same study. In a research tradition where the focus is on the point of view of the experiencing actor, anything that allows for greater sympathetic introspection is seen as potentially useful. Moreover, a multimethod approach offers different lines of sight on the same social situation or process, thereby giving analysts an opportunity to triangulate. Interpretations and generalizations that emerge out of the analysis of one source of data can be checked against other sources, leading not only to richer accounts but also to accounts in which the analyst can have greater confidence.

Interpreting Qualitative Data

Just as there are different theoretical perspectives guiding qualitative analysis in sociology, and different ways to collect data, there are different approaches to the handling, sorting, and interpretation of data. There is no clear set of rules that all qualitative sociologists follow. There are, however, certain practices that are commonly employed.

Memo Writing and Coding

Memos are essentially notes that researchers write to themselves as they are conducting a study. These notes, sometimes kept in the form of a research log, are typically a combination of procedural record, explaining the course that a study takes and the decisions made, and an analytical record consisting of reflections, ideas, impressions, reactions, and insights. Coding, is an analytical strategy aimed at sorting and organizing the data according to key themes, patterns, and concepts.

Open Coding

Coding can occur on different levels and take different forms. Typically, in the earliest phases of the research, codes are descriptive. Working line by line with the data, the researcher identifies the topics covered. What is the data about? The codes are not predetermined but based directly on what appears in the data. In fact, the codes may be *in vivo*, codes that use the very language of those studied. For example, in a study among people living with HIV, the talk in a line of transcribed data might have to do with telling a friend about one's seropositive status. The margin code may read "disclosure to friends" or simply "telling friends." This type of preliminary and unrestricted coding, referred to as open coding, invariably triggers ideas. Analytical observations that appear in the memos at this point are likely to take the form of tentative hunches about commonalities, recurrent themes, possible patterns, and what they may mean. These hunches, however, as they are considered in light of a rereading of already collected and incoming data become the basis for more conceptual, interpretive, or focused coding.

Focused Coding

Focused coding is more directed. Researchers concentrate on those codes with overriding significance, those that seem most interesting and show the most analytical promise, or those that work best at categorizing large portions of the data accurately and completely. They begin to theorize around them. Moving beyond a description of the data, the point now is to go through the data in a more selective way, searching for instances of a particular phenomenon; identifying its properties; searching for variations within it that may lead to the construction of a typology; specifying conditions under which it arises, continues to occur, and changes; considering its consequences; and exploring its relationship to other phenomena.

Returning to our example, a researcher might observe as a result of open coding that there are repeated references in the data to disclosing to others. The observation may lead to a more focused consideration of disclosure. What types of disclosure are individuals talking about—to partners, children, employers, coworkers, or neighbors? In relation to whom and in what circumstances does disclosure become an issue? What are the issues related to disclosure? How are disclosures done? To what extent does how they are done depend on whom one is telling? What are the consequences of telling or not telling? To what extent is disclosure linked to stigma or problem-solving strategies that people with HIV use?

It is this type of coding that often raises questions, reveals gaps in the data, and sends the analyst back into the field to do more theoretical sampling. Memo writing also begins to look different at this stage in an analysis, becoming increasingly more conceptual and taking on the quality of theoretical notes. These notes serve as a bridge from the coded data to the first draft of a completed analysis.

The Importance of Staying Close to the Data

There are variations among qualitative researchers in terms of how strictly they follow these strategies. There are variations in how far from the substantive case an analysis ventures and how theoretical generalizations become. There are also variations in terms of temporal sequencing: Some researchers adopt an analytical stance almost as soon as they have data to consider, whereas others orient themselves more toward analysis after most of the data have been obtained. Qualitative researchers have been criticized for not standardizing and codifying their analytic procedures. Some have responded by calling for greater articulation of the processes involved in the analysis of qualitative date as a first step toward codification. Others see codification as neither feasible nor desirable. In whatever way analysts choose to proceed in coding and interpreting their data, an intimate familiarity with the data is essential. If there is a cardinal rule in the analysis of qualitative data, it is "stay close and be true to the data." Although they are generally working with a huge volume of data in the form of hundreds if not thousands of pages of field notes, interview transcripts, and/or documents, by the time a study has been completed most researchers will have gone through the data several times. The data are read and reread, coded and recoded as the analyst moves with greater certainty, clarity, and depth toward the insights that will ultimately find expression in reports and published papers.

The Use of Computers

The task of analyzing qualitative data has been facilitated by the development of computer software programs. Some programs are generic word processors, text retrievers, or textbase managers, whereas others are designed specifically for qualitative data analysis. Many of the dedicated programs can perform sophisticated functions, such as coding the data automatically, searching for relationships among code categories, and creating hierarchical and graphic networks of codes, allowing for the automatic generation and testing of hypotheses.

The number of dedicated programs available is increasing rapidly, but sociologists who work with qualitative data have tended not to embrace them. There are concerns that the programs encourage simple counting and matching of code categories. Qualitative analysis would become more of a "word-crunching" exercise consistent with the logic of survey research as opposed to the in-depth, theme-based analysis it has always been. Although the programs may contribute to the greater standardization of procedures that some in the field would like to see, they are viewed by others as a threat to the long-standing tradition of imaginative theoretical work and intellectual craftsmanship on which qualitative researchers have prided themselves. Worse, there are fears that the demands of particular programs and what the programs can or cannot do may ultimately affect how a research project is designed and the types of questions that are asked. Although there are still qualitative researchers who do not use any sort of computer program and work largely by hand, there is no question but that computers will play an increasingly important role in qualitative analysis in the future. The hope of most qualitative analysts is that computer programs will enhance rather than control or compromise their studies, and that they will ease the more tedious aspects of working with qualitative data so that researchers can devote more time to interpretation.

Conclusion

rom the vantage point of someone who has never done it, qualitative analysis may appear easy. It is not. Both the data collection and analysis dimensions of the process can be labor-intensive and time-consuming. There may be innumerable challenges and frustrations. The lack of simple procedural recipes requires a capacity to tolerate ambiguity and the patience to hone fine interpretive skills. In a passage that refers specifically to fieldwork but could easily be adapted to fit other aspects of qualitative analysis, Shaffir *et al*. (1980) write,

> *[It] must certainly rank with the more disagreeable activities that humanity has fashioned for itself. It is usually inconvenient, to say the least, sometimes physically uncomfortable, frequently embarrassing, and, to a degree, always tense. Sociologists and anthropologists, among others in the social sciences, have voluntarily immersed themselves for the sake of research in situations that all but a tiny minority of humanity goes to great lengths to avoid.* (p. 3)

It is not unreasonable to wonder why a researcher would ever undertake qualitative analysis. On this point, if no other, there is consensus. Qualitative analysis is worth doing because its intellectual payoffs and personal rewards are great. Coming full circle to the essential assumptions underlying most qualitative analyses, if "society" is about individuals intersubjectively creating the

meanings that constitute the context for all human action, and if our purpose as researchers is to study society, there is simply no other way to do it.

See Also the Following Articles

Coding Variables • Computer Simulation • Computer-Based Mapping • Computer-Based Testing • Computerized Adaptive Testing • Computerized Record Linkage and Statistical Matching • Data Collection, Primary vs. Secondary • Ethnography • Innovative Computerized Test Items • Interviews

Further Reading

Becker, H. S. (1998). *Tricks of the Trade: How to Think about Your Research While You're Doing It.* University of Chicago Press, Chicago.

Berg, B. L. (2004). *Qualitative Research Methods for the Social Sciences.* 5th Ed. Allyn & Bacon, Boston.

Bryman, A., and Burgess, R. G. (eds.) (1999). *Qualitative Research*, Vols. 1–4. Sage, London.

Corbin, J. M., and Strauss, A. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory.* 2nd Ed. Sage, Thousand Oaks, CA.

Denzin, N., and Lincoln, Y. (eds.) (2000). *Handbook of Qualitative Research*. 2nd Ed. Sage, Thousand Oaks, CA.

Esterberg, K. G. (2002). *Qualitative Methods in Social Research.* McGraw-Hill, Boston.

Fine, G. A. (1993). Ten lies of ethnography: Moral dilemmas of field research. *J. Contemp. Ethnogr.* **22,** 267–294.

Glaser, B. G., and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine, Chicago.

Gubrium, J. F., and Holstein, J. A. (eds.) (2002). *Handbook of Interview Research: Context and Method*. Sage, Thousand Oaks, CA.

Hesse-Biber, S. N., and Leavy, P. (eds.) (2004). *Approaches to Qualitative Research: A Reader on Theory and Practice*. Oxford University Press, New York.

Prus, R. (1996). *Symbolic Interaction and Ethnographic Research: Intersubjectivity and the Study of Human Lived Experience.* State University of New York Press, Albany.

Schwandt, T. A. (2001). *Dictionary of Qualitative Inquiry.* 2nd Ed. Sage, London.

Shaffir, W. B., Stebbins, R. A., and Turowetz, A. (1980). *Fieldwork Experience: Qualitative Approaches to Social Research.* St. Martin's, New York.

Silverman, D. (2001). *Interpreting Qualitative Data: Methods for Analysing Talk, Text and Interaction.* 2nd Ed. Sage, London.

Taylor, S. J., and Bogdan, R. (1998). *Introduction to Qualitative Research Methods: A Guidebook and Resource.* 3rd Ed. Wiley, New York.

Quantitative Analysis, Anthropology

Dwight W. Read
University of California at Los Angeles, Los Angeles, California, USA

Glossary

clustering procedure An algorithm for grouping data points in such a way that the groups are internally cohesive and externally isolated from one another.
correlation coefficient A measure of the extent to which the values of two variables tend to covary.
descriptive statistics Methods for describing and summarizing patterns through parameter values based on measurements made over all members of a population.
histogram A graph in which a bar represents a value for a variable and the length of the bar is proportional to the frequency of that value in the population.
inferential statistics Methods for relating measurements made over samples to population parameters.
normal distribution A frequently occurring data pattern important for inferential statistics.
population mean, μ The value obtained by averaging all the population values.
population parameter A numerical value based on measurements made on all members of a population.
population standard deviation, σ The square root of the average of the square of the deviation for each measurement in the population and the population mean, μ.
scattergram plot A graph of two variables with each member of a population located at the point determined by variable values for the members of the population.

Quantitative analysis in anthropology has had an uneasy relationship with anthropological research stemming from a fundamental conflict between the underlying assumption of cultural as shared knowledge among culture bearers and statistical methods based on the assumption that patterning is not displayed on individual cases but in the aggregate. Culture as shared knowledge implies that data on cultural systems can be obtained from a few knowledgeable informants, whereas the presumption of patterning in the aggregate assumes that pattering is discerned through the relationships of individual cases to one another and not on single cases. Ethnographies are implicitly based on the assumption that discerning patterning at a cultural level does not require statistical sampling. At the same time, although there are many questions relating to behavior where patterning in the aggregate is the appropriate metaphor, it is often difficult to determine in advance what should be considered as the aggregate. This discordance between the underlying assumptions of anthropological research and statistical methodology implies that effective use of statistical methodology depends on determining ways in which anthropological reasoning can be expanded and enriched through the use of quantitative methods. At the level of the formation of a data set to be brought forward for analysis, this has led to preanalysis of data so as to subdivide an initially heterogeneous data set into data sets homogeneous in terms of both cases included in the data set and the variables measured over those cases. At the level of analysis, methods such as cultural consensus analysis have been formulated in accordance with assumptions about culture as shared knowledge. Anthropological research has made extensive use of statistical methods ranging from simple, single-variable descriptive methods to sophisticated, multivariate statistical modeling.

Statistical Methods and Anthropology: An Uneasy Relationship

As a discipline, anthropology, with the notable exception of biological anthropology and to a lesser extent archeology, has long had an uneasy relationship with

quantitative methods and analysis. The uneasy relationship stems, in part, from a presumed incompatibility between the full richness of human experience as it might be captured and expressed qualitatively in an ethnography and the supposed loss of that richness when behavior is reduced to summary, quantitative measures. Whether or not the uneasiness is valid, the trend in cultural anthropology toward a more humanistic, interpretive, and literary approach has substantially reduced the usage of quantitative methods in anthropology as measured by the percentage of articles that make use of any kind of quantitative analysis or reporting of quantitative data. A study of six of the major journals of anthropology by Michael Chibnik (1999) concludes, "The gap between quantifiers and nonquantifiers in sociocultural anthropology may be widening. There was a striking drop in the use of simple descriptive statistics from 1985–1986 to 1995–1996" (p. 155). Yet anthropological research has a long tradition of both using and developing quantitative methods of analysis and continues to use and develop quantitative methods that are part of statistical methodology.

Broadly speaking, the utility of statistical methods for anthropological research arises from the methods providing a way to represent and analyze patterns in phenomena through quantitative measurements. As a discipline, statistics is concerned with numerically expressed data that are individually idiosyncratic but display patterning in the aggregate. This core concept of patterning in the aggregate identifies and brings together two key methodological features central to the application of statistical methods in general and to anthropology in particular: (i) the use of quantitative measures as the basis for the display and discernment of pattern in phenomena and (ii) the use of an aggregate, or a population, rather than individual cases, as the reference point for discerning and expressing patterning found in measurements made over phenomena.

Of these two features, the first raises the issue of implementation (What are appropriate quantitative measures in anthropology, especially with regard to culture?), and the second raises a conceptual issue relating to a discrepancy between the way cultural patterning is presumed to be distributed with respect to culture bearers versus the statistical concern with patterning found in an aggregate but not on individual cases. Together, these two issues highlight the unease that many anthropologists have found with quantitative methods and analysis.

In fact, there is legitimate concern regarding the fit between the concepts underlying statistical methods and the understanding anthropologists have about the nature of culture and its distribution among societal members. The unease many anthropologists have had with quantitative analysis is not due solely to a shift toward a humanistic and interpretive approach. In part, it stems from a conceptual discordance between the underpinnings of statistical methods and the assumptions anthropologists make about how phenomena of interest to anthropologists are structured. Consequently, this article focuses primarily on the relationship between the conceptual foundation of statistical methods and the conceptual foundation that anthropologists bring forward in their analysis of human societies and human culture. The range of quantitative forms of analysis that have been implemented in anthropology will be discussed briefly.

The Basic Dilemma: Patterning in the Aggregate versus Patterning on Individual Cases

The difference between patterning found in the aggregate and patterning expressed on individual cases lies in the relationship of individual cases to the pattern. By patterning in the aggregate is meant patterning in which individual cases may deviate from the overall pattern that is discerned when the aggregate is considered as a whole. The patterning is defined by the relative position of each case to the other cases included in the aggregate; hence, the patterning is aggregate specific. Consequently, for patterns based on aggregated data, generalizations made from individual cases may differ widely, depending on the individual case that has been selected for generalization. In contrast, patterning expressed through properties of individual cases allows for patterning exhibited on an individual case to be generalized directly to other, like cases. This kind of patterning is often found in the physical world. An experimenter might drop an object with specified mass in a vacuum, for example, and graph the relationship between time and distance the object has traversed due to the effect of gravity acting on the mass. The pattern displayed in the graph of time versus distance is the same, keeping the experimental conditions (such as dropping the objects in a vacuum) unchanged and so the patterning found in a single case of a falling object characterizes the patterning found in other instances of falling objects as well. No aggregate or population needs to be identified to discern the pattern, and so no sampling protocol is required because there is no population to be sampled.

The focus on an aggregate, or population, in statistical methodology as a reference point for discerning patterning is not in keeping with assumptions made in cultural anthropology, however. The data elicitation method of participant/observation that underlines much of ethnographic fieldwork is based on a different notion regarding the locus of patterning for the phenomena of interest.

Discordance between Ethnographic Observation and Statistical Methodology

Discordance between ethnographic observation and statistical methodology stems from assumptions made in cultural anthropology about the distribution of cultural, and to a lesser extent behavioral, patterns across the members of a society. Cultural patterns are presumed to be shared by the members from the same cultural domain; hence, discernment of these patterns does not require the "patterning in the aggregate" feature that underlies statistical methods. Rather than requiring an aggregate to discern patterning, it is assumed that patterning obtained from an individual case—the informant—can be generalized to other cases as noted by W. Penn Handwerker and Danielle Wozniak (1997): "The *socially constructed* nature of cultural phenomena makes the classical sampling criterion of independent case selection not only impossible to attain but also undesirable" (p. 874, italics added).

The notion of patterning found on an individual case representing the patterning to be found on other cases has been the basis of traditional ethnographic research with its focus on the use of a few, knowledgeable informants. Since the cultural pattern obtained through the use of a single informant is assumed to be essentially the same regardless of the informant, the criterion for a "good informant" is someone knowledgeable about one's culture. Discussions in the literature that focus on the choice of informants pay attention to knowledge of one's culture and individual characteristics that make a person a good informant, not on statistical sampling procedures for selecting informants. Even when sampling procedures are used, it is the knowledge of informants and not the use of random samples that is paramount. As observed by Robert Trotter and Jean Schensul (1998), "It is important to talk to individuals who are carefully selected for their expertise . . . rather than randomly selecting someone from the general population" (p. 703). In effect, Durkheim's notion of culture as a collective representation presumes that individual variation in cultural knowledge relates primarily to the extent to which one is knowledgeable about one's culture and not to the pattern of culture per se. Even statistical techniques that have been developed for discerning cultural patterning, such as cultural consensus analysis, operate from the presumption that what distinguishes cultural knowledge is a high degree of consistency from one individual to another, thereby making variation in knowledge from one individual to another of secondary importance.

The discordance between the conceptual basis for statistical methodology and the presumption of shared knowledge underlying cultural phenomena, however, has received relatively little attention in the application of quantitative methods in sociocultural anthropology. The emphasis, instead, has mainly been on methods for discerning pattern in numerical data. Within this framework, considerable work has been done on issues that arise from the viewpoint of statistical methodology, such as Galton's problem with drawing a representative sample of world societies from the population of all societies since societies in a region may be historically related and hence are not independent observations from a statistical viewpoint. In archaeology, however, more attention has been placed on the discordance between the conceptual foundations of statistical methodology and the concepts anthropologists have about the nature of human societies.

Use of Statistical Methods to Extend Anthropological and Archaeological Arguments

In either case, anthropology or archaeology, we may consider the underlying rational for the use of numerical data and statistical methods to be a way to extend and increase the scope of anthropological and archaeological arguments regarding the nature and form of societies and their underlying culture, including the conditions under which societal and cultural change takes place. Use of statistical methods is not a goal in and of itself; rather, the goal is to determine ways in which anthropological reasoning can be expanded and enriched through the use of quantitative methods. Nor are quantitative methods considered to be inherently preferable to more qualitatively based methods (or vice versa). Instead, the choice of method arises from the questions being asked and the form of the data that can be brought forward to bear on those questions.

In ordinary discourse, we make use of both qualitative and quantitative concepts. Consider a statement such as "If a sexually mature female engages in sexual intercourse, she may become pregnant and give birth in 9 months to a baby weighing about 9 pounds." Both quality and quantities are integrated into our commonsense understanding of matters such as reproduction. The relative importance of the qualitative versus the quantitative aspects of this statement depends on the context. For someone talking to teenagers about the risk of unprotected sex, the emphasis would likely be on the qualitative aspects of reproduction: "sexually mature female," "become pregnant," and "give birth." On the other hand, a doctor talking with a teenager who has just become pregnant is likely to focus on more quantitative aspects, such as "give birth in 9 months" and "baby weighing about 9 pounds." Similarly, quantitative anthropology depends on a context in which the questions of interest can be best addressed by quantitative

measures. What constitutes quantitative measures, though, is not self-evident and subject to change as the degree of understanding by anthropologists of a research domain increases.

Measurement Scales

Initial qualitative distinctions made among observations, such as classification of societies into hunting/gathering, horticultural, agricultural, and industrial on the basis of the primary means of production and/or procurement of resources, may later be augmented by an ordinal dimension. We might, for example, rank the kinds of society on the basis of increasing complexity—hunting/gatherer < horticultural < agricultural < industrial—according to some definition of complexity. The indeterminacy of the latter brings to the fore a complex issue, namely the construction of measurement scales. For a measure such as complexity, the underlying difficulty lies in the fact that we typically measure behavior or the consequences of behavior, such as the number of interconnected parts, but the matter of concern may be at more of an ideational level of how a society is conceptually structured rather than at the material level of how a society is organized. The two viewpoints—ideational versus material—have different implications for the application of statistical methods because the former is concerned more with patterning in the form of shared knowledge, hence patterning generalizable from the pattern found on an individual case (such as a well-informed informant), and the later relates to patterning found in the aggregate, hence on patterning falling within the domain of statistical methods.

More problematic than an ordinal scale measurement in anthropological research is the formation of ratio scale measurements. Even when a satisfactory definition of a ratio scale measurement has been made, data collected by others may not permit assigning a ratio scale measurement. Carol Ember and Melvin Ember (1998) illustrate the problems that may arise through an example based on devising and implementing a measurement scale using the concept of an extended family. They suggest the operational definition that an extended family consists of "two or more constituent families united by a blood tie," where a family is a "social and economic unit consisting minimally of at least one or more parents and children" (p. 665). Although the definition provides the basis for determining the instances of social units that are to be considered as a family or as an extended family, it implicitly raises the question as to whether or not the definition should be an emic or an etic construction. Like other, similar operational definitions, the definition suggested by them is an etic one because it is presumed to be universally applicable. In contrast, Ira Buchler and Henry Selby (1968) had previously suggested that "the shape of the family exists *solely in the minds of the informants*, and the ethnographer *merely translates* the results of his rigorous and culture-specific methodology to establish what the family form of X culture may be" (p. 22, italics added). Their comment is echoed in David Schneider's (1984) assertion that "[t]he first task of anthropology, prerequisite to all others, is to understand and formulate the symbols and meanings and their configuration that a particular culture consists of" (p. 196). In brief, even in what appears to be a simple task of operationally defining a quantitative measure, fundamental theoretical issues regarding the nature and form of anthropological research quickly arise and present issues that do not have easy resolution within the domain of quantitative research and statistical methods. The latter presumes that adequate quantitative measures have already been defined or determined because statistical methods do not provide the means for determining or defining satisfactory quantitative measurements.

Even if we accept the kind of operational definition suggested by the Embers, implementation is still problematic, as they point out, when working with ethnographies that are likely to use descriptions such as "most households are extended families" and not the precise census data presumed by their definition of an extended family. Ethnographic comments of the form, "most households are extended families," provide, at best, an ordinal level of measurement scale even though the operational definition can, in principle, be implemented as a ratio scale measurement.

Statistical Methods and Ethnographic Research Questions

With these concerns in mind, we can identify the range of questions and quantitative methods that have been fruitfully used in anthropology for analytical purposes. Table I indicates how a series of statistical methods relate to questions typically asked about anthropological data. As can be seen from the table, a wide variety of methods are employed that address the range of questions that typically arise when doing ethnographic research. Leaving aside the details of specific methods, several broad analytical goals have been incorporated into anthropological research using quantitative methods that are briefly reviewed here.

Patterning

Histograms and variants such as stem and leaf plots are often used to visually display the patterning in the numerical values of a single variable. The patterning is then expressed in terms of qualitative measures, such as the

Table I Research Questions and Relevant Methods[a]

	Research question					
Analytical method	*What is there to be explained?*	*Who agrees with whom about what and to what degree?*	*What is the agreement about?*	*Who (What) acts (looks) like whom (what) and to what degree?*	*What goes with what and to what degree?*	*Can we see a suspected relationship even after we control for everything else we can think of?*
Patterning						
Summary statistics	X	X	X		X	
Visual representation of patterning of single variables (histograms, stem and leaf plots)	X					
Visual representations of patterning of two or more variables (scatterplots)	X			X		X
Similarity coefficients (correlation and association)	X					
Model construction						
Regression analysis	X					X
Dimensionality reduction						
Principal component and factor analysis (ratio scale)	X	X	X	X	X	
Multidimensional scaling (ordinal scale)	X	X	X	X		
Correspondence analysis (nominal scale)	X	X	X	X		
Grouping of cases						
Cluster analysis	X	X	X	X		
Correspondence analysis (nominal scale)	X	X	X	X		
Patterning at the level of the individual case						
Consensus analysis	X	X				

[a] Modified from Table I in Handwerker and Borgatti (1998; p. 556).

shape of a pattern, and quantitatively through the use of summary statistics. Common qualitative distinctions made for the shape of a histogram include unimodal versus multimodal distributions, symmetric versus skewed distributions, and, where relevant, the presence of unusually extreme values or outliers. Numerical expression of patterning primarily focuses on measures of central tendency (mean for ratio scale variables, median for ordinal scale variables, and mode for nominal scale variables). Measures of dispersion include the standard deviation for ratio scale variables. The location of an individual case in the aggregated data (z scores for ratio variables, percentile for ratio and ordinal variables, and percent for nominal data) is, however, less often used, with the exception of percents for nominal data (which can include ordinal and ratio scale variables). When more than one variable is considered, scattergram plots are used for expressing the pattern displayed in the relationship between the variables, although in practice scattergram plots are used less frequently than might be expected even though they provide the primary means for discerning the patterning between two or more variables. Measures of association (e.g., Pearson product–moment correlation for ratio variables, Spearman's rho for ordinal variables, and the phi coefficient for nominal variables) are used to express the strength of the association between pairs of variables.

Statistical Inference and Bootstrapping

Cross-cutting all statistical methods are inferential procedures aimed at relating measurements made on sample data to the population from which the sample data were obtained. Samples are used, for example, to make inferences about population parameters such as the population mean, μ, or population standard deviation, σ. When measured over an entire population, these parameters are exact and not subject to sampling error. Similarly, the overall pattern for the values obtained when a variable is measured over all the members of a population is simply whatever is the pattern displayed in the population of measurements.

Despite the fact that the parameters of interest are computed over measurements made over all members of a population, researchers usually work with samples. Sampling is conducted for pragmatic reasons, such as limited time to take measurements, problems with getting access to all members of a population, cost of taking measurements, and the like. The sample is then used to obtain estimates of population parameters or population patterns based on measurements made over the members of a sample. With estimation comes assessment of the likelihood that a sample value (e.g., the sample mean) is within some specified numerical distance from the population parameter being estimated from sample data (e.g., the population mean, μ). The assessment is expressed in terms of the probability, p, of getting a deviation as small as, or smaller than, the difference between the measurement made over sample data and the population parameter being estimated. Computation of p typically assumes a random sample from the population and a particular pattern, such as a normal distribution, for the data values in the population.

Sampling that fully meets statistical requirements for both estimating population parameters from sample data and computing the value of p is problematic in anthropological research. Often, populations are not well specified, may vary with time (e.g., the members of a community), or may not be specifiable in advance of doing at least some preliminary research. In addition, the shape of the pattern for the population values may not match the shape assumed for computing the value of p. One way to resolve this sampling problem is to use nonparametric statistics that do not depend on the shape of the distribution of values in a population. For many of the more common statistical measurements, there are equivalent nonparametric statistics. The drawback of using nonparametric statistics, though, is that they generally are less powerful than their parametric counterparts; that is, one is more likely to accept a false hypothesis with nonparametric statistics.

One recently devised method for getting around the discordance between the actual pattern for the data from a population and the assumed pattern underlying a particular estimation procedure is to resample from the sample data. Resampling is used to construct a frequency distribution for the measurement, and this constructed frequency distribution is used to assess the likelihood of obtaining a deviation of a specified magnitude. Methods of this kind are called bootstrapping estimation procedures. They provide a way of dealing more realistically with populations that do not match statistical assumptions, although they do not solve the problems that may be introduced through nonrandom sampling. In addition, bootstrapping does not resolve problems that may arise when the population brought forward for analysis is heterogeneous with respect to processes postulated or known to be the basis for patterning observed in the data. Methods for addressing this problem are discussed in the following section.

Model Construction

Methods devised in statistics based on regression analysis (RA) are used as a way to construct a model for a set of variables as predictors (independent variables) of one or more response (dependent) variables. Typically, the goal is to account for the variance in the values of the dependent variable(s) by determining the relationship of the dependent variable(s) to one or more independent

variables. The canonical form of a RA model for two variables, X and Y, where X is the independent variable and Y is the dependent variable, is given by the expression, $y = \mu_{Y|X=x} + \varepsilon$, where the value, y, for the variable, Y, is decomposed into two parts: (i) $\mu_{Y|X=x}$ ("the mean of the values, y, measured on those cases in the aggregate where the variable X takes on the value, x") and (ii) ε ("residual"). The locus of the set of points $(x, \mu_{Y|X=x})$ forms the (deterministic) regression curve for the data points (x, y) measured over the aggregate being analyzed. The regression curve is then expressed explicitly as an equation whose graph is the regression curve. When the regression curve is a straight line, it becomes $\mu_{Y|X=x} = \alpha + \beta x$. This leads to the expression, $y = \alpha + \beta x + \varepsilon$, as a model for a scattergram plot in which the variables X and Y appear to have a linear trend. Typically, the values for the parameters α and β are estimated using least squares estimators. Furthermore, it is assumed that the actual value, y, of the variable Y for a given value, x, of the variable X deviates from the underlying, deterministic relationship, $\mu_{Y|X=x} = \alpha + \beta x$, in a "random" manner—that is, by some amount, ε, with the value of ε obtained on a given member of the aggregate assumed to be a random observation from a normal distribution of values with mean, $\mu = 0$, and fixed variance, σ^2.

In practice, most applications of regression models in anthropology do not rigorously determine whether or not the context in question satisfies all the assumptions of the regression model. Instead, a procedure such as ordinary least squares regression is used to estimate the parameter values in a linear model, and the parameter values are tested for statistical significance. Often, the residuals are not tested for homeoscedasticity (unchanging variance) and whether they are more or less normally distributed, even though this is a fundamental assumption of the regression model.

Because one can construct a regression model for whatever variables are selected as the independent variables and whatever variables are selected as the dependent variables, even when the parameters in a linear model fit to the pattern displayed in a scattergram plot are statistically significant, the pattern need not have anthropological significance. Therein lies the discordance between statistical methods and anthropological theory discussed in the archaeological literature.

Anthropological Significance versus Statistical Significance

One can mechanically apply a regression model to any set of variables and possibly arrive at a model with parameter values that are statistically different than zero—that is, a meaningful model from a statistical viewpoint. However, whether the model has anthropological significance depends on first having selected a population in which the dependent variable(s) has a single, underlying relationship to the values of the independent variable(s). Here is the problem, though. Consider a typical question for which one might use a regression model: Does the value of one variable predict the value, or outcome, of another variable? For example, cross-cultural studies have made extensive use of questions such as the following: Does the mode of production predict the form of social organization? The purpose of the question, however, is not exhausted by the specific result that is obtained. Rather, the underlying intent is to use the pattern of relationship that is found as a way to gain insight into the structuring processes operative in human societies.

Consider in detail a more prosaic example relevant to the goal of admitting students to a university: Does the SAT score of a graduating high school senior predict the grade point average (GPA) that senior may receive should he or she complete a college education and receive a BA degree? An admissions committee may be satisfied with whatever statistical result is obtained for the purposes of making decisions about admission, but from an anthropological perspective the statistical result would simply be an initial step in an analysis aimed at trying to work out the way in which a variety of factors—family background, personal attributes, prior education, ethnicity, and the like—relate to the way students become engaged in university life and how that engagement translates into performance measured by variables such as GPA. Consider the problems that arise for the anthropological analysis when the statistical analysis is not constructed in terms of the kinds of insights anthropologists have developed about human societies.

For simplicity, assume that the relationship between SAT and GPA appears to have a linear trend in a scattergram plot of these two variables. In the form of a model, it is posited that $\text{GPA} = \alpha + \beta\ \text{SAT} + \varepsilon$, where the relationship $\alpha + \beta$ SAT represents the deterministic relationship between SAT and the average GPA, $\mu_{\text{GPA|SAT}=x}$, received by all graduating college students with an SAT given by the value x; for example, if $x = 550$, then $\mu_{\text{GPA|SAT}=550}$ would be the mean GPA for all graduating college students who entered college with an SAT score of 550. The value ε represents the magnitude of all those effects that relate to the actual GPA of a graduating college student deviating from the predicted value based on the deterministic relationship between x and the means $\mu_{\text{GPA|SAT}=x}$. The relationship between SAT score and GPA, however, may result from a number of different "causes." Some students may study hard and have a GPA that reflects their level of involvement with their studies. Other students might decide to be involved in a number of different activities that conflict with intensive studying and so have a lower GPA than

would be the case had they studied in the same manner as the first group. Yet other students may decide that they will be satisfied with a minimal GPA and thus spend little time studying. Each group may have a different set of parameter values for expressing the relationship between SAT score and GPA at time of graduation. Hence, each group has its own model for relating SAT scores to GPA. Taken collectively, the parameter values for all graduating students considered as the aggregate will reflect an average of the parameter values for each group considered separately weighted by the size of each group. Thus, the parameter values will mainly have descriptive value and fail to provide insight into the processes underlying the relationship between SAT scores and GPA.

This example illustrates the underlying problem that can arise with application of statistical methods to anthropological data. At one level, statistical methods can be applied to any aggregate, but in so doing the results may be descriptively accurate but lack interpretation in terms of underlying processes. Interpretation depends on defining an aggregate within which the same structuring process(es) applies equally to all cases within the aggregate. However, therein lies a double bind. On the one hand, one may want to use statistical methods to determine the appropriate aggregates; on the other hand, methods for doing so depend on the aggregates already being determined. Clustering procedures have been advocated as a way to address the problem of heterogeneity, but they make unrealistic assumptions about the way in which phenomena considered by anthropologists are structured.

Clustering Procedures

Clustering procedures have had wide application in anthropological and archaeological data analysis because they are based on the concept of dividing a heterogeneous data set into homogeneous subsets. However, as has been demonstrated with archaeological data (but the problem is not limited to archaeological data), clustering algorithms for determining aggregates typically make computations that assume the variables used for the clustering jointly determine the space within which the clustering may be identified. Clustering procedures have presumed that as more variables are included in the analysis the clustering algorithm will tend to converge on the distinct aggregates that may be present in the data set brought forward for analytic purposes. In brief, clustering procedures assume a paradigmatic data structure in which all variables are equally useful for the determination of clusters or aggregates. However, it is likely that the data set brought forward for analysis has a taxonomic structure in which different variables are relevant for determining subaggregates that reflect underlying structuring processes.

To continue with the SAT example, we might define a series of "background" variables, such as *S*, the time spent on studies, and *E*, the time spent on extracurricular activities. A typical cluster analysis would assume that clusters, or aggregates, should be found in the two-dimensional space defined by *S* and *E*. However, suppose for illustrative purposes that there are two groups of three students—one group having a high value for *S* and the other group a low value for *S* as shown in Fig. 1A, with each group having a different pattern for the relationship between SAT scores and GPA. Initially, we are not knowledgeable of this difference among the students, and so our analysis begins by taking the six students graphed in Fig. 1A as the aggregate. Suppose that we want to determine whether there are subaggregates of students that should be distinguished prior to beginning the analysis of the relationship between SAT scores and GPA by using a cluster analysis based on the variables *S* and *E*. Consider, however, data structured in the following manner with respect to variables *S* and *E*.

With respect to the variable *S*, the students graphed in Fig. 1A form two well-defined clusters; that is, clusters satisfying the criterion of "internally cohesive and externally isolated." Suppose further that the values for the variable *E* are random with regard to the two clusters; hence, some of the students in each of the two clusters might have a large value for *E* as shown in Fig. 1B. Note that the pair of students *C* and *a* in Fig. 1B, although isolated from each other when considering the space determined by the variable *S* alone, are close to each other in the two-dimensional space determined by *S* and *E*. Thus, they form a third cluster in the two-dimensional space. However, the cluster they form has no bearing on the relationship between SAT score and GPA. The cluster appears in the two-dimensional space only because of the way the geometry of the clusters has been altered by introducing a variable for which the distribution of values is independent of the clusters that are present in a space in which this variable does not occur. However, the analysis would proceed under the assumption that

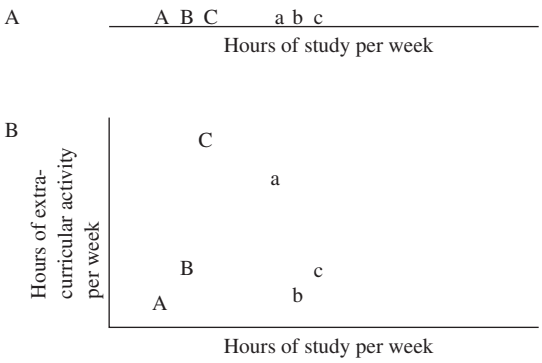


Figure 1

this third cluster found in the two-dimensional space is homogeneous when in fact it is heterogeneous with respect to the relationship between SAT score and GPA. If other variables having a pattern for the distribution of values independent of the two well-defined clusters are introduced, then the distorting effect becomes more pronounced.

The distorting effect is a consequence of having clusters that form a taxonomic, rather than a paradigmatic, structure with regard to the variables selected for measurement. I have demonstrated mathematically that the introduction of variables for which the distribution of values does not reflect clusters that may be present in a low dimension space leads to divergence from the embedded structure when considering the higher dimensional space determined by the full set of variables. This is contrary to the underlying assumption of most applications of clustering procedures as a way to reduce heterogeneity in data brought forward for analysis.

Dimensionality Reduction

With cultural data, measurements made at the level of observable behavior are used to infer processes at a cultural level through dimensionality reduction methods, such as principal component or factor analysis (ratio scale variables), multidimensional scaling (ordinal variables), and correspondence analysis (nominal variables). These procedures are used both descriptively as a way to reduce a higher dimensional space to a lower dimensional space in which most of the original variability in the data is present and inferentially as a way to construct new dimensions that reflect processes underlying the values of the variables as measured. Multidimensional scaling has been used extensively for this purpose in anthropology because it requires only measurements in terms of ordinal scale variables. Consequently, multidimensional scaling has made feasible the analysis of data in the form of paired comparisons ("Is *A* similar or dissimilar to *B*"), triads (" Is *A* more similar to *B* or to *C*"), pile sorts ("Sort the entities into groups of similar entities"), and the like as a way to discern an underlying cultural model for pattern displayed in the data. The culmination of this approach is represented by cultural consensus analysis as a way of determining whether or not responses to questions about a knowledge domain reflect shared, cultural knowledge about the domain or whether interrespondent agreement is better interpreted as resulting from factors such as shared experiences. The underlying assumption of cultural consensus analysis is that cultural patterning represents shared knowledge and hence should be patterning exhibited at the level of individual cases, whereas interrespondent agreement due to similar experiences should only exhibit patterning at the level of the aggregate. Interestingly, quantitative methods make a full circle here since cultural consensus analysis brings quantitative methods back to the underlying assumption of cultural anthropology that was implicitly, if not explicitly, rejected through the application of statistical methods to anthropology. Cultural consensus analysis returns to the notion of patterning on individual cases, but it does so through analytic demonstration, not by assumption.

The problem of data homogeneity is critical to dimensionality reduction applications that intend to go beyond description. Without data homogeneity, the dimensionality reduction procedures are responding to both the data structure within a homogeneous subaggregation and the data structure of the differences between subaggregations. As with the parameters in model fitting to heterogeneous data, dimensionality reduction procedures done over heterogeneous data will be a weighted average of the effects expressed in each of the subaggregates, thus making interpretation of the constructed variables difficult. The solution suggested in the archaeological literature to this problem lies in preanalysis of the data brought forward for analysis before utilizing statistical methods to work out the patterning in the data.

Preanalysis of Data

The problems previously discussed arise through variables that are not well defined over the data set brought forward for analysis, with data sets that lack homogeneity, or both. Four contexts for data analysis can be distinguished according to whether or not the data set is well defined or whether or not the variables are well defined, as shown in Table II. Mode 1 is where quantitative analysis should begin, whereas mode 4 is a common beginning point for data analysis. Preanalysis refers to steps that can be taken to reformulate modes 2–4 to mode 1 prior to beginning quantitative data analysis. In general, a shift from the right to the left in the table may utilize dimensionality reduction procedures, especially for variables that are known to only partially measure the structuring process that is of anthropological interest. A shift from the lower to the upper row typically involves identification of clusters by considering different ways the data structure may be viewed, such as using histograms with single variables or scattergrams with pairs of variables, where the variables may be as originally

Table II Modes of Analysis

	Variables	
Data set	*Well-defined variables*	*Not well-defined variables*
Well-defined data set	Mode 1	Mode 2
Not well-defined data set	Mode 3	Mode 4

measured or as modified through trying to formulate well-defined variables. A key aspect to preanalysis is using an iterative approach in which every time a subdivision of the data set is made the preanalysis is repeated on the newly formed data sets.

The shift from mode 4 to mode 1 depends on both restructuring the data set and reconstructing variables, yet each of these operations depends on the other for its successful implementation. An iterative approach allows for convergence to mode 1 through incremental steps rather than presuming a single algorithm that can simultaneously restructure the data set and reconfigure the variables brought forward for analysis.

Conclusion

Many of the questions asked by anthropologists naturally use quantitative measures, even if expressed in more qualitative terms such as many, most, and few. Often, one wants to know whether or not, when one has more of one thing, if there more of something else. In contrast, qualitative relationships generally take on an either/or form. At a theoretical level, the opposition between material and ideational approaches to modeling social phenomena parallels the opposition between quantitative and qualitative analyses. When societal properties arise in response to external conditions, societal patterning tends to have an underlying quantitative dimension because responses may be sensitive to the quantity and amount of external factors, whether it be individuals optimizing returns as hunters and gatherers or maximizing gains in a market context. In contrast, behaviors that arise out of ideational properties such as identities individuals take on as part of becoming a culture bearer tend to be differentiated by qualitative distinctions, such as kinship relationships and the social grouping to which one belongs. Quantitative analysis has tended to focus on the material side of human social groups or measures of behavior, both of which are more amenable to statistical methodology. To challenge perceptions of the declining use of quantitative analysis in anthropology, more methods need to be developed that are both methodologically rigorous from the standpoint of statistical requirements and conceptually sensitive to the full range of factors—ideational and material—that affect, constrain, and guide the form of human social systems and their associated cultural constructs.

See Also the Following Articles

Clustering • Correlations • Ethnography • Population vs. Sample • Scales and Indexes, Types of

Further Reading

Aldenderfer, M. (ed.) (1987). *Quantitative Research in Archaeology: Progress and Prospects*. Sage, Newbury Park, CA.

Bernard, H. R. (ed.) (1994). *Research Methods in Cultural Anthropology: Qualitative and Quantitative Approaches*, 2nd Ed. AltaMira, Walnut Creek, CA.

Bernard, H. R. (ed.) (1998). *Handbook of Methods in Cultural Anthropology*. AltaMira, Walnut Creek, CA.

Borgatti, S. P. (1992). *ANTHROPAC 4.0.* Analytic Technologies, Harvard, MA.

Buchler, I. R., and Selby, H. A. (1968). *Kinship and Social Organization.* Macmillan, New York.

Carr, C. (ed.) (1985). *Analysis of Archaeological Data Structures: Toward Logical Consistency between Data Structure, Technique and Theory*. Westport.

Chibnik, M. (1999). Quantification and statistics in six anthropology journals. *Field Methods* **11**(2), 146–157.

Ember, C. R., and Ember, M. (1998). Cross-cultural research. In *Handbook of Methods in Cultural Anthropology* (H. R. Bernard, ed.), pp. 647–687. AltaMira, Walnut Creek, CA.

Handwerker, W. P., and Wozniak, D. F. (1997). Sampling strategies for the collection of cultural data: An extension of Boas's answer to Galton's problem. *Curr. Anthropol.* **38**(5), 869–875.

Johnson, J. C. (1989). Inaugural issue. *J. Quant. Anthropol.* **1**(1), 1–224.

Johnson, J. C., Weller, S., and Brewer, D. D. (eds.) (2002). Special issue: Systematic data collection and analysis. *Field Methods* **14**(1), 3–118.

Pelto, P., and Pelto, G. (1978). *Anthropological Research: The Structure of Inquiry.* Cambridge University Press, New York.

Schneider, D. M. (1984). *A Critique of the Study of Kinship.* University of Michigan Press, Ann Arbor.

Trotter, R. T., and Schensul, J. J. (1998). Methods in applied anthropology. In *Handbook of Methods in Cultural Anthropology* (H. R. Bernard, ed.), pp. 691–735. AltaMira, Walnut Creek, CA.

Voorrips, A. (ed.) (1990). *Mathematics and Information Science in Archaeology: A Flexible Framework. Studies in Modern Archaeology*, Vol. 3. Helos, Bonn, Germany.

Weller, S. C., and Romney, A. K. (1988). *Systematic Data Collection. Qualitative Research Methods Series*, Vol. 10. Sage, Newbury Park, CA.

Quantitative Analysis, Economics

Charles G. Renfro
Journal of Economic and Social Measurement, New York, USA

Glossary

business cycle The tendency for an economy to experience perceptibly regular periods of expansion and contraction, often graphically represented as an undulating pattern around an upward trend.

econometrics Originally, in the 1930s, the unification of economic theory, economic statistics, and mathematics. Today, a body of techniques that tend to focus on the methods of estimating the unknown, usually presumed constant parameters of behavioral, quantitative relationships between economic variables.

game theory A body of essentially mathematical techniques seen as governing the competitive behavior of economic agents under conditions of scarce resources and unknown, but essentially probabilistic, outcomes.

index numbers A counting technique that attempts to measure relative quantities or prices of collections of non-homogenous goods and services.

input-output A technique originally developed by Wassily Leontief in the 1930s and 1940s that measures both the distribution of the production of an industry to all purchasers of that production, both intermediate and final purchasers, and the set of purchases by an industry of all inputs to that industry's production from itself and other industries. The result, under certain assumptions, permits an evaluation of the effects of changes in the scale of economic activity.

linear programming A mathematical technique designed to determine the optimal allocation of resources between competitive uses, or else the optimal configuration of activities, under conditions of limited resources.

This article traces the development of the methods of quantitative economic analysis from the earliest days of economics as an identifiable discipline, roughly during the 16th century, to the present day. It places particular stress upon the development of macroeconomic quantitative techniques, reflecting the historical development of economics first as concerned with macroeconomic phenomena and only later with the behavior of individual economic agents. In its approach, this account demonstrates various aspects of the interplay between the process of measurement and the development of theories of economic behavior.

Introduction

In the most basic as well as the broadest sense of the term, quantitative economic analysis refers to the process of quantifying, or measuring, the activities of economic agents, individually or in the aggregate, and then attempting to explain the characteristics of that observed behavior using these measurements, collectively called data. A more precise definition of what constitutes quantitative analysis in economics is in part a question of whether the primary focus is placed upon the purpose of the analysis or the specific techniques used. The techniques employed by economists today are in most cases relatively recent in origin, essentially from the mid-20th century, and are now closely associated with the use of both the modern computer and, increasingly, the Internet. Econometrics, for example, possibly the most well-developed quantitative sub-discipline of economics, dates to only the early 1930s as an identifiable body of techniques with the establishment of the Econometric Society, even if certain of the underlying methods originate at the beginning of the 19th century. When the relative extent of its development is considered, econometrics more specifically belongs to the period following World War II. A similar statement could be made about each of the other quantitative techniques now commonly used by economists, with perhaps one or

two exceptions. These other techniques include input-output analysis, linear and dynamic programming, representational measurement techniques such as index numbers, and in recent years, the developing field of experimental economics.

At the beginning of the 21st century, some 50 years after the first use of the electronic computer by economists, it is tempting to consider quantitative economic analysis almost exclusively in this context. Not only is the computer now employed to make most, if not all, of the necessary computations, but it has also increasingly become the means of organizing, distributing, and even collecting the data used, to the point that today, not just governments and other organizations but even individual economic entities such as supermarkets and other retail establishments have increasingly begun to amass point-of-sale and other data as a by-product of doing business—often for the explicit purpose of performing some type of subsequent analysis. If the techniques now used by economists are the specific focus, there is even less need to consider more than a handful of scholarly contributions prior to 1940. However, this argument, if adopted, not only isolates the study of economics from the methods of empirical analysis used, but also ignores the degree to which these techniques have developed as a consequence of the development of economics as a discipline.

The Interplay of Theory and Measurement

Economics as a discipline is the outcome of more than 400 years of collective effort to explain why economic agents behave as they do, and in what circumstances. During this time, economic concepts have progressively sharpened, but in certain general respects the issues addressed even in the early years are still familiar to the modern economist.

Macroeconomics, the study of the behavior of groups of economic agents, began as an attempt to explain economic growth and industrial development as a consequence of what is today called international trade. As early as the second half of the 16th century, Jean Bodin in France and John Hales in England produced treatises on money and economic development that addressed then-topical issues such as inflation and the effect of the agricultural revolution of that time. With the increasing power of central governments of nation states, the growth of trade during the next 100 years and its promotion, regulation, and taxation increasingly became a matter of policy, leading to the first attempts by writers such as Josiah Child and Thomas Mun during the 17th century to explain its determinants and effects. Of course, these centuries were also the period of the discovery and colonization of the Americas, which additionally helps to explain both the trade expansion and the growing degree of general awareness of its role in commercial life.

Trade was an obvious source of wealth for the individual and the nation, but so was the increased production of goods and services. William Petty, who lived in the second half of the 17th century, may to be regarded as one of the founders of political economy, today called economics, but the important element of his contribution is not only his attempt to interrelate the concepts of money, interest, taxation, and the value of land and labor as productive forces, but also his contributions to the foundation of statistics, particularly economic statistics. The most novel aspect of his argument was to bring empirical economic measurements to bear. He determined, as he said in the preface to his 1690 book *Political Arithmetick*, to "express [himself] in terms of *number*, *weight* or *measure*; to use only arguments of sense, and to consider only such causes, as have visible foundations in nature." In France in the middle of the next century, this theme was taken up by Quesney and the physiocrats who also stressed production but focused on agriculture as the principal source of a nation's wealth. The important methodological similarity was the attempt, in the form of Quesney's *Tableau Economique*, to explain empirically an economy's characteristics, using economic measurements to estimate the relationships. Subsequently, John Sinclair in Great Britain, the first to use the word "statistics," and Charles Alexander de Colonne in France were among those who attempted during the last 20 years of the 18th century to produce detailed compilations of their respective governments' accounts, as well as quantitative statements of economic and social conditions.

The 19th century, in contrast, became more conceptual in approach, following Adam Smith's synthesizing 1776 inquiry into the cause of the wealth of nations, which he famously cast in terms of the beneficial outcome of the selfish actions of individual economic agents. At the start of the century, Robert Mathus struck a discordant note, asserting that in as much as food production grows arithmetically but population geometrically, famine and want are the inevitable lot of mankind. Less well known are his writings on economic rent and related price concepts. The focus of such 19th century economists as Mathus, Mill and Ricardo, and even Marx was the more detailed conceptual explanation of microeconomic phenomena, culminating in the work of Bohm-Bawerk, Jevons, Marshall, Menger, and Walras at the end of the century. By then, production and cost were joined to the concept of marginal utility, leading to an explanation of supply and demand as the joint determinants of prices. Speaking in 1907, Marshall (pp. 7–8) proclaimed a "general agreement as to the characters and directions

of the changes which various economic forces tend to produce," but recognized that "much less progress has indeed been made towards the quantitative determination of the relative strength of different economic forces. That higher and more difficult task must wait upon the slow growth of thorough realistic statistics." In the 19th century, there were attempts to collect statistics on such things as earnings and hours worked, and there were a few methodological innovations, such as Marshall's chain index formula, that are relevant even today. But this work proceeded for the most part as isolated, individual efforts, even if often as the result of governmentally established inquiries and commissions.

The 20th century opened quietly, as Marshall's words indicated, but a shock to economic analysis soon came in the form of the first of two world wars. The years 1914–1918 marked a watershed. The sudden need to mobilize entire societies to support the war effort meant that it immediately became necessary to understand and, ideally, stimulate and control actual economic performance. In the United States, for example, entities such as the War Industries Board were established and, in this case, led to the construction by Day, Stewart, and others of the first reasonably extensive set of single industry production and price indices. But this work proceeded slowly. The fact that the major industrial countries had blundered into a conflict beyond previous imagination meant that few prior preparations had been made. Consequently, only after the war did statistical reports begin to appear in any significant numbers. Even then, immediately afterward, the war was at first seen as a temporary aberration. However, the early postwar years produced widespread unemployment, hyperinflation in Germany, and then, following a boom in the second half of the 1920s, a pervasive depression in the 1930s. This accumulation of circumstances progressively established the general recognition of needs, even if no immediate solutions were forthcoming.

From Occasional Observations to Modern Economic Statistics

Bricks cannot be made without straw. In a parallel sense, quantitative economic analysis presupposes measurement, including not just the process of observing prices, quantities, and other primary observations, but also the construction of economic statistics using these numbers. It is actually difficult to be certain in detail about the early modern history of economic measurement, because this activity is almost invisible in the late 19th and early 20th century economics literature, although particular pioneering efforts can be identified. These include Irving Fisher's 1906 attempt to establish appropriate concepts for the construction of economic accounts, Wesley Mitchell's *Business Cycles* in 1913, as well as King's first calculations from the U.S. Census of estimates of total income and wealth in 1915. Then, beginning in the early 1920s, came contributions by Bowley, Day, Mitchell, Stamp, Stewart, Williams, and Working, among others. Although it involves a certain degree of idealization, the work of these people can be collectively viewed as mapping a progressive movement, during the period from 1906 to the 1930s, from single industry indexes to combined indexes, expressing business conditions or aggregate industry behavior, to the computation of national income measurements and, finally, the basis for the subsequent construction of national income and product accounts, as well as a broad range of other economic statistics.

In order to comprehend what such work involved, it must first be recognized that analytical judgment is an integral component. For example, there are a number of economic concepts, such as inflation and the cost of living, that in the abstract appear to have definite meaning, yet are difficult to measure empirically. Inflation refers to a general rise in prices, but just how to construct a good measure of inflation is a continuing subject of debate: for instance, which prices should be included, and how should they be combined? Similarly, the concept of cost of living raises a number of questions, among them, whose cost of living? The meaning of the resulting statistics clearly depends on what is bought, which in turn obviously depends on who is buying, and possibly other particular circumstances. Each of these concepts is measured using a weighted combination of its constituent elements, the latter being individual prices that collectively express these measurements. The composite value is known as an index number. Obviously, when abstract concepts are involved, the index number measurement problem consists of determining both the constituent elements and the weights.

However, measurement problems do not disappear when simpler cases are considered. The general role of index numbers can be seen immediately by considering such ordinary objects as apples and oranges. Considering these items separately, there is obviously no problem making quantity measurements. So long as quality and type are fixed, price measurement is similarly straightforward, in principle. However, collectively, apples and oranges are fruit, and assigning a single number to measure the quantity or the price of "fruit" is clearly immediately problematic. By extension, the difficulty is much the same when it becomes necessary to add up the quantities of the entire range of goods and services produced by an economy, so as to measure both the total production and the value of that production: ingots of steel, numbers of computers, bushels of wheat, and so on cannot of course be simply added together. The value of what is produced in

principle can be added, but simply knowing the value of production over time does not provide a good measure of economic performance, unless prices remain constant. In a world in which prices of goods and services change at varying rates, it is necessary to measure both production values and aggregate price level changes in order to be able to distinguish the real changes in production.

The Marriage of Concepts and Analytic Tools

During the early 20th century, aggregate economic statistics began to be created as an organized activity, as just indicated. However, the tendency for an industrial economy to experience progressively somewhat regular periods of expansion and contraction was a recognized phenomenon throughout the 19th century, and came to be called the business cycle. As a subject for quantitative analysis, it attracted the attention of Wesley Mitchell and others even prior to World War I. Their attempt to document this phenomenon received considerable support from the postwar development of the production and price indices, together with estimates of employment, unemployment, and other measures of economic performance that existed, including the relatively easily measured stock price indices, such as the well-known Dow Jones indexes. By the process of first gathering performance statistics for production, employment, and other business conditions, next mapping stages of a prototypical cycle, and then classifying these various measures into the categories of "leading," "lagging," and "coincident" indicators, the business cycle can be both monitored and, to a degree, predicted. Or at least that is the theory. In practice, not only is each individual cycle in important respects unique, but a heavy dose of judgment is still needed in order to classify the individual indicator variables, both in their selection and in the specific way the information is combined to explain an economy's performance pattern. Notwithstanding such qualifications, business cycle analysis promised to Mitchell and his colleagues the possibility of ameliorating, if not eliminating, at least some of the negative effects of the cycle.

From this description, albeit brief, it should be evident that one of the characteristics of these business cycle indicators is that they form a relatively parsimonious collection of economic statistics. This trait was beneficial during the first third of the 20th century, when nothing even approaching a complete statement of national income and product accounts was available. But such parsimony also highlights the knowledge gaps that existed then, due to the lack of other economic measurements—such as business expenditures for plants, equipment, and inventories, consumer expenditures, imports, exports, and government purchases. The significance of these gaps is especially easy to appreciate in hindsight, given the degree to which these particular statistics are regularly reported today on the television and in newspapers, not to mention the extent to which they have become critical in the formulation of government economic policies. But their present familiarity can also obscure the fact that the concepts just mentioned—investment, consumption, imports, exports, and government purchases—did not exist as elements of an integrated explanation of the performance of an economy until 1936, with the publication of John Maynard Keynes' *General Theory of Employment, Interest and Money*. In fact, Keynes did not create these concepts out of nothing. During the period from 1920 to the early 1940s, independently of Keynes, the underpinnings of national economic accounts were in the process of being assembled, as indicated earlier. Consumers, governments, and businesses existed, and their purchases were recognized as accounting for the bulk of a nation's production. Nonetheless, until the *General Theory*, what was missing was the conceptual spark that locked these concepts into the now familiar macroeconomic intellectual framework and made it commonplace to think of the change in the aggregate demand for goods and services as integral to any complete explanation of the business cycle.

The immediate effect of Keynes' book was muted. At that point, his analytic framework was new, and it would take a further 10 years for Colin Clark, Simon Kuznets, James Meade, Richard Stone, and others to clothe these concepts with empirical meaning and thus develop a database for future quantitative analysis. Nevertheless, at the beginning of World War II, the combination of Keynes' *General Theory* and the publication in 1940 of his equally influential, if much less well known, *How to Pay for the War* set the stage for the expanded development of quantitative economic analysis in the postwar years. His *General Theory* drew upon many strands of the economic thought developed during the preceding three centuries to provide a conceptual basis for coordinated governmental economic policies. It incorporated a consistent explanation for the behavior of an economy, particularly in the context of a severe recession or depression, and provided a number of policy prescriptions. *How to Pay for the War*, which mapped out the empirical connections more precisely, was also closely associated with the British National Income and Expenditure Accounts that were first produced in 1940. Subsequently, this work helped to stimulate to the creation of national income and product accounts worldwide, particularly in the post war period. In his 1984 Nobel prize lecture, Richard Stone, who received this prize in recognition of the role he played in the development of modern national income and product accounting, described in some detail the development of economic accounting systems from the 17th century to the 1980s.

The Offspring of the Marriage

With the foregoing description as background, it is now possible to consider how an accounting system both translates into an analytic device and supports quantitative economic analysis. It so happens that even the earliest economic accounts, such as those of Petty and Quesnay, directly provide a means of representing at least certain of the characteristics of an economy. The closest modern parallel to Quesnay's *Economic Tableaux*, but a very substantial improvement on it, is the work on the inter-industry structure of the U.S. economy by Wassily Leontief, also first described in 1936. This work complements that of Keynes, providing a way to view the relationships between industries. For each industry included, what is called an input-output table specifies, row by row, the value of the industry's product purchased as an intermediate product by itself and every other industry, as well as direct purchases by final purchasers. These direct purchases, referred to as final demand, are the point of connection with the Keynesian analytical framework. Once such a table is constructed, it is immediately possible, using a simple arithmetic operation, to compute the proportion of each industry's product accounted for by each purchaser, as a consequence deriving a table of decimal coefficients. Although effectively requiring the assumption that these coefficients do not change as production rises and falls, once constructed, this table of coefficients immediately allows the consideration of the effects on industry output of changes in final demand and production levels. Other, related relationships between final demand, prices, and employment can be derived in a somewhat similar manner, extending Leontief's original approach, and in the process providing a reasonably complete quantitative, interrelated representation of an economy. Input-output models have been developed and applied to regional as well as national economies. Leontief's work began in advance of the publication of formal economic accounts by the United States and other governments, and in fact was influential in the development of those accounts, once they began to incorporate industry detail. On their own, input-output models do not, however, provide the means to consider analytically, or as a means of policy analysis, the demand side behavior of economic agents.

The macroeconometric model is another one of the analytic tools that has been developed in order to attempt to explain the behavior of economies, the first of which was created in 1939 by Jan Tinbergen. The national income and product accounts, combined with data on employment and other statistical measures, provide the skeleton for these models, particularly the structural models created during the period from 1950 to the present day by Lawrence Klein and others. These models normally incorporate Keynesian concepts, and combine statistically estimated behavioral equation specifications with accounting identities, as the income and product accounts are called in this context. The result is to provide a representation of, usually, the macro economy at any of several levels, national, multi-national, or regional. However, a number of model structures are possible, and by the end of the 1970s, there were a variety of forms of quantitative economic models, using econometric techniques and reflecting competing beliefs among economists as to how economic phenomena should be represented, as well as progress in the development of both the electronic computer and the underlying database. Models created during the early 1950s through the early to mid-1960s were estimated and solved using desktop electromechanical calculators and were generally no larger than 20–30 equations, but by the late 1960s, models of 200–400 equations were attempted. By the 1980s, computer technology had advanced to the point that models as large as 10,000 and more equations were developed, although most contained less than 1000 equations.

It is obvious that the type of quantitative analysis associated with both econometric and input-output models has been both enabled and stimulated by the invention of the computer, since they often require large quantities of data and intensive calculations in their construction and use. The computer itself is a child of World War II, with the first electronic computer having been created at the Moore School of the University of Pennsylvania for the purpose of computing ballistic trajectories for artillery shells. Although Leontief began his work in advance of its construction, he quickly saw the benefit of this machine to his work: he was the first economist to use the computer, although the particular one he used was in fact electromechanical in operation.

An immediate post-war development that, like the computer, also originated as a technique to serve the needs of the military was the mathematical technique developed by George Danzig in 1947, known as linear programming. Although developed independently, there is in fact a close mathematical connection between Danzig's and Leontief's work, at least given certain assumptions. It is characteristic of the input-output model that, by construction, there is a unique relationship between the sectorial outputs and final demand: the data used constitute observations on what occurred historically, so that this relationship is determined by that history. In contrast, Danzig addressed the situation, originally in the context of U.S. Air Force planning, in which a variety of inputs and outputs are *a priori* mutually compatible, posing the problem of choosing the best activity level: the question is, given distinct sets of inputs and outputs, each of which is feasible, what is the optimum configuration? Interestingly, there is also a connection between both Danzig and Leontief's work and another

area of quantitative economic analysis, known as game theory, the mathematical underpinnings of each coincidentally being the same. Although recently connected in the public mind with the work of John Nash, the central theorem and original development of game theory was due to John von Neumann in 1928; von Neumann also played an important role in the development of the first electronic computers, particularly the modern form known as a stored program computer, an ubiquitous example of which is the now-familiar microcomputer.

During the past 50 years, there have been further extensions of the techniques of econometrics, linear programming, and game theory. Econometric techniques are now applied to the explanation of stock prices and other financial variables; they have been applied by political scientists and historians, among others. Linear programming has been applied to a broad variety of problems, but has also been more generally extended, in the form of nonlinear programming. Game theory, similarly, is a technique with a host of applications beyond economic analysis, and such terms as a "zero sum game" are now a part of everyday language. Broadly speaking, these techniques not only have been developed and applied by economists, but also have been incorporated into the subject areas of such fields as operations research and decision sciences, and even to some degree in the physical and biological sciences. The applications are now extremely widespread.

However, while these developments occurred, economists additionally began to create an increasing number of distinct data sets containing observations on the behavior of individual economic agents. As mentioned previously, economists characteristically attempt to explain the behavior of people in their various roles as economic agents. Just as in the 19th century many economists began to focus on microeconomic behavior as an underlying conceptual explanation for macroeconomic phenomena, once database and computer technology permitted, a similar progression occurred, now focused on empirical research. An initial impediment was that some of the data sets collected, such as those administratively generated by tax authorities and other governmental agencies, are by their nature confidential. However, with the progressive development of the computer, beginning particularly during the late 1960s and early 1970s, it has become increasingly feasible to produce sample data bases, stripped of individual identifying information, that can then be used quite generally in applied economic research. Other governmentally collected data sets of this type include the population census and surveys of consumer purchasing patterns that underlie the construction of government price indexes, industry specific censuses and surveys, and the like, the latter of which also provide data for the construction of input-output tables.

Of course, not all microeconomic applied research has depended upon the availability of such administrative data, or data collected as a by-product of official censuses and surveys: during the past 40 years, there have also been an increasing number of microeconomic data sets collected independently, specifically to enable various types of economic and social research. And, of course, as mentioned earlier, during the past 20 years in particular, businesses have begun to develop point-of-sale and other data sets, derived from their operations, that are increasingly being used for proprietary research. The result has been a general, substantial increase in data resources that are often available for general research purposes.

The modern availability of both microeconomic and macroeconomic data, which during the past 10 years has become increasingly available via the Internet, has obviously supported the development of a broad range of quantitative analysis in economics, to the point that it is now difficult to classify this work, beyond simply saying that both the individual techniques and applications have become profuse. The behavior and decisions of economic agents acting as consumers, investors, importers, exporters, and even government bureaucrats is now quantitatively analyzed by economists on both the microeconomic and macroeconomic level. Economists consider productivity as a measure of the efficiency of production for a given level of human, machine, and other inputs. The way in which consumers choose to purchase goods and services, given their income and accumulated wealth, when prices are set or change is increasingly evaluated quantitatively. More broadly, the effect of the presence or absence of information about the world on the part of economic agents, particularly when this involves costs of acquiring relevant information, generally known as search costs, has become an active area of quantitative analysis. The effects of economic agents' expectations about the future has been seen, particularly during the past 30 years, not only as determining their behavior as individuals but also as potentially affecting the ability of governments to effectively pursue both short-term and long-term economic policies.

There is, however, a distinction to be made between the non-experimental and experimental. Economics has traditionally been regarded as a non-experimental discipline. Obviously, it is generally not possible to institute a particular government policy, observe the results, then roll back the clock, and try some alternative instead. In principle, it is infeasible, if not impossible, to perform controlled macroeconomic experiments. Based upon such considerations, quantitative economic analysis has generally been viewed as limited in its application to data that are generated as a by-product of economic activity. Economic theory, in turn, particularly microeconomic theory, has developed on the basis of postulates about the motivations

and behavior of economic agents that can be characterized as logically abstract, and not based on assumptions tested empirically. Concepts such as utility, profit maximization, and other behavioral precepts abound in the theories of economists, but as theoretical constructs these are essentially assumptions that have been made about the motivations of economic agents.

Notwithstanding limitations, during the past 30–40 years, a new area of economic quantitative analysis has emerged that is generally known as experimental economics. It of course remains true that experimentation with governmental economic policies under anything approaching controlled laboratory conditions is difficult, if not impossible. But the behavior of individuals as consumers, employees, and managers, and in their other roles as economic agents, is potentially subject to experimental observation. Generally, experimental economists have adopted the methodologies of psychologists and others who investigate conscious, and perhaps unconscious, human behavior. This research can in fact be seen as an extension of the prior investigations of economists: since the 18th century, they have considered markets in terms of the behavior of buyers and sellers, and in the past 50 years, game theory has sharpened the concepts; the difference is that in recent years it has proved possible to evaluate this microeconomic behavior experimentally. Strategies of behavior, as well as both the motivation of market participants and the degree of contextual consistency and rationality of that behavior, have been examined. It is too early to begin to evaluate to what degree these new investigations will feed back to the development of economic theory, but there is certainly scope for this. Microeconomic theory has generally been developed axiomatically, with the conclusions during the 20th century increasingly being stated in mathematical terms, resting upon a particular set of assumptions that can be shown to imply those conclusions. For example, consumers have usually been assumed to be "rational" and "consistent" in their behavior. Experimental economics has questioned these assumptions. This assault upon the assumptions may in time lead either to their confirmation or to the formulation of new behavioral theories that rest upon tested assumptions rather than what can at least sometimes be seen as mathematically convenient axioms.

Conclusion

Economists, because they study behavior that involves quantities, prices as well as the number of apples, oranges, tons of steel, bushels of wheat, and the like that may be demanded at a particular price level, have since the early 20th century led social scientists in the development of empirical analytic techniques. Many of the specific techniques developed have also been applied more widely, not only by other social scientists, but also by historians and even certain physical scientists. Particularly in the case of game theory and linear and dynamic programming, scholars with a background in operations research and other such disciplines might even view these techniques as so much a part of their own tradition as to cause them to question the degree to which these are "economic" in origin. However, it is also true that with the development of experimental economics in particular, economists have begun to import others' experimental techniques. Cross-fertilization has become endemic. The purpose of this account is not to make any particular claims, nor stimulate debate as to the origin of a particular technique. Instead its purpose is simply to describe, in rather broad terms, the development of quantitative economic analysis as an activity pursued by economists.

See Also the Following Articles

Economics, Strategies in Social Science • Game Theory, Overview • Regional Input–Output Analysis • Spatial Econometrics

Further Reading

Bodkin, R. G., Klein, L. R., and Marwah, K. (1991). *A History of Macroeconometric Model-Building.* Edward Elgar, Brookfield, VT.

Davis, D. D., and Holt, C. A. (1993). *Experimental Economics.* Princeton University Press, Princeton, NJ.

Dorfman, R., Samuelson, P. A., and Solow, R. M. (1958). *Linear Programming and Economic Analysis.* McGraw-Hill, New York.

Friedman, D., and Sunder, S. (1994). *Experimental Methods. A Primer for Economists.* Cambridge University Press, Cambridge, UK.

Greene, W. H. (2003). *Econometric Analysis,* 5th Ed. Prentice Hall, Upper Saddle River, NJ.

Marshall, A. (1907). The social possibilities of economic chivalry. *Econ. J.* **17,** 7–29.

Renfro, C. G. (1997). Economic data base systems: Further reflections on the state of the art. *J. Econ. Soc. Measure.* **23,** 43–85.

Renfro, C. G. (2004). Econometric software: The first fifty years in perspective. *J. Econ. Soc. Measure.*, in press.

Schumpeter, J. A. (1954). In *History of Economic Analysis*, (E. B. Schumpeter, ed.), Oxford University Press, Oxford, UK.

Smith, V. (1982). Microeconomic systems as an experimental science. *Am. Econ. Rev.* **72,** 923–955.

Stewart, W. W. (1921). An index number of production. *Am. Econ. Rev.* **11,** 57–70.

Stone, R. (1997). The accounts of society. Nobel Memorial Lecture, 8 December 1984. *Am. Econ. Rev.* **87,** 17–29.

Taeuber, R. C., and Rockwell, R. C. (1982). National social data series: A compendium of brief descriptions. *Rev. Public Data Use* **10,** 23–111.

Quasi-Experiment

Linda Heath
Loyola University Chicago, Chicago, Illinois, USA

Glossary

covariate A variable that is related to the independent variable of interest that could bias the estimation of causal relationship between the independent and dependent variable if not controlled.
randomization The process of assigning participants to experimental or control conditions based on a random process. Also called random assignment.
randomized experiment Any experiment in which participants are randomly assigned to experimental and comparison conditions by either the researcher or naturally occurring random processes.
threat to validity A process or factor that could undercut the veracity of the suggested causal relationship between the independent and dependent variables.
validity The veracity of the claims made of research findings.

Broadly speaking, quasi-experiments can be conceived of as research designs that operate without random assignment of participants to condition (generally because random assignment is either impossible or unethical to employ in that particular research area) but that seek to approach the causal strength of true experiments by research design techniques. Because the researcher cannot rely on randomization to create comparability of respondents in the experimental and control conditions, additional controls must be devised and employed to account for potential extraneous differences between the experimental and control conditions. At the simplest level of quasi-experimentation (i.e., the simple time series and the simple nonequivalent control group design), the experimenter simply describes the presence of potential differences between the experimental and comparison group participants and outlines strategies to minimize the occurrence of such differences. In more sophisticated quasi-experimentation, statistical controls and/or extensive design strategies are used to minimize, eliminate, or account for preexisting differences between experimental and control conditions. This article provides an overview of these simple and more complex quasi-experimental designs.

The Logic Underlying Experimental and Quasi-Experimental Research

Historical Context

Two very different research approaches set the stage for quasi-experimentation. The first, classical experimental design, formed the basis for much of the research in the natural sciences and for a great deal of the early work in certain areas of the social sciences (e.g., psychology). By rigorously controlling the experimental setting, researchers could isolate variables of interest and make strong causal claims. Both single participant experiments (such as those done by the early perception researchers Hering and Helmholtz and the early memory researcher Ebbinghaus) and carefully controlled between-subject experiments (such as those done by the learning researcher B. F. Skinner) drew upon the scientific logic that undergirded much of the natural sciences. The second approach that appeared in early research in the social sciences was the case study, a careful analysis of a single person or group of people to inform generalizations about human nature. The early works by Freud and Piaget illustrate this approach.

As the social sciences matured, however, researchers sought to go beyond the case study approach and bring experimental rigor to important social and psychological

issues that defied capture in the scientific laboratory. Both practical and ethical problems stymied these researchers' efforts. For example, in seeking to understand the etiology of mental disorders, researchers could not test hypotheses about early childhood abuses, family patterns, or socialization disruptions within a rigorous experimental design because it was patently impossible for researchers to control all the extraneous variables in a child's upbringing. Even if the practical constraints could have been overcome, the researchers still faced serious ethical limitations about what ills they could inflict on their respondents. Watson's Little Albert (if he really existed) illustrates this ethical problem. Numerous introductory psychology textbooks over the years have reported that Watson conditioned a young boy named Albert to fear white rats via traditional classical conditioning techniques, and that this fear generalized to white hair and Santa's beard. If this story is true, serious ethical concerns about inducing psychological disorders arise. Thus, researchers who sought to introduce scientific rigor to the investigation of serious pathologies and other social problems faced the dual problems of the feasibility of implementing an experimental design and ethical constraints if the design were feasible.

Thus was born the quasi-experimental approach to research design, which trades off some experimental rigor to allow the examination of important areas that defy study using classical experimental design.

The Cost of Forgoing Randomization

The strength of most classical experimental design lies in the process of randomization (the exceptions being in the physical sciences, in which control, not randomization, allows causal inference). By following a random procedure of assigning participants to the experimental and control conditions of the study, the researcher distributes systematic error (unrelated to the topic of study) evenly across conditions in most instances. That is, if the researcher conducting a study of upper body strength enhancement assigns men and women to experimental or control conditions by following a random procedure, without accounting for gender, on average, half of the women will be placed into the experimental condition and half will be placed in the control condition. The same would be true for men, on average. Rare quirky randomization outcomes do occur, but, on average, the outcome of a random process is equal distribution of irrelevant variables across experimental and control conditions.

In situations in which researchers have reason to suspect that a particular outside variable (such as gender) could seriously affect or interact with the variable of interest, they should force that variable to be evenly distributed across conditions by randomizing within levels of that variable (a process known as blocking). However, the strength of randomization is that, on average, it evenly distributes between conditions extraneous variables about which the researchers are concerned and even those variables about which they are not concerned (even if they should be). Thus, handedness, eye color, left brain dominance, and even zodiac sign would all be evenly distributed across conditions, on average, by following a random assignment procedure.

Distributing these irrelevant variables evenly across experimental conditions is important to avoid introducing systematic differences that could masquerade as treatment effects. For example, if new strength-building programs were being tested in a college setting, researchers who created the experimental group entirely from one class and the control group entirely from another class could introduce serious selection differences into the findings if one class was a nursing class that enrolled primarily women and the other was an engineering class that enrolled primarily men. Even if both classes were mathematics classes, the same degree of initial difference could be introduced if one class was a required math course for nursing (and therefore enrolled more women) and the other was a required course for engineering (and enrolled more men). We know that gender is related to body strength, so in this instance the type of selection difference we would be introducing is obvious. Another, more subtle type of selection difference could be introduced even if we used two sections of the identical course as our basis for assignment to experimental or control conditions. Athletes are often allowed to register early for classes in order to avoid times that conflict with their training schedules. If a researcher were to assign students in a 10 AM section of Math 118 to the strength-building condition and assign students in a 4 PM section of Math 118 to the control condition, selection difference could still be introduced into the findings if significantly more athletes were in the 10 AM section compared to the 4 PM section. Even if the researcher blocked on gender and made sure that the gender ratio was identical in each class, thereby controlling for differences attributable to gender, differences associated with athlete status (and therefore presumably level of physical fitness) could still produce spurious results in this design. Even if the researcher blocked on both gender and athlete status, differences could still be introduced into the results if the testing on the dependent variable occurred early in the morning and the students who sign up for 4 PM classes are particularly sluggish during morning hours.

The point is that researchers who employ designs without random assignment of participants to experimental and control conditions risk introducing systematic differences into their findings that can be misinterpreted as effects of the independent variable. Variables that

systematically covary with (or are confounded with) the variable of interest are potentially the true causal variable that produced the results observed in the research. Quasi-experimentation is the art of controlling for as many of these potential confounds as possible or, at least, testing the plausibility of those confounds as the true causal force. The strongest quasi-experimental designs are those that can control for a whole host of potential confounds rather than just laboriously testing one potential confound's effect at a time. Some of the most dangerous confounding variables may be those of which the researcher is totally unaware. Because these variables may never be identified as topics for further analyses, they might be able to introduce considerable confusion into the research literature.

Types of Validity

Validity refers to the truth or veracity of the claims about research findings, as opposed to reliability, which refers to the consistency of the research findings. Validity has many dimensions. At the basic level, internal validity (as described in the classic work by Campbell and Stanley) refers to the truth of the assertion that A caused B. Imagine that we have two groups of people: One group gets high levels of A and a second group gets no A. The group that gets lots of A also displays high levels of B, whereas the group that gets no A shows low levels of B. Internal validity asks, "Is A the cause of the observed variation in B?" Concretely, imagine one group consumes lots of vitamin A and displays high income, whereas a second group consumes no vitamin A and has low levels of income. Is vitamin A the cause of the different levels of income, or is there some other variable that covaries with vitamin A consumption that is really the causal agent for the differences in income levels?

Construct validity of the cause and the effect asks, "What exactly is A, anyway?" and "What exactly is B, anyway?" If a researcher randomly assigns children to a special preschool program or to a comparison condition in which they stay home with their caregivers and later observes that the children in the preschool condition score higher on a test of number recognition, the internal validity of the study might be strong. (We would actually need to know many more details before we could make that judgment.) However, what exactly is the treatment and what exactly is the effect? Is the treatment the enrichment activities provided by the preschool (the most obvious construct) or is it having a nutritious lunch and snack every day, or is it being out of the home for a period every day, giving the caregiver a sorely needed respite? Even when one has determined that something about the preschool assignment led to improved scores on the test of number recognition, the issue of construct validity can remain open to question. Similarly, what does the score on the test of number recognition really measure? Knowledge of numbers? The ability to follow testing instructions? Motivation to take this type of test? Familiarity with this type of test? Again, construct validity questions remain.

Another type of validity is termed external validity and concerns the generalizability of the research findings to other instances, settings, populations, and times. A causal relationship between an independent and dependent variable might be valid for a particular type of person, a particular place, or a particular point in time but invalid for other people, places, or times. For example, researchers might validly conclude that a particular antibiotic kills a particular bacterium in 1999, but this conclusion might not be valid in 2000 (if the bacterium has developed resistance to that antibiotic). Similarly, a drug prevention program might be effective for sixth-graders but not for tenth-graders. Also, the use of powdered formula for infants might convey a health benefit in places with access to clean water but be a health liability in places without clean water supplies.

A final type of validity that is of concern in quasi-experiments is termed statistical conclusion validity by Cook and Campbell in their oft-cited 1979 text. This type of validity addresses the appropriateness of the statistical procedures employed and the degree to which statistical assumptions have been satisfied. Although a researcher might correctly conclude that in the particular research being reported the independent variable did not produce a statistically significant change in the dependent variable as measured, this does not necessarily mean that the independent variable does not cause a change in the dependent variable. Too few participants (resulting in low statistical power), a poorly measured dependent variable (also resulting in a loss of statistical power), or failure to meet statistical assumptions can all mask a causal relationship between variables. On the other hand, exceedingly high statistical power can reveal a statistically significant relationship where no meaningful or practically significant relationship exists.

Different research designs are susceptible to different threats to validity that could compromise the internal validity, construct validity of cause and effect, external validity, and statistical conclusion validity. The strongest design for a particular research project is the one that minimizes the number of serious threats to validity that operate, but identifying that design involves a variety of judgments about the priorities among the different types of validity. Should external validity be compromised to improve internal validity? Should construct validity be so rigorously protected that threats to internal validity occur, or would a lesser level of construct validity that avoids problems with internal validity suffice? Following is a discussion of some of the classic quasi-experimental designs, along with explication of the main threats to

validity associated with these designs. The final decision about which design best suits a particular research situation must take into account the unique research situation and the research goals.

Bivariate Quasi-Experimental Designs

Nonequivalent Control Group Design

One of the most frequently used quasi-experimental designs is the nonequivalent control group design (NECG) (sometimes referred to as the nonequivalent comparison group design or the nonequivalent group design). This design is similar to the randomized experiment but does not employ random assignment to experimental and control conditions. (Recall from the previous discussion that this is an extremely important difference.) Probably the majority of studies that employ a NECG design use just one treatment group and one no-treatment comparison group, with a pretest and posttest observation for each condition. More than one treatment group representing different treatments or different levels of treatment are possible, as are more than one control or comparison groups (such as a placebo treatment group and a pure no-treatment group). Occasionally, researchers employ a NECG design with no pretest observation, although this type of design is extremely weak and termed by Campbell and Stanley a "preexperimental" design rather than a quasi-experimental design.

In the typical NECG study, random assignment of participants to experimental or control conditions is impossible or extremely difficult without seriously compromising the strength of the independent variable. For example, in one of the early studies of the effects of television violence on aggressive behavior, boys in one cottage at a reform school were allowed to watch their normal television programs, and boys in another cottage were allowed only to watch nonviolent programming. It would have been impossible to assign individual boys within a cottage to type of viewing because the cottage had only one television and the repercussions of allowing some boys in the cottage unfettered access to television programming while strictly limiting the access of other boys in the cottage would have introduced serious construct validity problems. Similarly, researchers interested in examining the effects of childhood events such as abuse or parental imprisonment cannot randomly assign children to such conditions.

Another situation that frequently results in NECG research involves an innovation that cannot be tailored to one person, such as a new classroom technique. A researcher with unlimited resources could randomly assign classroom or school to the new technique (thus using classroom or school as the unit of analysis), but researchers rarely have sufficient resources to study the dozens of schools or classrooms that would be necessary for sufficient power with those units of analysis. Consequently, the more common design is for all the classrooms in school A to get the new technique and all the classrooms in school B to continue using the old technique, producing a NECG design.

NECG designs are susceptible to all the threats to construct validity and statistical conclusion validity that apply to randomized experiments, plus the additional selection threats to internal validity that do not present major problems for randomized experiments. NECG designs, however, generally have fewer problems with threats to external validity than do most randomized experiments in the same research area.

By far the most prevalent threat to the NECG design is that of selection. A selection threat refers to the selection differences discussed previously as costs of foregoing randomization. The threat rests on the possibility that the type of participants who are selected for the treatment condition are in some important way different from the type of participants selected for the comparison condition. Imagine an attractive program offered to the first 40 children who sign up. That group would receive the program and form the treatment group. The next 40 children whose parents tried to sign them up after the program was already full would constitute the comparison group. It is not difficult to create a long list of ways that the parents who succeed in enrolling their children in the program might differ from the parents who fail to enroll their children. The first group of parents might be more motivated; might be more attentive to announcements; might have better transportation or better day care options, allowing them to arrive at the enrollment location earlier; might have fewer personal life crises that could detract from their ability to get their child enrolled; and so on. One can easily imagine that even before the treatment began, the children of such parents would have many advantages over the children whose parents cannot get them enrolled.

Some researchers mistakenly believe that having a pretest measure will overcome such selection problems, that they can simply subtract out the pretest difference or otherwise make corrections statistically using a pretest score. This assumption is generally weak, but this assumption is particularly problematic if maturational processes are operating. The pretest observation might fail to capture preexisting differences between the groups that have not yet manifested themselves in a measurable way. For example, if our attractive program from the previous example was a reading readiness program, the pretest measure might not show any differences between the groups if none of the children had yet developed reading readiness skills (meaning the pretest is basically all error). Both

groups could score identically at pretest, but real underlying differences between the groups could already exist (although not in a way that was measured by the pretest instrument). This selection by maturation interaction could distort the findings and produce the appearance of a treatment effect where none existed.

A selection by instrumentation interaction threat to validity could occur if the pretest were paper-and-pencil and the posttest were completed at a computer. For this to create a selection by instrumentation interaction, the participants in one group would need to be more comfortable with one of the measures than participants in the other group. Returning to the example of the first 40 children to enroll in the special program, if those children are more likely to have computers in their homes or to use computers at the local library than are the children in the comparison group, the treatment group might score equal to the comparison group at pretest and significantly higher than the comparison group at posttest for reasons that have nothing to do with the treatment. Many of the other threats to internal validity can similarly operate via a selection by threat interaction in the NECG design.

In addition to selection threats to validity, the NECG design can also encounter validity problems if participants in the different conditions are aware of the alternative condition. The study of boys in reform schools mentioned earlier fell prey to the "resentful demoralization of group with less desirable treatment" threat to validity. The boys in cottages who had their television viewing curtailed, causing them to miss their favorite show (Batman), created such uproar at the institution that the study had to be prematurely ended. The obverse could also happen if the group with the less desirable treatment tries to "show up" the researcher by performing especially well, termed the compensatory rivalry or "John Henry" threat to validity.

If others in positions of power are aware of the different experimental conditions, they can introduce two types of threats to validity. First, the control group might receive other special programs or benefits to compensate for not receiving the treatment in the research project. Such compensatory equalization of treatments can occur if, for example, a grade school principal creates a special activity or club for the students who were not chosen to be in the treatment group. If the compensatory option has the same effect as the identified treatment, both groups should score equally on the posttest, which could lead the researcher to erroneously conclude that the treatment was ineffective. Similarly, the principal could notice the children in the treatment condition improving in performance or behavior and decide to offer the treatment to all the students in the school. This diffusion or imitation of treatment threat to validity could make the identified treatment appear ineffective by raising the scores of the comparison group to those of the treatment group. Such a problem with the research implementation might not be evident from the data analysis, especially if a naturally occurring growth trend in outcome scores would be expected (e.g., in the case of reading scores among schoolchildren).

Simple Interrupted Time Series

The simple interrupted time series design makes use of data systematically collected at regular intervals over an extended number of observations. (The old rule of thumb was 50 or more observations, but recent analytic developments might allow that number to be much lower, especially with more than one observation per time point.) The time series can be based on measures taken yearly, monthly, daily, or even by minute or second. The interruption should be well specified and abrupt to produce interpretable results. For example, one of the classic time series analyses involved the examination of the introduction of a Breathalyzer law on the rates of traffic fatalities. As with most laws, the legislation instituting the Breathalyzer examinations of suspected drunk drivers took effect on a precise day at a precise time that was well publicized. Traffic accidents that result in fatalities can also be well specified as to time and location. The simplest version of this study examined the traffic fatality rate for many time periods before and after the introduction of the new law. A statistically significant change in level or slope of the time series of observations after the introduction of the new law would be interpreted as an effect of the law.

As is probably apparent, the major threat to the validity of the simple interrupted time series design is history. Some other event might coincide with the law change in the Breathalyzer example. For example, another law regarding the number of infractions before license revocation could have occurred simultaneously. Another possibility is that police procedures for dealing with drunk drivers changed, and those procedures, rather than the legal change, caused changes in fatality rates. Another potential history threat would be improved ambulance service, resulting in fewer fatalities even if the number of accidents remained constant.

Another very plausible threat to internal validity in simple interrupted time series designs is instrumentation. If the manner of data collection changes at a time that coincides with the presumed treatment, that change in instrumentation could masquerade as a treatment effect. For example, in response to a crime wave, a city might institute new policing procedures and change the way it records and tracks crimes. Also, in response to community concerns about sexual assaults, a city might open rape crisis centers and provide victim advocates to support victims in filing charges. A researcher who used the rape statistics to examine the effects of the crisis centers without careful consideration of the effects of the victim

advocacy program would risk drawing improper conclusions.

Multivariate Quasi-Experimental Designs

Nonequivalent Control Group Designs with Covariate

Although the NECG design allows the researcher to examine the effects of real-life, socially relevant, high-impact independent variables, the key concern is that some variable other than the putative independent variable is the real causal force on the dependent variable. For example, research on the effects of television viewing on criminal behavior cannot be examined in a true experimental design for a wide variety of ethical and practical reasons. Comparing the criminal paths of people who chose to watch a lot of television as children with those of people who chose to watch less is one variant of the NECG (a sort of epidemiological NECG). The major concern with such a design, however, is that some factor other than television viewing is the true causal variable. For example, children who exhibit high levels of aggressiveness toward their peers might not have many playmates and might retreat into television viewing as a result of their aggressive nature. If this aggressive nature then leads them to commit violent crimes, an NECG analysis could mistakenly appear to reveal that television viewing caused the criminal behavior instead of the aggressive personality. Similarly, time spent watching television is time not spent on other activities that might have a preventative effect on criminal behavior. Again, an NECG analysis could mistake the effects of not participating in those other preventative activities with an effect of television viewing.

The use of covariates in conjunction with the NECG design can reduce some of the risks of misinterpreting findings or misspecifying causal relationships. For example, a researcher who had access to earlier measures of aggressive personality characteristics could use these as a covariate to determine if the relationship between television viewing and criminal activity remained after controlling for aggressive personality characteristics.

The strongest covariate is one that controls for a whole host of potential threats to validity. Regarding the example of the parents who manage to enroll their children in a desirable program, a researcher could use a strong measure of "family functioning" as a covariate for several of those potential troublesome variables. In all cases, however, researchers' understandings of the research area and their careful deliberation about potential threats to validity are necessary precursors to any salutary effects of covariate analysis.

Interrupted Time Series with Control Group or Switching Replications

In the same way that covariates can help eliminate many threats to validity in the NECG design, the use of a comparison series or a series with switching replications can address concerns about threats to validity in the interrupted time series design. For example, in the Breathalyzer study mentioned earlier, researchers addressed a whole array of history threats to validity by including a comparison of traffic fatalities during hours that bars or pubs were open with traffic fatalities during the hours they were closed. If changes in weather, road conditions, or ambulance service had coincided with the introduction of the Breathalyzer law, those changes should be as evident during hours when the pubs are closed as when they are open. The Breathalyzer, on the other hand, should only show effects during the time the pubs are open (and just after). Results showed no changes during hours that pubs were closed and significant changes during hours that pubs were open.

Similarly, an examination of the impact of the introduction of television on crime rates compared cities that received broadcasting licenses just before the Federal Communications Commission froze licenses with those that received licenses just after the freeze was lifted. Early adopter cities showed an increase in theft rates immediately after the introduction of television that was not matched by the late-adopter cities. Immediately after they began receiving television signals, the late-adopter cities showed an increase in theft rates, rising to the level of those of the early adopter cities. Crime data from the late-adopter cities, therefore, served as the comparison series at the first intervention point, and crime data from the early adopter cities served as the comparison series at the second intervention point. This time series with switching replications design is one of the strongest quasi-experimental designs because for a threat to validity to be plausible, it would have to impact the first series (but not the second) at a time that coincided with the early intervention, and it would have to impact the second series (but not the first) at a time that coincided with the late intervention. Few, if any, problems with instrumentation, maturation, selection, history, and so on are plausibly posited to follow such a complicated pattern.

Conclusions

Quasi-experiments allow researchers to deal with serious social problems with scientific rigor that, although not equal to that of randomized experiments, controls for many of the most important threats to validity. In many instances, the gain in external validity from adopting

a quasi-experimental approach outweighs the loss in internal validity. In all cases, however, the researcher should strive to maximize both internal and external validity as much as possible. Thoughtful consideration about ways to identify and test potential confounding variables is crucial. Quasi-experiments can be one important component of a well-conceived program of research on important social issues.

See Also the Following Articles

Randomization • Time-Series–Cross-Section Data • Validity, Data Sources

Further Reading

Bryant, F. B., Edwards, J., Tindale, R. S., Posavac, E. J., Heath, L., Herderson-King, E., and Suarez-Balcazar, Y. (eds.) (1992). *Methodological Issues in Applied Social Research*. (Social Psychological Applications to Social Issues, Vol. 2). Plenum, New York.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Settings.* Houghton Mifflin, Boston.

Kendall, P. C., Butcher, J. N., and Holmbeck, G. N. (eds.) (1999). *Handbook of Research Methods in Clinical Psychology*. 2nd Ed. Wiley, New York.

Reichardt, C. S., and Mark, M. M. (1998). Quasi-experimentation. In *Handbook of Applied Social Research Methods* (L. Bickman and D. J. Rog, eds.), pp. 193–228. Sage, Thousand Oaks, CA.

Shadish, W. R., Cook, T. D., and Campbell, D. T., (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton-Mifflin, Boston.

Tinsley, H. E. A., and Brown, S. D. (2000). *Handbook of Applied Multivariate Statistics and Mathematical Modeling.* Academic Press, New York.

Quetelet, Adolphe

Alain Desrosières
National Institute for Statistics and Economic Studies, Paris, France

Glossary

average Approximating or resembling an arithmetic mean; approximately midway between extremes.
average man A fictitious man, invented by Quetelet, whose characteristics (e.g., size) are the averages of all individuals. It represents, for Quetelet, an ideal of perfection.
International Statistical Congress Created by Quetelet, it met every 2 or 3 years between 1853 and 1875 and included the main statisticians of the world. After 1885, it was replaced by the International Statistical Institute, which still exists.
law of errors The astronomers of the 18th and 19th centuries call the law of errors the probabilistic distribution, which was later called normal law by 20th-century statisticians.
mean A quantity of the same kind as the members of a set that in some sense is representative of them all and that is located within their range in accordance with a set rule.
probability The chance that a given event will occur.
statistical regularities According to Quetelet, some statistical series, such as crimes or suicides, are almost like constants. For him, it was the proof that there could be a *physique sociale*, with forecasting possibilities, as in astronomy.

Adolphe Quetelet lay the foundations of modern statistics in the two meanings of the word: (i) the organized collection and arrangement of quantitative data on the social world and (ii) the interpretation and analysis of data on large numbers by means of mathematical tools. He transposed to the social sciences the formalism of Gauss's law, derived from the theory of errors in astronomy. The concepts of "average man" and the regularity of statistical series are important contributions to the theory of social measurement.

Introduction

The Belgian statistician Adolphe Quetelet (1796–1874) played an essential role in the development and dissemination of "social measurement." He also largely helped to give statistics the importance it has today—"statistics" being understood here in the two standard present-day meanings of the term: (i) the organized collection and arrangement of quantitative data on the social world (usually under an official aegis) and (ii) the interpretation and analysis of data on large numbers by means of mathematical tools. For these purposes, Quetelet brought together and combined several previously distinct scientific traditions: that of the measurements practiced in 18th-century natural science (astronomy and physics) and that of the social surveys and administrative enumerations undertaken at the behest of states—also since the 18th century—within the framework of English "political arithmetic" and German "statistics." This reconfiguration of knowledge sectors and analytical techniques makes Quetelet one of the founding fathers of the social measurement that, in the 20th century, came to prevail in the social sciences and in the invention of new forms of governmentality. Statistics indeed implies the systematic organization of periodic large-scale measurements in the collection phase upstream. In the interpretation phase downstream, it requires the use of mathematical techniques, probabilistic or not, such as the calculation of means or the application of the law of large numbers, which make it possible to establish and identify macrosocial regularities.

Quetelet's initial training was both in the humanities and in the sciences. In 1819, he received a doctorate in mathematics with a thesis on focal conics, a class of curves used in optics. Having developed an interest in astronomy, he succeeded in the 1820s in convincing the authorities of the Kingdom of the Netherlands (to which

Belgium belonged until 1830) of the need to build an astronomical observatory in Brussels similar to those already operating in France and England. In 1823, to prepare the project, he spent a few months in Paris, where he met French astronomers and mathematicians, including Alexis Bouvard, François Arago, Pierre Simon Laplace, Joseph Fourier, and Siméon Denis Poisson. This stay was to have a decisive influence on the rest of his career. Quetelet discovered how the scientists he met used probability theory to assess measurement errors in astronomy. The core of this mathematical methodology was formalized by Carl Friedrich Gauss and Laplace around three closely linked notions: (i) the Gaussian distribution (later referred to as the normal law or distribution) of the measurement errors for an astronomical value, (ii) the choice of the arithmetic mean as the most probable value of the quantity measured, and (iii) the least-squares method as an optimization criterion. Quetelet imported and popularized the first two notions—the Gauss curve and the mean—in the social sciences. The third notion, the least-squares method, was not accepted until the late 19th century, thanks to the British statistician Udny Yule. The Gauss curve, which was assumed to reflect the distribution of measurement errors, was dubbed the law of possibilities. It did not receive the name normal law until the end of the 19th century.

From Astronomy to Social Science

Most of Quetelet's biographers state that he assimilated one major idea from his contacts with French scientists in 1823: the transfer of the most spectacular advances in astronomy—then the queen of sciences—to the social sciences. According to this somewhat simplified version of Quetelet's ideas, mathematical determinism, and hence the predictability of astronomical phenomena, can be transferred to physical and moral (i.e., social) measurements of human beings and, consequently, to the regularities observed in these measurements. Stephen Stigler paints a more nuanced picture of Quetelet's thought. Stigler helps us to understand the intellectual context in which this "knowledge transfer" was possible: The intermediate link in the sequence appears to have been meteorology. At that time, astronomy and meteorology were not as distinct as they are today. Scientists studying time and climate were struck by the complexity of these phenomena and the impossibility of reducing them to simple laws—in the same way as astronomers, for example, did with Isaac Newton's laws of gravity. Many interlocking causes, impossible to enumerate in full, were involved, making it even more necessary to think in terms of probability. Quetelet, it is claimed, remembered this when he later sought to construct a "science of man." He even made the comparison explicitly in 1853, when he organized an International Statistics Congress modeled on the meteorology conferences held to improve navigation at sea.

This interpretation of the cognitive and institutional shift achieved by Quetelet, from natural science to the nascent social sciences, is of great value because it directly introduces the multiplicity of causes and uncertainty in those very areas in which the simplifying determinism of natural science are manifestly inapplicable. Meteorology and social science share the impossibility of arriving at a comprehensive knowledge and measurement of the many factors involved in the phenomena described and, hence, of reducing them to simple laws. The only solution in either field is to multiply the observation loci by interconnecting them via a dense network of exchanges of standardized information. In this convergence between two seemingly different scientific fields, we find the two main features of the statistical approach promoted by Quetelet all his life. The first is the organization of observations, and the second is the treatment of the large numbers thus collected by applying tools derived from natural science: the normal law and the mean. In fact, the main element transferred from Quetelet's Paris visit was the high generality of possible uses of the Gaussian distribution postulated by astronomers to characterize the dispersion of their measurement errors. The Gaussian distribution is supposed to result from the combination of many small, independent effects. However, the novelty of the transfer lay in the radical change in the epistemological status of the normal distribution (for the sake of convenience, this now standard designation is used, despite its anachronism when applied to Quetelet's ideas).

Quetelet was indeed the first to observe that other measurements—for example, the heights of conscripts in a regiment—are distributed in the same "gendarme's cap" (i.e., bell-shaped) pattern. His stroke of genius was to bring together the two similar distributions (those of measurement errors in astronomy and those of conscripts' heights), thus creating a totally new entity: the average man. Astronomers know that a real star lies at the source of their imperfect, scattered observations. By calculating a mean, they can estimate the star's most likely position. Similarly, Quetelet argues that a new but no less real being—the "average man"—is the ideal model behind individuals, all of whom are different. For astronomers, the star's real position is the constant cause of their successive observations, which explains the normal shape of the distribution of measurement errors. Likewise, the average man is the "constant cause" of the height distribution. The reasoning was thus reversed: It is the normal profile of the height distribution that implies the prior existence of a constant cause—that is, the average man, whose most probable height is the mean of the observed heights.

Average Man and the Gladiator Metaphor

In fact, the constant cause argument was derived from Bernoulli's theorem or law of large numbers (1713). Consider an urn containing an unknown but constant proportion of black and white balls. If we draw a sequence of balls from the urn, the proportion of balls of either color converges toward that of the balls in the urn. The proportion in the urn appears to be the constant cause of the shares observed in the draws. A chain of statements established a continuum between Bernoulli's probabilistic formulations, the astronomers' error computations, the measurement of conscripts' heights, and the propensities to crime or suicide and their regularities. The metaphor of the "Gladiator's statue" is one of the ways to concatenate this series.

To promote an understanding of the apparently unexpected convergence between the distribution of measurement errors and that of different individuals' heights, Quetelet conceived the following metaphor, inspired by his youthful interest in the arts and letters. The King of Prussia admires a beautiful statue of a gladiator. He commissions 1000 copies by 1000 sculptors in his kingdom for distribution to his loyal courtiers. The copies are imperfect and, of course, all different, but they all broadly resemble the model. Thus, the model acts as the constant cause of the copies since the 1000 sculptors have sought to imitate it. According to the metaphor, the average man is the equivalent of the original, perfect statue of the Gladiator. Actual human beings are the imperfect reproductions of this average man, described as an ideal of perfection. Quetelet elaborates the metaphor in detail in his "Letter No. 20," included in his main book on probability theory published in 1846, "*Lettres à S. A. R. Le Duc Régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques*" ("Letters Addressed to H. R. H. the Grand Duke of Saxe Coburg and Gotha, on the Theory of Probability as Applied to the Moral and Political Sciences"). By transposing the diversity of the sculptors' copies to the diversity of measurements performed on human beings, he even adds another cause of variability: the effect of successive imperfect measurements taken on a single individual. Quetelet thus combines in the same model the two sets of causes of variation: the real differences between individuals and the imperfections of measurement procedures, analogous to those factored in by astronomers since the 18th century.

The Gladiator model thus provided the link between two uses of the notion of mean (or average) that had hitherto been very distinct, including in the vocabulary employed. In the 18th century, scientists distinguished between (i) the "mean proportionals" (*moyennes proportionnelles*) adopted by astronomers as the best possible estimate (calculated from several observations) of a single real but unknown value and (ii) the common values chosen to summarize and proxy a diversity of values of different objects. In the second case, the substitution was justified by figures of speech such as "by and large," "taking one year with another," "more or less," and "the strong carrying the weak" (*le fort portant le faible*), which effectively express the notion of mutual "compensation" implied by this use of the mean. In a sense, Quetelet resorted to epistemological strong-arm tactics by linking the two applications of means. However, he did stipulate an important restrictive condition: The computation of a mean of measurements of different beings is justified, according to him, only if the distribution of the measurements exhibits the profile of the much vaunted law of possibilities (i.e., a normal shape). Only the presence of this form justifies the assumption of a constant cause preceding the contingent and imperfect manifestations of the tangible beings observed. In so doing, Quetelet introduces a distinction between true means, whose computation is justified by the presumed existence of a constant cause, and false means, whose computation and use should be prohibited in cases in which the distribution is not normal, particularly if it is bimodal. Two examples of the latter are often quoted by Quetelet: the distribution of heights of buildings in a town and that of the life spans of a population of children born in a given year. In either case, speaking of a mean would make no sense.

Observed Regularities of Moral Statistics

Quetelet's transfer of cognitive tools relies on a second characteristic of means: their relative stability over time. The average height of conscripts possesses this important property. The height dispersion in any given year is fairly wide. In contrast, the average height of new conscripts is roughly stable from year to year, or at least it varies in a far smaller interval than that of the individual heights of conscripts in a particular year. The stability of the mean, as against the dispersion of individual cases, was to serve as the foundation for the use of statistics in the social sciences. A comparable time invariance was observed for other totalizations supplied by nascent administrative statistics. Exposing regularities justifies the practice of forecasting, which is one of the main tasks that the world of action demands from the world of the social sciences.

The identification of regularities was made possible by the publication of administrative statistics assembled under the heading of what was then known as moral

statistics. Such publications were issued on an increasingly regular basis from the 1830s on. They cover a range of diverse events, such as births, marriages, suicides, and crimes. Although these events may seem to be due to decisions governed solely by individual freedom, their annual numbers, revealed by official statistics, were remarkably stable, as was the average height of conscripts. The "inexorable budget of crime," prophesized by Quetelet, heralds the sociological determinism of Émile Durkheim and his successors. Just as each individual has a height and weight, he or she is also endowed with a "propensity" to marry, commit suicide, or kill another person. The statistics bureaus may be compared to astronomical observatories: They record facts that are stable and thus predictable. The discipline of social measurement established its scientific credentials by modeling them on the unquestioned credentials of astronomy. Indeed, in the 1830s Quetelet was just as busy lobbying for the construction of an observatory in Brussels as in organizing a statistical system that included population censuses as well as the collection of vital statistics (births, marriages, and deaths) and crime and health statistics.

Quetelet's importance in promoting social measurements, both in the social sciences and in the political management of the social world, results from the strength of that association between the state and science. The two sides of the link are (i) the administrative machine and its registrations, stabilized in conformity with codified procedures, and (ii) the simplicity and generality of the law of large numbers, issued from the application of probability theory to astronomical measurements. However, in the process, the State partly changes nature. To the generality arising from judicial law—taken as the social rule applicable to all citizens of a State—was now added the notion of statistical law, observed in society, and accordingly treated as independent of the State. Statistics, synonymous with the State when introduced in Germany in the 18th century, was now emancipated from the State and laid the foundation for an autonomous social domain. The State had to make allowance for these new "laws," which it had not decreed. Sociology, particularly that of Durkheim, was constructed partly in opposition to the science of legal experts and jurists; it relied instead on the statistical regularities that Quetelet was so good at orchestrating. The quantitative social sciences—with their variables whose effects are analyzed—descend directly from the table of constant causes and propensities of individuals, invented by Quetelet. The terms in quotes are shorthand forms for naming new and previously unthinkable aggregates and assemblages. The epistemological status of these social measurements was encapsulated in a box, whose invisible cogs consisted of the recording procedures, and whose outputs, naturalized by statistical language, were propensities or variables producing effects. These could now be measured and separated, for example, by logistic regressions.

Determinism and Free Will

Quetelet's ambition was to import into the social sciences what was the pride of specialists of natural science in the first half of the 19th century—the possibility of starting from systematic measurements to arrive at general laws of phenomena and hence to predict their future course. The most brilliant example of this was the astronomy born of the endeavors of Newton, Laplace, and Gauss. However, the macrosocial "regularities" exhibited by Quetelet concerning marriages, crimes, or suicides seemed to clash head-on with the idealistic philosophy of freedom and individual responsibility that thinkers—religiously inspired or not—were seeking to promote in the same period. If the total number of crimes and suicides perpetrated in a given year is stable and predictable, what remains of the free will of the moral individual? This question was debated by philosophers, theologians, and novelists, particularly in the German world. In 1912, the Belgian Jesuit Joseph Lottin devoted a large volume to "*Quetelet, statisticien et sociologue*" ("Quetelet, Statistician and Sociologist"). He admired Quetelet but went into a lengthy discussion on the link between the moral freedom of the individual and the mean regularities described by Quetelet. From a different perspective, the French Durkheimian sociologist Maurice Halbwachs posed similar questions in 1912 in his supplementary doctoral dissertation titled "*La théorie de l'homme moyen. Essai sur Quetelet et la statistique morale*" ("The Theory of the Average Man: An Essay on Quetelet and Moral Statistics"), which took a more critical stance. Since that period, such criticism has often been directed against quantitative sociology—in particular of the Durkheimian variety—although Quetelet is no longer mentioned. Although the man is frequently forgotten, the dispute over macrosocial determinisms triggered by his work endures. A detailed study of the preoccupation with this issue in 20th-century European literature is found in a 1993 book by the French philosopher Jacques Bouveresse, in which he discusses the influence of Quetelet and his philosophy of statistics on the Austrian writer Robert Musil. Significantly titled "*L'homme probable. Robert Musil, le hasard, la moyenne et l'escargot de l'histoire*" ("Probable Man: Robert Musil, Chance, Averages, and History's Snail"), the book shows that the theme of Musil's novel "The Man without Qualities" (i.e., without specific singularities) was heavily influenced by the German debates over average man and individual freedom.

Implementation of a Coordinated Network of Social Measurement

Although his work sparked an ideological debate of this kind, Quetelet was less a philosopher than a scientific and administrative entrepreneur. By the 1820s, he was working to set up coordinated observations in meteorology. In 1853, he initiated the first meeting in Brussels of the International Statistical Congress to bring together statisticians from the statistical bureaus created (often under his influence) in several countries during the two preceding decades. The goal was to establish a coordinated network that would gather information on population, the economy, and moral statistics using the most standardized definitions, classifications, and recording procedures possible. On several later occasions, he drew attention to the concomitance of two international conferences—one on statistics and the other on meteorology—in Brussels in 1853. In both cases, "governments wanted to reach agreement through their special delegates." The first conference "deals with the general statistics of the different countries and with the means of unifying the official documents intended for the administration and for science," whereas the second "deals with the navy and the agreement that needs to be achieved between the efforts of different nations to arrive at a knowledge of the laws governing the movements of the seas and the atmosphere, the depth and temperature of waters, and, in general, all matters of interest to the navigator" [Quetelet (1860) as quoted in Brian (1989, p. 122)]. This parallel confirms Stigler's suggested interpretation of the 1823 Paris trip. It covers four aspects of the international statistical system that Quetelet was seeking to promote: the establishment of a measurement network, the standardization of procedures in the network, the role of States in this implementation, and their usefulness (the proper administration of society, sea navigation), and not only scientific in the "pure science" sense.

Quetelet in the 20th Century: Posterity and Oblivion

Quetelet's ideas were incorporated into the social sciences of the 20th century, for example, through Durkheim's work, but his name has not been preserved in the pantheon of the founding fathers of sociology or economics, as are those of Durkheim, Max Weber, Adam Smith, and David Ricardo. The reasons for this relative oblivion are philosophical and technical. Intellectually speaking, the theme of the regularity of observed averages, which cast all philosophies of liberty and action to the wind, was perceived by some as highly simplistic. It is striking that after "*Le Suicide*" (1897) and until his death in 1917, Durkheim made no further allusion to Quetelet. However, as discussed previously, Quetelet is important for having promoted the implementation of observation and measurement networks. History shows that statistics—an essential phase in the construction of the social sciences—is destined to become anonymous and invisible after its networks are up and running.

In another area—the techniques of statistical analysis—Quetelet's simple tools (calculation of means and rudimentary test of the stability of time series) were left far behind by the mathematical tools (variance, correlation, regression, and chi-square test) invented in the late 19th century by British biometricians Francis Galton, Karl Pearson, Udny Yule, and Ronald Fisher. These tools focus on the analysis of distributions and on the relations between them. The "normal" law, dear to Quetelet, remained important but was now interpreted in an entirely different manner. Henceforth, attention would focus on the differences and hierarchic relationships between individuals, not on averages supposedly reflecting a macrosocial whole endowed with specific properties of regularity and predictability.

Quetelet is therefore an important figure in two fields: epistemology and the sociology of the social sciences. For the epistemologist, he was the first to theorize the use of official statistical records to deduce the formulation of empirical laws, in the sense of observed regularities that can be extrapolated to the future. When Quetelet's name is mentioned today, it is generally in this connection. In contrast, the sociologist and the historian of the social sciences view Quetelet as the tireless advocate of the establishment of statistical offices and commissions at the national and international levels, followed by the standardization of measurement procedures to allow comparisons in space and time. This second aspect is often forgotten when the history of the social sciences is told exclusively from the standpoint of the history of ideas. Quetelet's distinctive achievement was to fully unite these two dimensions—cognitive and organizational. This undoubtedly explains his success in the 19th century but also his relative fall into obscurity in the 20th century, once his ideas came to be perceived as simplistic and once the forms of institutional organization that he had promoted had become completely routine.

See Also the Following Articles

Bernoulli, Jakob • Durkheim, Émile • Human Growth and Development • Nightingale, Florence

Further Reading

Académie Royale, de Belgique. (1997). *Actualité et universalité de la pensée scientifique d'Adolphe Quetelet*. Proceedings of the conference held on October 24–25, 1996, Mémoire

de la Classe des Sciences, 3rd series, tome XIII. Académie Royale de Belgique, Belgium.

Armatte, M., and Droesbeke, J. J. (1997). Quetelet et les probabilités. le sens de la formule. In *Actualité et universalité de la pensée scientifique d'Adolphe Quetelet*, pp. 107–135. Académie Royale de Belgique, Belgium.

Bouveresse, J. (1993). *L'homme probable. Robert Musil, le hasard, la moyenne et l'escargot de l'histoire.* L'Éclat, Combas, France.

Brian, E. (1989). Observation sur les origines et sur les activités du Congrès International de Statistique (1853–1876). *Bull. Inst. Int. Statistique,* 47th Session, IIS**,** 121–138.

Desrosières, A. (1997). Quetelet et la sociologie quantitative. du piédestal à l'oubli. In *Actualité et Universalité de la Pensée Scientifique d'Adolphe Quetelet*, pp. 179–198. Académie Royale de Belgique, Belgium.

Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning.* Harvard University Press, Cambridge, MA.

Halbwachs, M. (1912). *La théorie de l'homme moyen. Essai sur Quetelet et la statistique morale.* Alcan, Paris.

Lazarsfeld, P. F. (1961). Notes on the history of quantification in sociology. Trends, sources and problems. In *Quantification. A History of the Meaning of Measurement in the Natural and Social Sciences* (H. Woolf, ed.), pp. 147–203. Bobbs-Merrill, Indianapolis, IN.

Lottin, J. (1912). *Quetelet, Statisticien et Sociologue.* Institut Supérieur de Philosophie, Louvain, Belgium.

Perrot, J. C. (1992). *Une histoire intellectuelle de l'économie politique.* Éditions de l'EHESS, Paris.

Porter, T. M. (1986). *The Rise of Statistical Thinking.* Princeton University Press, Princeton, NJ.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale.* Bachelier, Paris.

Quetelet, A. (1842). *A Treatise on Man and the Development of His Faculties.* Chambers, Edinburgh, UK [Trans. of Quetelet, 1835].

Quetelet, A. (1846). *Lettres à S. A. R. le Duc Régnant de Saxe-Cobourg et Gotha, sur la théorie des probabilités, appliquée aux sciences morales et politiques.* Hayez, Brussels, Belgium.

Quetelet, A. (1849). *Letters Addressed to H. R. H. the Grand Duke of Saxe Coburg and Gotha, on the Theory of Probability as Applied to the Moral and Political Sciences.* (O. G. Downes, Trans.). Layton, London [Trans. of Quetelet, 1846].

Quetelet, A. (1860). Sur le Congrès international de statistique tenu à Londres le 16 juillet 1860 et les cinq jours suivants. *Bull. Commission Centrale de Statistique,* **9**.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Belknap Press of Harvard University Press, Cambridge, MA.

Stigler, S. M. (1997). Adolphe Quetelet: Statistician, scientist, builder of intellectual institutions. In *Actualité et universalité de la pensée scientifique d'Adolphe Quetelet*, pp. 47–61. Académie Royale de Belgique, Belgium.

Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods.* Harvard University Press, Cambridge, MA.

Randomization

Bryan F. J. Manly
Universidade de São Paulo, Piracicaba, Brazil

Glossary

experimental unit A unit (e.g., a human subject) receiving a prescribed treatment as part of an experiment, with a variable of interest measured on the unit as a response to the treatment.
randomization test A test of a hypothesis based on the process of randomization that is used to allocate subjects to treatments.
sample unit The basic unit (e.g., a household) selected for study and measurement as part of a sample survey.
simple random sample A sample selected in such a way that every item available for selection has the same probability of being included in the sample, independently of the selection of any other unit.
stratified random sampling A sampling method whereby the population of sample units that is of interest is divided into groups or strata that are expected to be similar in terms of the variables of interest, and a simple random sample is taken separately from each of the strata.
target population The collection of all sample units that are of interest for a sample survey.

Randomization is a process that is used in the allocation of treatments to experimental units as part of some experimental designs and in the selection of sample units as part of some sampling designs. The intention in both cases is to eliminate biases due to differences between units that are not explicitly taken into account in the study design. In certain circumstances, the process of randomization can also be used as the basis for testing statistical hypotheses by determining whether a set of observed data was likely to have arisen by chance as a result of a random selection of sample units.

Introduction

Since the publication of Fisher's 1935 book on experimental design, it has been recognized that randomization is an important and desirable part of many experiments. The basic idea in this case, as propounded by Fisher, is that whenever there is some choice about assigning experimental units to treatments, the units should be randomly allocated to ensure that the allocation is independent of any preexisting differences between the units, such as the likelihood of a high response. This ensures a fair comparison between the treatments even when there are important differences between the units that the experimenter is unaware of and therefore cannot explicitly allow for in the experimental design.

Fisher first introduced the idea of randomization in his 1925 book on statistical methods in general. Before that time, it seems that there were no widely accepted procedures for the layout and analysis of experiments, and systematic arrangements were commonly used. Fisher was a forceful proponent of randomization, but there was much controversy in the early days about whether it was desirable or not.

Randomization in sample selection has a slightly longer history. According to Folks, the first real definition of random sampling was provided by the American philosopher Peirce, who is quoted as saying in approximately 1896 that "a sample is a random one, provided it is drawn by such machinery, artificial or physiological, that in the long run any one individual of the whole would get taken as often as any other." Strictly, this definition is not altogether satisfactory because it does not specify the independence of the selection of different individuals. It would, for example, apply if two individuals were always either selected or not selected together.

An alternative to random sampling for surveys is purposive selection, whereby samples are chosen to be representative of the population of units of interest. Cochran notes that random sampling and purposive sampling were the two rival methods used for sample surveys by local and national governments in the early 20th century. A commission appointed by the International Statistical Institute seemed on balance to favor purposive selection. However, the method was eventually largely abandoned because it provides no measure of the accuracy of estimation and is not as flexible as random sampling. An exception is the use of systematic sampling, which is sometimes used to obtain "representative" samples over space or time.

Randomization in Experimental Designs

As a hypothetical example of the use of randomization as part of an experimental design, suppose that five drugs for the relief of arthritis pain are to be compared. The subjects available will be 20 men and 20 women, with 10 men and 10 women younger than 60 years old and the other men and women older than 60 years old.

Table I lists the subjects in the four age–gender groups; within each group the order is based on the date of enrollment into the study. It would obviously be possible to allocate out the five drugs systematically to the subjects in each age–gender group, with the first two subjects receiving drug A, the next two receiving drug B, etc. However, there would be concern about that allocation because, for example, drug A will always be assigned to patients enrolled into the study at an early date, and it is conceivable that these patients might have something unusual about them in terms of their response to the drugs.

Therefore, it is sensible to allocate out the drugs in a random order. One way to achieve this is to assign each subject a random number between 0 and 1, as shown in the "random" column in Table I. Such numbers can be obtained either from a table of random numbers, such as is often included in statistics textbooks, or by generating the numbers in a computer spreadsheet (as was done in this case). If the subjects are sorted in order within groups based on the random number, and the drugs are then allocated out in order (A, A, B, B, C, etc.), then this is effectively a completely random allocation. Table II shows the outcome obtained for the example.

If nothing else, this type of random allocation avoids the possibility of any bias, conscious or unconscious, on the part of the experimenters in the allocation of the different drugs. Indeed, it is preferable that the experimenters and subjects do not actually know what the drug allocation was so as to avoid any biases in the measurement of responses to the drugs. It then becomes a double-blind trial.

Sometimes, it is thought that it is possible to improve on a random allocation by altering the allocation to make it

Table I List of Subjects for an Experiment on Pain Relief in the Order (within Groups) of Enrollment into the Study[a]

Group	*Subject*	*Random*	*Group*	*Subject*	*Random*
Female, < 60	1	0.284	Male, < 60	1	0.095
Female, < 60	2	0.044	Male, < 60	2	0.056
Female, < 60	3	0.745	Male, < 60	3	0.174
Female, < 60	4	0.808	Male, < 60	4	0.001
Female, < 60	5	0.593	Male, < 60	5	0.661
Female, < 60	6	0.075	Male, < 60	6	0.065
Female, < 60	7	0.368	Male, < 60	7	0.306
Female, < 60	8	0.627	Male, < 60	8	0.407
Female, < 60	9	0.950	Male, < 60	9	0.514
Female, < 60	10	0.899	Male, < 60	10	0.437
Female, > 60	1	0.132	Male, > 60	1	0.704
Female, > 60	2	0.314	Male, > 60	2	0.852
Female, > 60	3	0.160	Male, > 60	3	0.027
Female, > 60	4	0.862	Male, > 60	4	0.071
Female, > 60	5	0.866	Male, > 60	5	0.798
Female, > 60	6	0.890	Male, > 60	6	0.150
Female, > 60	7	0.008	Male, > 60	7	0.258
Female, > 60	8	0.663	Male, > 60	8	0.499
Female, > 60	9	0.818	Male, > 60	9	0.733
Female, > 60	10	0.202	Male, > 60	10	0.238

[a] The random values are uniform random numbers between 0 and 1.

Table II Random Allocation of the Drugs to Subjects Based on the Random Numbers Assigned to the Subjects

Group	*Subject*	*Random*	*Drug*	*Group*	*Subject*	*Random*	*Drug*
Female, < 60	2	0.044	A	Male, < 60	4	0.001	A
Female, < 60	6	0.075	A	Male, < 60	2	0.056	A
Female, < 60	1	0.284	B	Male, < 60	6	0.065	B
Female, < 60	7	0.368	B	Male, < 60	1	0.095	B
Female, < 60	5	0.593	C	Male, < 60	3	0.174	C
Female, < 60	8	0.627	C	Male, < 60	7	0.306	C
Female, < 60	3	0.745	D	Male, < 60	8	0.407	D
Female, < 60	4	0.808	D	Male, < 60	10	0.437	D
Female, < 60	10	0.899	E	Male, < 60	9	0.514	E
Female, < 60	9	0.950	E	Male, < 60	5	0.661	E
Female, > 60	7	0.008	A	Male, > 60	3	0.027	A
Female, > 60	1	0.132	A	Male, > 60	4	0.071	A
Female, > 60	3	0.160	B	Male, > 60	6	0.150	B
Female, > 60	10	0.202	B	Male, > 60	10	0.238	B
Female, > 60	2	0.314	C	Male, > 60	7	0.258	C
Female, > 60	8	0.663	C	Male, > 60	8	0.499	C
Female, > 60	9	0.818	D	Male, > 60	1	0.704	D
Female, > 60	4	0.862	D	Male, > 60	9	0.733	D
Female, > 60	5	0.866	E	Male, > 60	5	0.798	E
Female, > 60	6	0.890	E	Male, > 60	2	0.852	E

"fairer." An example of the disaster that this can cause is provided by the famous Lanarkshire milk experiment. This took place over 4 months in 1930 in Lanarkshire, Scotland, and was intended to compare the growth of schoolchildren given raw milk, pasteurized milk, or no milk. Unfortunately, within a school the selection of children as "feeders" (receiving milk) or "controls" (not receiving milk) was left to the principal of the school. Initially, a more or less random allocation was made, but unfortunately in the description of the experiment it is stated that "in any particular school where there was any group to which these methods had given an undue proportion of well-fed or ill-nourished children, others were substituted in order to obtain a more level selection." Given this flexibility, it seems that the teachers tended to allocate milk to poorly nourished children whom they thought needed it. This, together with other problems, turned a perfectly reasonable experiment into one in which the validity of the final results is very questionable.

Randomization in Sampling Designs

A well-designed sampling program will always include a clear definition of the sample units, which are the basic units selected for study such as the individual households in a survey of household expenditure, and of the target population, which is the collection of all sample units that are of interest. The sampling plan then describes how the sample units will be selected from the target population.

It is sometimes assumed that human beings are able to take a representative sample of units without needing to be concerned with randomization schemes, and often the word random is applied to any method of choosing sample units that appears to be arbitrary and without purpose. However, statisticians realized in the early 20th century that an arbitrary selection of units is often biased in comparison with what would be obtained by a truly random method. Some early examples are provided by Kendall and Stuart.

Thus, to overcome biases in sample selection, a proper randomization process is always desirable. Often, simple random sampling is used, whereby it is ensured that every unit in the target population has the same chance of being included in the sample, independently of all the other units. This is still often done by a mechanical process. For example, the names of the N units in the target population can be written on N cards, one per unit. These are then put in a container, and n cards are drawn to determine which units will be included in the sample to be taken. This produces an essentially random sample providing that the N cards are mixed up sufficiently in the container.

In practice, it will often be more convenient to use random numbers in a similar way to the random allocation of experimental treatments to patients that was described in the previous section and illustrated in

Tables I and II. One way to do this is to assign each of the N units in the target population a random number between 0 and 1, sort the units in order according to the value of the random number, and then take the first n units in the resulting list as the ones to be sampled. This randomization process ensures that each of the N units is equally likely to appear anywhere in the final list, independently of all the other units, and therefore gives a truly random sample.

Random sampling can be included in sample designs that are much more complicated than what has just been described. For example, stratified random sampling is often used, where the target population of sample units is divided into several groups or strata, such that within each of these strata the units are expected to be similar for the variable or variables of interest. Thus, the households in a city might be stratified on the basis of the suburbs where they occur, on the assumption that this will produce strata within which the households will tend to be similar in terms of their socioeconomic status. A simple random sample would then be taken separately from each of the strata, and the data that are collected from each household would be analyzed taking into account the way that the sample was determined. Stratified random sampling and other modifications of simple random sampling are reviewed by Manly.

Many populations that need to be sampled consist of units that are not easily identified. No list of the items is available, and therefore choosing a random sample is not just a matter of selecting units at random using random numbers or some mechanical process. In these cases, ingenious methods may be needed to attempt to obtain something approximately equivalent to a random sample. Some examples are discussed by Manly.

Inference Based on Randomization

When data are collected using a design that includes some process of randomization, it is sometimes possible to draw conclusions from the data based on the randomization by following a procedure first suggested by Fisher in 1936. The following example illustrates how this can be done.

Suppose that 30 adult females recently diagnosed as having schizophrenia are available for the trial of a new treatment for the disorder. Fifteen of the 30 patients are randomly chosen to be assigned the new treatment, and the others receive the standard treatment. After 1 month, the results for the two treatments are compared.

For each patient, there is a score at the start and the end of the month based on her symptoms, with high scores indicating many symptoms of schizophrenia. The effectiveness of the treatment received is therefore measured by the difference between the final score and the initial score (the change), with a large negative value being desirable. Table III shows these differences for the two groups of patients.

Table III Results from the Trial for a New Treatment for Schizophrenia

New treatment		*Standard treatment*	
Patient no.	*Change in score*[a]	*Patient no.*	*Change in score*
1	−7	1	−11
2	−25	2	−1
3	−3	3	−16
4	−4	4	−14
5	−7	5	12
6	−11	6	−18
7	−9	7	−11
8	−12	8	−9
9	−28	9	−10
10	−17	10	−20
11	−16	11	−23
12	−22	12	−13
13	−22	13	−10
14	−27	14	4
15	−12	15	−27
Mean	−14.80		−11.13
SD	8.36		10.10

[a] The difference between a score for symptoms at the start and end of 1 month of treatment, with large negative values indicating an effective treatment.

The mean change for the standard treatment is −11.13, indicating that there was some improvement on average for the 15 patients given this treatment. For the new treatment, the mean change is −14.80, suggesting that this treatment may be better than the standard treatment for reducing symptoms. However, there is obvious interest in knowing whether the mean difference, $D = (-11.13) - (-14.80) = 3.67$, is large enough to give real evidence that the new treatment is better.

The usual method for assessing the significance of the observed mean difference of 3.67 would involve performing either a two-sample t test or a nonparametric alternative, such as the Wilcoxon rank-sum test. However, an alternative procedure that has the merit of being very easy to understand examines the problem from the standpoint of asking whether the mean difference of 3.67 could reasonably have occurred as a result of the random allocation of the 30 women to the two groups.

The null hypothesis for the test based on the randomization is that the observed change for each woman would be the same, irrespective of the treatment given. If this is true, then the observed mean difference of 3.67

arose simply because of the particular randomization used in the experiment. To determine the probability of obtaining a mean difference as large as 3.67 on this basis, it is simply necessary to use a computer to carry out a very large number of alternative randomizations and find the proportion of times that a mean difference of 3.67 or more occurs. This can be done, for example, using an add-on to the Excel spreadsheet.

When 20,000 alternative randomizations were carried out, it was found that a mean difference of 3.67 or more occurred 15% of the time. Consequently, this size of difference is slightly unusual but not to the extent that it suggests that the difference is not just due to the randomization procedure. In the terminology of a test of significance, the significance level is $p = 0.15$ (the probability of getting a result as extreme as that observed by chance in the absence of a difference between the effects of the treatments). This would not usually be regarded as evidence against the null hypothesis.

More information about randomization-based methods of inference can be found in the texts by Edgington, Good, and Manly. See also the discussion by Lunneborg of the commonly occurring experiment, such as the one described previously, in which subjects are randomly allocated to two groups that are then given different treatment. Lunneborg notes that in certain circumstances conventional and randomization tests comparing the means of treatment will be too conservative, giving too few significant results.

See Also the Following Articles

Explore, Explain, Design • Population vs. Sample • Stratified Sampling Types • Units of Analysis

Further Reading

Blank, S., Seiter, C., and Bruce, P. (2001). *Resampling Stats Add-In for Excel User's Guide.* Resampling Stats, Arlington, Virginia.

Cochran, W. G. (1977). *Sampling Techniques.* 3rd Ed. Wiley, New York.

Edgington, E. S. (1995). *Randomization Tests.* 3rd Ed. Dekker, New York.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, UK.

Fisher, R. A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh, UK.

Fisher, R. A. (1936). The coefficient of racial likeness and the future of craniometry. *J. R. Anthropol. Inst.* **66,** 57–63.

Folks, J. L. (1981). *Ideas of Statistics.* Wiley, New York.

Good, P. (1994). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer-Verlag, New York.

Holschuh, N. (1980). Randomization and design I. In *R. A. Fisher: An Appreciation* (S. E. Fineberg and D. V. Hinkley, eds.), Lecture Notes in Statistics, Vol. 1, pp. 35–45. Springer-Verlag, Berlin.

Jensen, A. (1926). Report on the representative method in statistics. *Bull. Int. Statistical Inst.* **22,** 359–377.

Kendall, M. G., and Stuart, A. (1977). *The Advanced Theory of Statistics. Vol. 1: Distribution Theory.* 4th Ed. Macmillan, New York.

Lunneborg, C. E. (2001). Random assignment of available cases: Bootstrap standard errors and confidence intervals. *Psychol. Methods* **6,** 402–412.

Manly, B. F. J. (1992). *The Design and Analysis of Research Studies.* Cambridge University Press, Cambridge, UK.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology,* 2nd Ed. Chapman & Hall, London.

Picard, R. (1980). Randomization and design II. In *R. A. Fisher: An Appreciation* (S. E. Fineberg and D. V. Hinkley, eds.), Lecture Notes in Statistics, Vol. 1, pp. 46–58. Springer-Verlag, Berlin.

Ranking

Wade D. Cook
Seymour Schulich School of Business, York University, Toronto, Canada

Glossary

consensus ranking This term is used to describe the ranking of a set of alternatives that is in best agreement with the rankings supplied by a set of voters.
data representation This term refers to the manner in which one displays ordinal responses (e.g., pairwise comparisons and vectors).
intransitivity Inconsistency in preference specifications whereby the voter prefers for example **a** to **b**, and **b** to **c**, but also **c** to **a**.
ordinal data Data or information in the form of binary pairwise comparisons, or as rank positions on an *N*-point scale.
round robin tournament Refers to any competition or comparison in which each pair of members (teams, individuals, objects) competes exactly once.

The theory of ordinal ranking, data, and consensus involves (1) the evaluation of methods for formatting data when only ordinal preferences can be specified, (2) the resolution of possible inconsistencies in that specification, and (3) the derivation of a consensus of opinion when multiple sets of preferences are present.

Introduction

Data in real world decision problems appear in many different forms. Generally, data can be classified as falling into one of four groups.

(1) *Nominal data* are measurements that simply classify the units of the sample into categories. An example would be the political party affiliation of each individual in a sample of 50 business executives.

(2) *Ordinal data* are measurements that enable the units of the sample to be ordered with respect to the variable of interest. An example would be the size of car rented by each individual in a sample of 30 business travelers; compact, subcompact, midsize, or full-size.

(3) *Interval data* are measurements that enable the determination of how much more or less of the characteristic being measured is possessed by one unit of the sample than another. An example is the temperature at which each of a sample of 20 pieces of heat-resistant plastic begins to melt.

(4) *Ratio data* are measurements that enable the determination of how many times as much of the characteristic being measured is possessed by one unit of the sample than another. An example is the sales revenue for each firm in a sample of 100 U.S. firms.

In this article, attention is confined to ordinal data.

Information or data in the form of individual preferences are common in a wide variety of real-world comparisons and choice situations. When a consumer is asked, for example, to rate or compare several flavors of pudding, it is natural that an ordinal response be given: "I prefer the flavor of chocolate to that of vanilla, the flavor of strawberry to that of chocolate." It can be unreasonable in such situations to expect an individual to be able to quantify his or her responses to these stimuli (flavors) on an absolute cardinal scale (chocolate rates at 6 while vanilla rates at 5). Another example is the fitting of lenses by an optometrist. At each trial, the optometrist presents to the patient two lenses from which the patient chooses the one through which he or she sees the clearest. In this case, the patient is not even required to provide an ordinal ranking of the lenses, but simply to express ordinal preferences between two samples at a time. The result of this set of pairwise comparisons is the identification of the most suitable lens.

The analysis of stimuli impacts, therefore, constitutes an important area where the data are inherently ordinal.

Such ordinal preference data have been the subject of study for over two centuries. Initially, a problem arising in the theorizing on preferential elections of the 18th century, it has evolved into the social choice theory of today. An important group of problems involving ordinal data and ranking concern the aggregation of preferences provided by a set of individuals into a group preference function or a consensus. Numerous authors have investigated problems of ranking and consensus, including Blin and Whinston, Cook *et al*., Kemeny and Snell, and Kendall.

Dealing with problems involving ordinal data provided by individual responses, generally involve three issues. The first issue pertains to the format in which data concerning a set of alternatives should be collected. This depends upon the application under investigation and the number of alternatives at hand. The second issue, in some settings, involves resolving inconsistencies in preferences supplied by the respondent. Such inconsistencies arise when the data format chosen involves pairwise comparisons of alternatives. Finally, when multiple respondents supply preferences concerning a set of alternatives, there is the issue of combining these preferences into a consensus ranking of those alternatives.

Problem Settings

Using Consumer Preferences to Evaluate Products

Opinions of consumers play a dominant role in the development and marketing of new products. Private enterprise relies on such opinions to aid in targeting their products toward particular segments of society according to age, sex, economic status, and so on. Because of the enormous financial implications associated with new product development and changes to product formulations, the obtaining and processing of consumer perceptions are crucial elements in the product's success or failure.

To have a particular problem setting as a backdrop, consider the situation faced by a researcher who is collecting consumer responses pertaining to preferences among five formulations of a pudding mix, which a company is considering for production. Denote the alternative formulations as $a, b, c, d,$ and e. The formulations vary in texture, flavor, smoothness, and the like. The company is attempting to determine which formulation would be most favorable with the public and which segments of the public should be targeted as the primary market for the product in question. In simple terms, therefore, the ultimate purpose of the survey is to arrive at a preferred product from among the five options, or more generally to obtain a ranking of the products.

This problem has been the focus of extensive research in the field of marketing for decades. Hundreds of papers have been written on the subject, and scores of models for characterizing and evaluating consumer preferences have been advanced. A number of major issues must be confronted in dealing with this problem, including choosing an appropriate data format, resolving inconsistencies in responses from consumers, and deriving a compromise or consensus of opinions among multiple respondents.

Aggregating Voter Responses in a Preferential Ballot Election

In a preferential election, each voter selects a subset of k candidates from a ballot of m choices, and rank orders these k candidates from most to least preferred. Such a voting format is common in municipal elections where a number of candidates are required to fill various positions.

A number of models for aggregating preferential votes have arisen from parliamentary settings. These methods all utilize the set of values v_{ij}, giving the number of jth place votes received by the ith candidate on the ballot. Models that go under such names as the American, English, and West Australian systems, for example, have different ways of utilizing these statistics to arrive at a final list of winners. A sample of these methods is discussed below.

Ranking Players in a Round Robin Tournament

Many sports events involve the direct competition of one individual or team against another. While many variations exist, one form of competition, the round-robin tournament, characterizes several games. Consider the case of a chess tournament, where a group of players competes in a pairwise fashion. Each match between a pair of players results in a win of one player over the other or in a draw, and generally every pair is involved in exactly one match.

The outcomes of all matches can be represented by a binary preference matrix of a form similar to that presented in the previous application. The rows and columns are labeled by the players $P_1, P_2, \ldots, P_N$, and the entries $a_{p_i p_j}$ are 1s, 1/2s, and 0s, depending on whether P_i won against P_j, tied, or lost.

$$A = \begin{array}{c|cccc} & P_1 & P_2 & \cdots & P_N \\ \hline P_1 & 0 & 1 & \cdots & 0 \\ P_2 & 0 & 0 & & \\ \vdots & & & & \\ P_N & 1 & 0 & \cdots & \end{array}$$

Ultimately, it is necessary to determine a clear winner out of the competition. As observed in the previous application, however, cycles can occur. Cycles are present here, not because of an inconsistency in stating preferences, but rather that player P_1 defeats P_2 who defeats P_3 who defeats P_1. Such a phenomenon is often the rule rather than the exception, partially because the teams or players are generally evenly matched.

While standard procedures have been adopted to score players in tournaments such as chess, many different approaches have been developed and studied in the literature that attempt to rank players in a manner that best captures the outcomes from the matches. Some of these account for the strength of the players: a win against a strong player is valued more than one against a weak player.

In cases where not all pairs of players or teams compete, opponents must be ranked using an incomplete set of outcomes. Here it may be necessary to evaluate indirect results from matches. If P_1 beats P_2 and P_2 beats P_3, but P_1 and P_3 never compete, P_1 could be viewed as superior to P_3 because of the chain of outcomes, which may be assumed to be transitive. On the other side of the scale, there can be multiple matches: two players may compete several times. Furthermore, this number of matches may be different for some pairs of players than for others.

The problem of ranking players in a tournament is then clearly one where binary pairwise comparison data must be evaluated. In its most generic form, it can be viewed in the same manner as the problem of resolving intransitivities in a consumer's stated preferences. At the same time, in a multiple-match tournament, consensus issues also arise.

Representation of Ordinal Preferences

A number of different models for dealing with ordinal preference have been developed. Three of the most common are presented below.

Object-to-Object Representation

One of the most common frameworks for eliciting individual preferences is the pairwise comparison method in which each pair of alternatives or objects is compared in an ordinal sense. Specifically, preferences concerning n alternatives are represented in an $n \times n$ pairwise comparison matrix $A = (a_{ij})$ where

$$a_{ij} = \begin{cases} 1 & \text{if alternative } i \text{ is preferred to } j \\ 1/2 & \text{if } i \text{ and } j \text{ are tied} \\ 0 & \text{otherwise.} \end{cases}$$

In the case of 4 alternatives, a, b, c, d, where a is in second place, b in first, c in fourth, and d in third, the preference matrix is given by

A_1:

0	a	b	c	d
a	0	0	1	1
b	1	0	1	1
c	0	0	0	0
d	0	0	1	0

Alternatively, if a is in first place, b in fourth and c and d are tied for second and third positions,

A_2:

	a	b	c	d
a	0	1	1	1
b	0	0	0	0
c	0	1	0	1/2
d	0	1	1/2	0

It is noted that this representation is equivalent to the matrix representation of Kemeny and Snell, where $a_{ij} = 0$ if i and j are tied, $a_{ij} = 1$ if i is preferred to j, and $a_{ij} = -1$ if j is preferred to i.

Vector Representation

Cook and Seiford offer a vector representation $A = (a_1, a_2, \ldots, a_n)$, where a_i is the rank or priority assigned to alternative i. The vector representation of the preference relation given by A_1 above is (2, 1, 4, 3). That is, alternative a is ranked second, b first, c fourth, and d third. The priority vector corresponding to A_2 is (1, 4, 2.5, 2.5). The 2.5 designation indicates that alternative c and d are tied for second and third place.

Object-to-Rank Binary Matrix Representation

Blin has suggested an alternative to the Kemeny and Snell model for complete orderings. Armstrong *et al.* have extended this to include ties (weak orderings). Specifically, an ordering is defined by matrix $P = (p_{ij})$, where

$$p_{ij} = \begin{cases} 1 & \text{if alternative } i \text{ has rank } j \\ 0 & \text{otherwise.} \end{cases}$$

It is assumed here that i takes on the values 1, 1.5, 2, 2.5, $\ldots$, $n-1$, $n-0.5$, n. The P-matrices corresponding to A_1 and A_2 above would then be given by

P_{A1}:

	1	1.5	2	2.5	3	3.5	4
a	0	0	1	0	0	0	0
b	1	0	0	0	0	0	0
c	0	0	0	0	0	0	1
d	0	0	0	1	0	0	0

and

$$P_{A2} = \begin{array}{c|ccccccc} & 1 & 1.5 & 2 & 2.5 & 3 & 3.5 & 4 \\ \hline a & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ c & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ d & 9 & 9 & 9 & 1 & 0 & 0 & 0 \end{array}$$

Geometric Interpretations

Ordinal preferences represented as priority vectors have a convenient description in a geometric sense. A necessary (but not sufficient) condition that a ranking vector satisfies is

$$\frac{p(p+1)}{2} \le \sum_{i \in K_p} a_i \le \frac{2np + p - p^2}{2}, \quad p = 1, \ldots, n$$

for every subset K_p, of the indices, of cardinality p.

In Figs. 1 and 2, this condition is utilized to illustrate the space of rankings for $n = 2$ and 3. In each case the rankings lie on the $(n-1)$ dimensional simplex given by

$$\sum_{i=1}^{n} S_i = \frac{n(n+1)}{2} \quad x_i \exists 0 \quad \text{for all } i.$$

The complete rankings are the extreme points that result from imposing the additional constraints

$$\frac{p(p+1)}{2} \le \sum_{i \in K_p} x_i \le \frac{2np + p - p^2}{2} \quad \text{for } p = 1, 2, \ldots, n-1.$$

(For $n = 2$ and 3 these additional constraints reduce to 1 # x_i # n for all i.) The tied rankings correspond to midpoints of the faces. (The ranking in which all objects are tied is the midpoint of the polyhedron.) See Figs. 1 and 2. It should also be noted that the space of rankings for $n-1$ objects is naturally embedded in (as a face of) the space for n objects. This is illustrated for $n = 4$ in Fig. 3.

Resolving Intransitivities in Pairwise Comparison Preferences

When preferences are expressed in a pairwise comparison format, as discussed above (object-to-object representation), the problem of providing a ranking of the alternatives being considered is often hampered by intransitivities in specification (e.g., a is preferred to b, b to c, and c to a).

In the application to ranking players in a tournament, as discussed above, this issue was raised. The tournament ranking problem has attracted considerable attention since the early work of Zermelo.

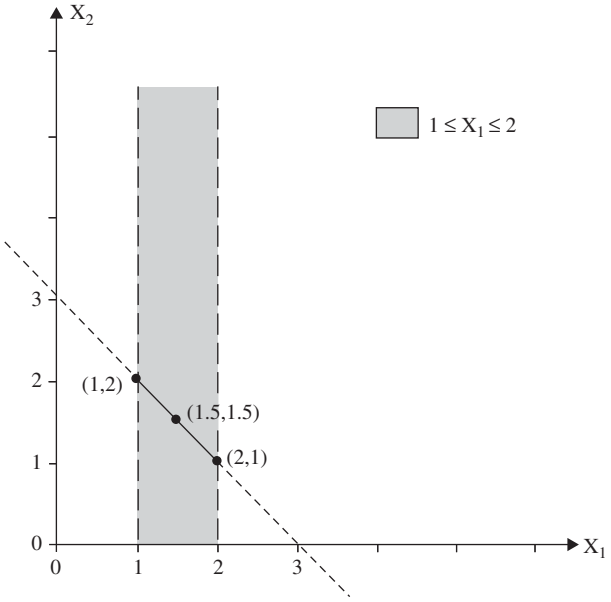


Figure 1 Ranking space for 2 objects.

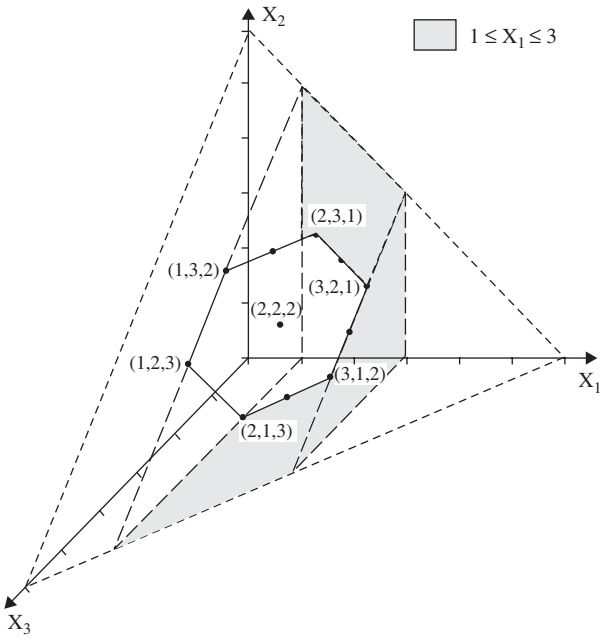


Figure 2 Ranking space for 3 objects.

The Kendall Scores Method

Numerous techniques have been suggested for deriving a ranking of the "players," with perhaps the most simple being the Kendall scores approach, which we briefly discuss here. The "score" of an alternative i is defined as the number of alternatives to which i is preferred, or (in sports) the number of players that were beaten by player i. This method was first proposed by Kendall.

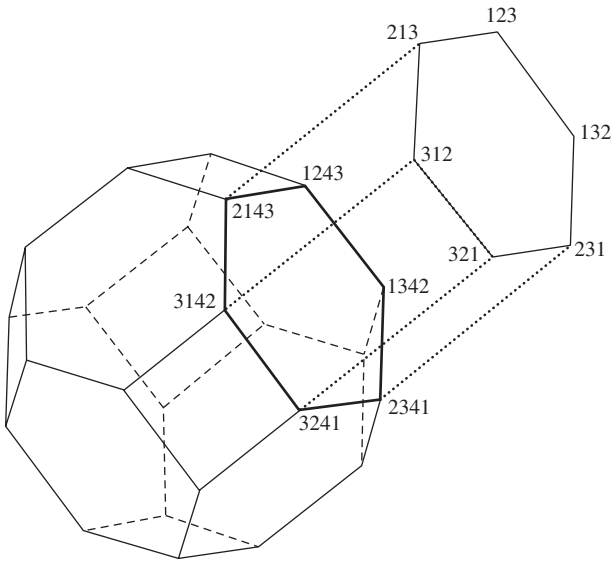


Figure 3 Ranking space for 3 objects embedded in space for 4 objects.

An Example

Consider the preference matrix

$$A = \begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 1 & 1 & 0 \\ b & 0 & 0 & 1 & 1 & 0 \\ c & 0 & 0 & 0 & 1 & 1 \\ d & 0 & 0 & 0 & 0 & 1 \\ e & 1 & 1 & 0 & 0 & 0 \end{array}$$

The KS vector for this preference structure is $(3, 2, 2, 1, 2)^{\mathrm{T}}$; that is, the sum of the elements of the first row is 3, that of the second row is 2, and so on. If the criterion for ranking the alternatives is "i is ranked higher than j if the score of i is greater than that of j and only if the score of i is at least as high as the score of j," then the resulting possible rankings are (a, b, c, e, d) or (a, b, e, c, d) or (a, c, b, e, d) or (a, c, e, b, d) or (a, e, b, c, d) or (a, e, c, b, d). Thus, six possible rankings result from this preference structure. In general, if there are k strings of tied scores where the ith string consists of n_i alternatives, then the number of possible rankings that are consistent with this set of scores is $(n_1)!(n_2)! \cdots (n_k)!$. In this example, there is one string with three alternatives tied; hence, there are $3! = 3 \times 2 \times 1 = 6$ rankings.

To demonstrate the possibility of multiple rankings, consider preference structures of order 5. There are $2^{10} = 1024$ possible 5×5 preference structures. Table I presents the frequencies of the numbers of possible KS rankings for those preference structures.

The preference structures with a single ranking number $5! = 120$, with KS equal to a permutation of the vector (1, 2, 3, 4, 5). The 24 preference structures each with 120 possible rankings represent the case where the KS vector is (2, 2, 2, 2, 2).

Breaking Ties: The Iterated Kendall Method

Given the high probability of ties among the Kendall scores, the question arises as to how to reduce the number of possible rankings. One approach is the Iterated Kendall ranking method. In this procedure, those alternatives with tied scores are pulled out from the preference structure and constitute a preference structure or substructure of their own. An attempt is made to rank order this substructure according to KS, and this procedure is repeated as far as possible. Formally, the method involves three steps:

1. Rank order the alternatives according to Kendall scores. If there are no ties, the resulting ranking has no violations and the procedure terminates. Otherwise, go to step 2.
2. Break the ties among any k tied alternatives by considering only the $\binom{k}{2}$ outcomes of these k choices, and then performing the ranking as in step 1.
3. If there are ℓ alternatives that are tied among themselves, select any one of these ℓ, and place it first in the subranking. The tie among the rest of the $\ell - 1$ alternatives is broken by performing step 2.

Applying this method to the above example, define the substructure that includes alternatives b, c, and e only. This substructure is

	b	c	e
b	0	1	0
c	0	0	1
e	1	0	0

The KS vector of this substructure is (1, 1, 1), and unless a winner is picked according to step 3, the tie still cannot be broken, and therefore we cannot eliminate any of the six rankings. Even though we can easily construct an example where this iterated method can reduce the number of possible rankings, in many cases it will not work.

Table I Frequencies of Multiple KS Rankings

No. of possible rankings	*Frequency*
1	120
4	480
6	400
120	24
Total	1024

Another approach is to rank order tied alternatives according to a Hamiltonian order.

DEFINITION 1. The ranking $(a_1, a_2, \ldots, a_n)$ is said to be a Hamiltonian order if and only if a_1 is preferred to a_2, a_2 is preferred to $a_3, \ldots, a_{n-1}$ is preferred to a_n. That is, the rank order of any consecutive pair of alternatives agrees with their pairwise comparison results in the preference structure.

Out of the six KS rankings in our example, only three rank the tied alternatives (b, c, e) in a Hamiltonian order. These rankings are (a, b, c, e, d), (a, c, e, b, d), and (a, e, b, c, d). Furthermore, since e is preferred to a and d is preferred to e, there is only one complete Hamiltonian ranking, (a, c, e, b, d). Therefore, by applying the reasonable Hamiltonian requirement to the tied alternatives, it is possible to reduce the number of KS rankings by one-half. By applying this rule to the entire set of alternatives, we could single out a unique KS ranking.

Ad hoc Consensus Methods

As discussed above, many problem settings require finding a consensus among a set of declared preferences.

Numerous approaches have been suggested in the literature for aggregating individual rankings in order to arrive at such a compromise or consensus. While some of these approaches can be linked to a particular piece of literature, e.g., Borda, others have simply evolved over time via parliamentary procedures, preferential voting needs, etc. In this section, some of these "*ad hoc*" approaches are very briefly examined.

These *ad hoc* methods can be grouped under two headings—elimination and nonelimination methods.

Nonelimination Methods of Consensus

Borda's Method of Marks

This approach, due to Borda, and later discussed at length by Kendall, is based on deriving the total of the ranks for each alternative as assigned by the voters. Consider the following 3-alternative, multivoter example

23 *votes*:	1	2	3
17 *votes*:	3	1	2
2 *votes*:	2	1	3
10 *votes*:	2	3	1
8 *votes*:	3	2	1

Total for

a: $23 \times 1 + 17 \times 3 + 2 \times 2 + 10 \times 2 + 8 \times 3 = 122$

b: $23 \times 2 + 17 \times 1 + 2 \times 1 + 10 \times 3 + 8 \times 2 = 111$

c: $23 \times 3 + 17 \times 2 + 2 \times 3 + 10 \times 1 + 8 \times 1 = 127$

The consensus here by Borda's Method is then $b > a > c$ or $A^* = (2, 1, 3)$.

Several modifications of the Borda Method have been developed, including those due to Cook and Seiford.

Simple Majority Rule or Condorcet's Method

Condorcet proposed a method whereby alternative x should be declared the winner if for all $y \neq x$, x is preferred to y by more voters than the number who prefer y to x. Similarly, y would be ranked second if for all $z \neq x$ or y, y is preferred to z by more voters than the number who prefer z to y. Consider the following example.

	Alternative		
	a	b	c
Voter#1	1	2	3
Voter#2	2	1	3
Voter#3	1	2	3
Voter#4	2	3	1
Voter#5	1	3	2

$a > b$ by 4 voters

$b > c$ by 3 voters

$a > c$ by 4 voters

Thus, the consensus ranking by simple majority rule is $a > b > c$ or $A^* = (1, 2, 3)$.

One problem cited by Condorcet, and one that is often encountered in applying this method is the occurrence of intransitivity, giving rise to the so-called "paradox of voting" or Condorcet effect. Example 1 above illustrates this phenomenon. There, $a > b$ by 33 voters out of 60, $b > c$ by 42 voters, and yet $c > a$ by 35 voters. Thus, a cycle arises, and the simple majority procedure breaks down. It has been shown that in the case of a uniform distribution of 3 alternatives, intransitivity occurs 8.8% of the time. For 4 alternatives, this probability is approximately 16%. Niemi and Weisberg and others have obtained estimates of such probabilities for a number of combinations of voters and alternatives.

Several "Condorcet completions" have been developed to deal with such intransitivities.

Elimination Methods

These methods gained popularity in parliamentary settings.

Runoff from Top Method

One such procedure consists of each individual first voting for the prospect he most prefers, and if there is no majority on a first ballot, a second vote is taken after eliminating the prospect with the fewest "first choice" votes on the first ballot. This appears to be identical with what is sometimes

called the "West Australian System" and very close to what is sometimes called the "English System," particularly for only three prospects.

Runoff from Bottom Method

This approach has approximately the same appeal as the runoff from top method. On each successive ballot the voters choose the prospect to eliminate.

The American System

The system sometimes identified as the "American System" is apparently designed only for use with preferential ballots which collect full rankings on the first round. Here, if there is no majority of first choice votes, the option with the fewest first choices is eliminated along with all those ballots whose first choice was the eliminated option.

Pairwise Majority Rule

Finally, there is the basic method defined by pairwise majority rule, which has special practical as well as theoretical appeal as long as it results in a determinate outcome. Some other methods have sought to make simple modifications to the majority rule approach. In Copeland's method, the prospect of x is more preferred the greater the number of prospects which lost to x relative to the number to which x loses.

Distance-Based Consensus Methods

In this section, aggregation or consensus among a set of preferences is examined from the point of view of a distance function. This concept has intuitive appeal in that a consensus is defined to be that set of preferences which is closest in a minimum distance sense to voter responses. This idea was first advanced by Kemeny and Snell, and was later adopted by Blin and by Cook and Seiford.

The approach is to define a distance function on the set of all preference orders which satisfies certain desirable properties. These properties or axioms are related to social choice properties. For purposes of presentation, the vector model of Cook and Seiford will be used as the preference representation.

Cook and Seiford propose that any distance function d_{cs} on the set of all priority vectors should satisfy the axioms:

Axiom 1. $d_{cs}(A, B) \geq 0$, with equality iff $A \equiv B$.

Axiom 2. $d_{cs}(A, B) = d_{cs}(B, A)$.

Axiom 3. $d_{cs}(A, C) \leq d_{cs}(A, B) + d_{cs}(B, C)$, with equality holding if and only if ranking B is between A and C. A, B and C are said to lie on a line in this case, hence d_{cs} is additive on lines.

Axiom 4. (Invariance) $d_{cs}(A, B) = d_{cs}(A', B')$, where A' and B' result from A and B, respectively, by the same permutation of the alternatives in each case.

Axiom 5. (Lifting from n to $(n+1)$-dimensional space). If A^* and B^* result from A and B by listing the same $(n+1)$st alternative in last place, then $d_{cs}(A^*, B^*) = d_{cs}(A, B)$.

Axiom 6. (Scaling) The minimum positive distance is 1.

It can be shown that the unique distance function which satisfies this set of properties is the ℓ^1 norm

$$d_{cs}(A, B) = \sum_{i=1}^{n} |a_i - b_i|.$$

Consensus among a set of voter priority vectors $\{A^\ell\}_{\ell=1}^m$ is then given by the vector $B^* = (b_1^*, b_2^*, \ldots, b_n^*)$, which solves the minimization problem

$$\sum_{\ell=1}^{m} d_{cs}(A^\ell, B^*) = \min_B \sum_{\ell=1}^{m} d_{cs}(A^\ell, B) = \min \sum_{\ell=1}^{m} \sum_{i=1}^{n} |a_i^\ell - b_i|.$$

A similar axiomatic structure has been proposed by Kemeny and Snell for pairwise comparison priorities and for the Blin model (see the work of Armstrong *et al.*). The Kemeny and Snell distance is defined as

$$d_{KS}(A, B) = \sum_i \sum_j |a_{ij} - b_{ij}|,$$

where A and B are pairwise comparison matrices. The Blin distance function is given by

$$d_{cs}(P, Q) = \sum_i \sum_j |p_{ij} - q_{ij}|,$$

where P, Q are object to rank binary matrices.

To better understand the distance models presented above, it is useful to examine an interesting connection that can be derived by starting with the Blin model and extending it to include degree of disagreement. Two of these extensions lead directly to the Kemeny and Snell and the Cook and Seiford models, respectively. Approaching the design of the latter two models from this direction lends an important insight into the level of complexity vis-à-vis the solution of these models.

Rank-Based Distance

One point of departure from the simple Blin representation is to define a function in which the aggregate disagreement between voters is measured according to the location of the alternatives relative to the various rank positions.

DEFINITION 2. The position j forward indicator vector $P^+(j)$ and the position j backward indicator vector $P^-(j)$

are those vectors whose kth components are given by

$$(P^+(j))_k = \begin{cases} 1 & \text{if object } k \text{ is ranked in a} \\ & \text{lower position than } j, \\ 0 & \text{otherwise} \end{cases}$$

$$(P^-(j))_k = \begin{cases} 1 & \text{if object } k \text{ is ranked in a} \\ & \text{higher position than } j, \\ 0 & \text{otherwise,} \end{cases}$$

respectively.

The consensus model arising from this position based extension of the Blin function is one of determining a permutation matrix $X = (X_{ij})$, which minimizes

$$\sum_{\ell=1}^{m} d_p(A^\ell, X) = \sum_{\ell=1}^{m} \langle n(n-1) - \sum_{j=1}^{n} [P_\ell^+(j), X^+(j)\rangle + \langle P_\ell^-(j), X^-(j)\rangle]].$$

This is equivalent to the linear assignment problem

$$\text{Max} \sum_{\ell=1}^{m}\sum_{j=1}^{n}\sum_{i=1}^{n}\left[\left(\sum_{t=j+1}^{n} p_{it}\right)\left(\sum_{t=j+1}^{n} x_{it}\right) + \left(\sum_{t=1}^{j-1} p_{it}\right)\left(\sum_{t=1}^{j-1} x_{it}\right)\right],$$

Subject to

$$\sum_{i=1}^{n} x_{ij} = \sum_{j=1}^{n} x_{ij} = 1, \quad x_{ij} \geq 0 \quad \text{for all } i,j.$$

Object-Based Distance

In a fashion similar to that presented above, one can construct an alternative-based distance function.

Definition 3. The alternative i forward indicator vector $0^+(i)$ and the alternative i backward indicator vector $0^-(i)$ are those vectors whose kth components are given by

$$(0^+(i))_k = \begin{cases} 1 & \text{if alternative } k \text{ is ranked} \\ & \text{lower than alternative } i, \\ 0 & \text{otherwise,} \end{cases}$$

$$(0^-(i))_k = \begin{cases} 1 & \text{if alternative } k \text{ is ranked} \\ & \text{higher than alternative } i, \\ 0 & \text{otherwise.} \end{cases}$$

The consensus model arising from this object based extension of Blin is the quadratic assignment problem:

Maximize

$$\sum_{\ell=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n} p_{ik}^{-\ell} x_{ij} \sum_{t=j+1}^{n} x_{kt},$$

Subject to

$$\sum_{i=1}^{n} x_{ij} = \sum_{j=1}^{n} x_{ij} = 1, \quad x_{ij} \geq 0 \quad \text{for all } i,j,$$

where

$$\bar{p}_{ik}^{\ell} = \sum_{t=j_i(\ell)+1}^{n} p_{kt}^{\ell}.$$

Some Comparisons

An issue which a number of authors have addressed has to do with the likelihood of different criteria giving rise to the same outcome (same winner or same consensus ranking). Fishburn, for example, has carried out a simulation study comparing Borda's method with that of Copeland. In this particular study, various combinations (n, m) were examined (n = number of voters and m = number of alternatives) from $n = 3-21$ and $m = 3-9$. For each such combination, 1000 cases were generated. A uniform distribution of ranked votes was assumed in carrying out the simulations. In comparing the two consensus methods, the issue was whether the winning candidates matched (2nd, 3rd, ..., etc. place standings were not compared).

Fishburn has found, for example, in the case of 21 voters and 3 alternatives, that in 81.8% of the 1000 cases examined, all winners via Borda (i.e., the set of alternatives tied for 1st place) were also the winners in Copeland and vice versa. In 12.8% of the cases at least some Borda winner matched a Copeland winner (but not all winners under Borda matched all Copeland winners). Finally, in the remaining 5.4% of the 1000 cases, no Borda winner was a Copeland winner.

The 3-Alternative Case

In the 3-alternative case, the six possible linear ordering of the 3 alternatives are assumed to follow a Dirichlet distribution. Figure 4 illustrates the shape of this distribution for $n = 3$. This single peaked distribution is particularly instructive in that there is a relative independence among options. At the same time, it is sufficiently general to permit a number of possible shapes.

For the 3 alternative case, and under the assumption of Dirichlet distributed rankings, it can be shown that the pairwise majority rule model always results in a transitive ranking. Thus, Copeland and majority rule are equivalent. It can also be shown that all of the aforementioned runoff or elimination methods will yield the same consensus ranking as well.

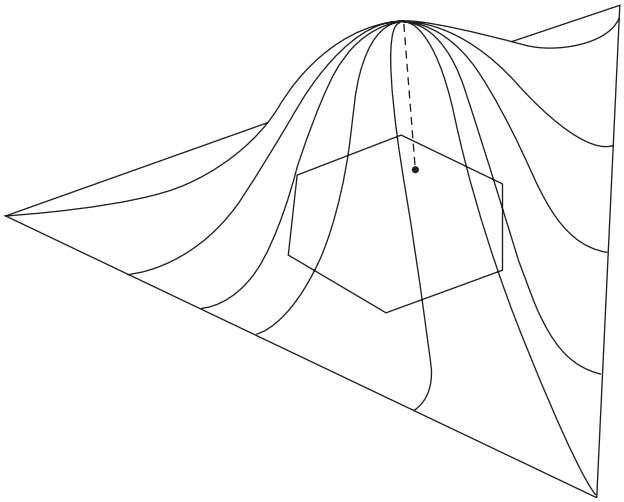


Figure 4 The Dirichlet distribution.

The General Case

Cook *et al*. examine the general case for n alternatives. For this case, let $\{R^\ell\}_{\ell=1}^{n!}$ denote the space of ordinal rankings, and $\{P^\ell\}_{\ell=1}^{n!}$ the corresponding proportions of voters (probabilities of the $n!$ rankings being chosen). Further, define the index set $M_{ij} = \{\ell \,|\, a_i^\ell < a_j^\ell\}$. Here, a_i^ℓ denotes the position assigned to alternative i by the ℓth voter. That is M_{ij} is the set of all rankings in which alternative i is preferred to alternative j. Note that $M_{ij} 1 M_{ji} = \phi$ and $M_{ij} \chi M_{ji} = \{1, 2, \ldots, n!\}$. With this notation, voter responses are said to be transitive if there exists an ordering of the n alternatives (assume this is say the natural ordering $(1, 2, \ldots, n)$) such that

$$\sum_{\ell \in M_{ij}} P^\ell > \sum_{\ell \in M_{ji}} P^\ell \quad \text{for all } i,j \text{ where } i<j.$$

Clearly, if such a transitivity property holds, then the consensus under pairwise majority rule (and Copeland's model) is the natural ordering

$$(r_1, r_2, \ldots, r_n) = (1, 2, \ldots, n).$$

In order to evaluate the Borda model, it is necessary to define another form of transitivity. Partition the index set M_{ij} into a series of different sets or levels. Formally, we define for alternatives i and j $(j\ i)$ the set of difference c (or level c) by

$$L_c = \{\ell \,|\, a_i^\ell - a_j^\ell = c\}.$$

Note that $M_{ij} = \cup_{c=-1}^{1-n} L_c$ and $M_{ji} = \cup_{c=1}^{n-1} L_c$.

Definition 4. Voter responses are said to be weighted transitive if there exists an ordering of the n objects such that

$$\sum_{c=1}^{n-1} c \sum_{\ell \in L_{(-c)}} P^\ell > \sum_{c=1}^{n-1} c \sum_{\ell \in L_c} P^\ell,$$

for all i, j where $i < j$.

Theorem 1. Under the condition that voter responses follow a Dirichlet distribution, transitivity on levels is always present, and Simple Majority Rule, Borda's method and Copelands method are equivalent.

Concluding Comments

Ordinal data in day to day decision making settings is a naturally occurring phenomenon. To deal with such data, numerous formats for its representation and models for aggregating responses have been developed. Many papers and books have been written on this subject including Cook and Kress. Many of the methods for dealing with data of this type have arisen over centuries in parliamentary voting settings. Others find their genesis in mathematical and social science settings, where axiomatic structures have been advanced to properly characterize consensus of opinions. This article attempts to capture some of these ideas.

See Also the Following Articles

Condorcet • Measurement Theory

Further Reading

Armstrong, R. D., Cook, W. D., and Seiford, L. (1982). Priority ranking and consensus formation: The case of ties. *Management Sci.* **28,** 638–645.

Blin, J. M., and Whinston, A. B. (1974). A note on majority rule under transitivity constraints. *Management Sci.* **20**(11), 1439–1440.

Blin, J. M. (1976). A linear assignment formulation of the multiattribute decision problem. *Rev. Automat. Inform. Recherche Operation.* **10,** 21–23.

Borda, J. C. (1781). Memoire sur les Elections au Scrutin. *Hist. Acad. Roy. Sci.*, Paris.

Condorcet, M. (1785). *Essai sur L'Applicaation de L'Analyse a la Probabilite des Decisions Rendues a la Pluralite des Voix.* Paris.

Cook, W. D., and Seiford, L. (1978). Priority ranking and consensus formation. *Management Sci.* **24,** 1721–1732.

Cook, W. D., and Kress, M. (1992). *Ordinal Information and Preference Structures: Decision Models and Applications.* Prentice-Hall, Englewood Cliffs, NJ.

Cook, W. D., Seiford, L., and Warner, S. (1983). Preference ranking models: Conditions for equivalence. *J. Math. Soc.* **9,** 125–127.

Copeland, A. (1945). John Von Newman and Oskar Morgenstern's theory of games and economic behavior. *Bull. Am. Math. Soc.* **51,** 498–504.

Fishburn, P. C. (1971). A comparative analysis of group decision methods. *Behav. Sci.* **16,** 538–544.

Kemeny, J. G., and Snell, L. J. (1962). Preference ranking: An axiomatic approach. *Mathematical Models in the Social Sciences*, pp. 9–23. Ginn, Boston.

Kendall, M. (1962). *Rank Correlation Methods*, 3rd Ed. Hafner, New York.

Niemi, R. G., and Weisberg, H. F. (1968). A mathematical solution for the probability of the paradox of voting. *Behav. Sci.* **13,** 317–323.

Zermello, E. (1926). Die Berechnung der Twinier-Ergebnisse als ein Maximum Problem de Wahrscheinlichkeitsrechnung. *Mathe. Zeitschrift*, 436–460.

Rapid Assessment Process

James Beebe
Gonzaga University, Spokane, Washington, USA

Glossary

bogus empowerment Letting people think they have control over outcomes and the power to act on their own judgments when they actually do not have this control or power; occurs whenever someone is asked for their input but there are no intentions of using it.

ethnography A descriptive study of an intact cultural or social group or an individual or individuals within the group, based primarily on participant observation and open-ended interviews; based on learning from people as opposed to studying people.

iterative process A process in which replications of a cycle produce results that approximate the desired result more and more closely. For rapid assessment, the process describes the cycle of data analysis and data collection, designed to produce a preliminary understanding of a situation from an insider's perspective.

participants Persons interviewed as part of the rapid assessment process. The term can be used interchangeably with "informants" or "respondents" when these terms are not modified with the words "individual" or "key." The term "subjects" is generally avoided.

rapid appraisal, rapid rural assessment, rapid rural appraisal (RRA), participatory rural appraisal (PRA) Different types of rapid qualitative research based on small multidisciplinary teams using semistructured interviews and direct observations to collect information in processes that can be completed in less than 6 weeks. As often implemented, these approaches may lack some of the methodological rigor of rapid assessment.

rapid assessment process (RAP), rapid assessment Intensive, team-based qualitative inquiry using triangulation, iterative data analysis, and additional data collection to quickly develop a preliminary understanding of a situation from the insider's perspective. Rigor is achieved by the use of multidisciplinary teams for both data collection and analysis; explicit use of an iterative process for data collection, analysis, and additional data collection; a defined role for "insiders" in the research team; member checking; documentation of the process, and attention to ethics.

triangulation A term from navigation and physical surveying that describes an operation for determining a position by use of bearings from two known fixed points. Triangulation is used as a metaphor by social scientists for the use of data from different sources, the use of several different researchers, the use of multiple perspectives to interpret a single set of data, and the use of multiple methods to study a single problem.

Rapid assessment allows a team of at least two researchers to develop a preliminary understanding of a complicated situation in which issues are not yet well defined. Rapid assessment is especially relevant when an insider's perspective is needed, and there is not sufficient time or other resources for long-term, traditional qualitative research. Rapid assessment is a type of participatory action research. It shares many of the characteristics of ethnographic research. However, rapid assessment, uses intensive, team interaction and multiple cycles of data collection followed by data review/analysis, instead of the prolonged fieldwork normally associated with traditional qualitative research. Results can be used for planning, monitoring, and evaluating activities and for the design of additional research. Rapid assessment will almost always produce results in a fraction of the time and at less cost than is required by traditional qualitative research.

Introduction

Rapid assessment is defined as intensive, team-based qualitative inquiry using triangulation, iterative data analysis, and additional data collection to quickly develop

a preliminary understanding of a situation from the insider's perspective. Data is collected by talking with people and getting them to tell their stories. The acronym for the rapid assessment process, "RAP," expresses well the need to communicate with participants using their vocabulary and rhythm and one definition of "rap" is "to talk freely and frankly." "Rapid" means a minimum of 4–5 days and, in most situations, a maximum of 6 weeks. Responding to the need for almost immediate results involves compromises and requires special attention to methodological rigor. "Process" means a series of actions or operations conducive to an end. A process approach suggests that at least as much attention is given to the way results are obtained as to the results. References to rapid appraisal, rapid rural assessment, rapid rural appraisal, participatory rural appraisal, and the acronym RAP have been widely used in the literature to identify different rapid qualitative research methods. Not everything labeled "rapid assessment" meets the methodological rigor of the rapid assessment process.

Evolution, Current Status, and Relationship to Other Approaches to Rapid Qualitative Research

Rapid assessment has its roots in farming systems research of the late 1970s. Farming systems research was based on a holistic consideration of people along with their plants and livestock and started with the assumption that local systems consisted of mutually related elements that constitute a whole. A systems approach initially considers all aspects of a local situation, but quickly moves toward the definition of a model that focuses on only the most important elements and their relationship to each other from the perspective of the local participants. In the initial research on local farming systems, neither research tourism or questionnaire survey research were able to produce solid and timely results. A 1979 paper by Peter Hildebrand described a farming systems research approach based on teamwork called "sondeo." His paper, along with others, had been presented at the 1979 Rapid Rural Appraisal conference at the Institute of Development Studies at the University of Sussex. Because of the title of the conference, "rapid rural appraisal," along with variants, including "rapid appraisal" and "rapid assessment," became associated with rapid qualitative team-based research. The publication of the *Proceeding of the 1985 International Conference on Rapid Rural Appraisals* by the Khon Kean University in 1987 made information available to a wider audience. Robert Chambers and his colleagues at the Institute of Development Studies have been at the forefront of formulating and disseminating information on rapid research methods for the past two decades. Significant publications have included Chamber's *Shortcut and Participatory Methods for Gaining Social Information for Projects*, Krishna Kumar's *Rapid Appraisal Methods*, Nevin Scrimshaw and Gary Gleason's *Rapid Assessment Procedure: Qualitative Methodologies for Planning and Evaluation of Health Related Programmes*, John Van Willigen and Timothy Finan's *Sounding: Rapid and Reliable Research Methods for Practicing Anthropologists*, and James Beebe's *Basic Concepts and Techniques of Rapid Appraisal* and *Rapid Assessment Process: An Introduction*. Despite differences in details, all of the different rapid qualitative research methods are based on small multidisciplinary teams using semistructured interviews, direct observation, and other techniques to collect information, with the entire process being completed in less than 6 weeks.

There is increasing use of rapid qualitative research methods as planning and evaluation tools in a variety of fields. More than 50 examples illustrate the range of topics that have been investigated using rapid research methods in areas as diverse as agriculture, community and rural development, conservation and natural resources, health and family planning, and marketing. These examples also illustrate how groups and organizations that have used these methods, and the extent to which the methods have been adapted to meet different needs. Increased recognition of a need for qualitative research has encouraged the use of rapid research methods as resources available for traditional long-term qualitative research have declined. A growing consensus on the need for participatory approaches to research activities has also contributed to expansion of rapid research methods. Chambers has observed that anticipatory language has become obligatory "donor-speak." Current issues concern the extent to which rapid qualitative methods should be participatory as well as how participation is implemented. The contrast of rapid assessment with participatory action research, and especially participatory rural appraisal (PRA) associated with Robert Chambers, illustrates these issues. In response to different needs, these approaches have evolved in different ways, with participatory action research focusing more on the empowerment of local participants who do research to satisfy local needs, and rapid assessment focusing more on methodological rigor and the involvement of decision makers at different levels. The two approaches complement each other and share methodological techniques. Rapid assessment differs from numerous forms of participatory action research by explicitly recognizing that in many situations local participants do not have control over the resources necessary for change. Rapid assessment intentionally involves decision makers in the research process and attempts to ensure sufficient rigor for credibility with decision makers.

The lack of methodological rigor has been one factor in the limited use of rapid qualitative research methods for research published in peer-reviewed journals. Publication in peer-reviewed journals continues to be viewed as validating research that contributes to the body of knowledge. Chambers has lamented that calling research methods rapid has "been used to justify and legitimize sloppy, biased, rushed, and unself-critical work." Rapid assessment explicitly deals with methodological rigor.

Questions raised by some anthropologists regarding the legitimacy of rapid research methods have limited their use. Almost all descriptions of ethnography refer to a requirement for prolonged fieldwork. There has been an unfortunate tendency to equate time with quality and to dismiss rapid results with pejorative phrases such as "quick and dirty." The case for prolonged fieldwork advanced by anthropologists such as H. Russell Bernard and Harry F. Wolcott is based on tradition and the argument that it takes times to develop intellectualized competence in another culture, to be accepted, to develop rapport, to be included in gossip, and to get information about social change.

Basic Characteristics

"Rapid" means producing results in 1 to 6 weeks. There is growing consensus among practitioners that completion of data collection, data analysis, additional data collection and the preparation of a report requires at least 5 days. "Rapid" does not mean "rushed." Schedules must be flexible to allow the team to take advantage of the unanticipated. Rapid assessment uses the techniques and shares many of the characteristics of ethnography, but differs in two important ways: (1) more than one researcher is always involved in data collection, with data triangulation based on teamwork, and (2) more than one researcher is involved in an iterative approach to data analysis and additional data collection. The intensive teamwork is necessary because of the shortened fieldwork.

Data Collection: Triangulation and Intensive Teamwork

The Team and Teamwork

Between two and six individuals are usually on the RAP team, and teams need to be multidisciplinary, diverse, and include at least one "insider" as well as "outsiders." The assumption is that two sets of eyes and ears are better than one and that the use of different techniques can help make the best use of the extra eyes and ears as part of intensive teamwork. The assumption is also that two heads are better than one in figuring out what has been seen and heard and for deciding on the next steps. Sensitivity to cultural differences is essential and team diversity improves cultural sensitivity and helps establish credibility with local communities. Whereas traditional research methodology has focused on helping outsiders to better understand insiders' knowledge, there is a growing appreciation of the role insiders should play in the design, implementation, and publication of research. The ability of the RAP team to quickly develop a preliminary understanding of a situation, from the perspective of the local participants, is facilitated by having an insider on the team. The insider needs to be a full team member and must be involved in planning, data collection, data analysis, and the preparation of the report.

Rapid assessment depends on teamwork and cannot be done by one person. All team members should be involved in data collection and data analysis, including the preparation of the report. Teamwork by a multidisciplinary team increases sensitivity to the insiders' categories and definitions. Because of the importance of team interaction, the RAP team should be together most of the time.

Directed Conversation (Semistructured Interviews)

The most important way of learning about local conditions is to get local people to tell what they know. The goal is to get people to talk on a subject and not simply answer questions. This process is often identified as a "semistructured interview," but it is better thought of as directed conversation. Directed group discussions involve the entire team interacting with each other as well as with the respondent. This is not sequential interviewing by individual team members. Relaxed, semistructured interviewing provides respondents with time to think and helps elicit stories.

Experience has shown the value of opening the conversation with a carefully articulated "grand tour" question. All RAP team members need to be active listeners. The grunts and noises the team members make as active listeners, such as the "umms," "uh-huhs," and "mmmmmms," improve rapport and encourage people to speak longer. The conversation is kept moving and on-track with probes that do not inject the views of the team. Nondirective probes are culture specific and need to be identified prior to conducting the first interview. Examples of nondirective probes that work in a United States cultural setting might include "Give me a description of . . . Tell me what goes on when you . . . Describe what it's like to . . . Say more, keep talking." The RAP team will likely want to develop short guidelines based on a few big issues. Guidelines are used instead of a list of questions prepared in advance of the conversation. Despite the guidelines, the direction of the study should emerge as information is collection. Guidelines should be viewed as a reminder of issues that should

not be missed, rather than as an agenda to be diligently worked through.

The RAP team purposefully selects individuals for directed conservations. Consistent with qualitative research, they are identified as "participants" or "respondents" and not as "subjects." They are not a sample. They are selected not because they are believed to be average, but because they are believed to represent the diversity found in the local situation. The RAP team should seek out the poorer, less articulate, more upset, and those least like the members of the RAP team.

Other Techniques for Data Triangulation

In addition to semistructured interviewing other specific research techniques for use in a given rapid assessment are chosen from among a wide range of techniques, based on the specific topic being investigated and the resources available. Observations and team interaction with respondents, based on what is seen and heard, are necessary. All interviews should be conducted in a relevant setting where listening can be combined with observing. Anthropologists note that participant observation can range from actually living and working in the field as a member of group over an extended period of time, to simply being an observer. The essential requirements for participant observation are that people must feel comfortable with the presence of the team. Even during a rapid assessment, there are opportunities for observing. Team members should try to be present at relevant times outside of normal business hours, including early morning and late evenings and weekends. Sharing of meals with respondents provides opportunities for combining observations with informal discussions and follow-up.

Groups of respondents as well as individual respondents can be interviewed using techniques associated with focus group research. Folktales, myths, songs, proverbs, riddles, and jokes can provide insights to local situations. Drawing diagrams and "rich pictures" allows respondents to express themselves in ways that are often more valid than talk. Maps drawn by respondents can be used for collecting data and planning action.

Field Notes, Transcripts, and Logs

Among social scientists, there is considerable disagreement on what constitutes field notes and how they should be organized. Because more than one person is involved in collecting and processing field notes, all parties need to agree on the format. The term "field notes" probably should be reserved for the usually handwritten notes prepared as data is being collected. Field notes must be readable by the person who wrote them and should clearly differentiate between (1) what has actually been said by the respondents and observed by the team and (2) comments by the researchers, including reflections, thoughts about conclusions, and other notes. Field notes need to include detailed observations and direct quotations, as opposed to summaries. Carefully done field notes can help the team avoid imputing false meaning. Field notes should include information about the interview process and should identify things that need to change during subsequent interviews. All team members should take notes, including the team member who is directing the conversation at any given moment. Note taking becomes a way to control the speed of the conservation and gives both the team members and the respondent time to think. Field notes should also include descriptions of the settings, who else was present, the overall demeanor of the respondent, and nonverbal communications such as a smile or yawn.

Interviews should always be taped unless the respondents specifically object to the use of a tape recorder. The written version of what has been recorded is usually identified as a "transcript." Ideally, transcripts should be made available to the team within 24 hours. Transcripts are not field notes. Even when there are not plans to transcribe interviews, tapes can be used to fill in missing information. Tape recorders should be expected to fail and should never be used in place of note taking.

The "log" is a combination of the field notes, the transcripts, and the reflections of the researchers in a format ready for analysis. When transcripts are not available, the log created from the different field notes of the team. Logs are most useful if they are typed (double-spaced), with every sentence beginning on a new line and with very wide margins on both sides. Many researchers place codes in the left margin and comments in the right margin. Anything in the log that is not a direct observation should be clearly identified. Logs should be prepared within 24 hours of the interview. Preparation of the log allows the team to carefully examine what has been heard and to consider explicitly the insider's perspective. This review also provides an opportunity to consider changes that need to be made in the next round of data collection, including changes in the way the interview was implemented.

Iterative Analysis and Additional Data Collection

Rapid assessment explicitly divides research time between blocks used for collecting information and blocks when the RAP team does data analysis and considers changes in the next round of data collection. Beginning on day 1, time is scheduled for team interaction. Usually more time is spent on team interaction than on data collection. An iterative process is defined as a process in which replications of a cycle produce results that approximate the desired result more and more closely. The constant shifting between data analysis and additional

data collection is an iterative, or recursive, process. For rapid assessment, the replication of the process of data collection, followed by analysis and additional data collection, contributes to the goal of understanding the situation under investigation from the perspective of the local participants.

Just as there is no one best way for data collection, there is no one best way for analysis. An approach that has worked for many rapid assessments and that can serve as a beginning point for modification is based on Matthew Miles and A. Michael Huberman's model. This model involves (1) coding the data and adding marginal remarks, (2) displaying the data, and (3) drawing conclusions. For rapid assessment, as for qualitative research in general, analysis is an ongoing process that begins with, or even before, the first round of data collection and continues through the preparation of the report.

Coding

The logs, combining the field notes of all members of the RAP team and, when available, the transcripts of the directed conversations, are the source of the data for understanding the situation from the perspective of the respondents. The first step in the analysis process is to read the logs several times. The next step is dividing the log into thought units and applying codes to these units. A unit of thought may be a sentence, paragraph, several paragraphs, or even an individual word. Coding can be thought of as cutting the logs into strips and placing the strips into piles. The codes are the labels that the RAP team gives to the individual piles. Developing a coding system is based on trial and error. The coding system should remain flexible enough that codes can be added, combined, and removed as needed. The RAP team looks for threads that tie together bits of data and seeks to identify recurring words or phrases. These words often become the labels for the codes. Not everything in the log is coded and a single unit of thought will often have multiple codes. The team can always change the codes. Experience suggests it is better to start with only five or six codes and to then subdivide these when necessary.

After codes have been assigned, the next step is to consolidate everything relating to each code. Materials from all interviews, as well as other observations relating to codes, can then be considered together. Cut-and-paste functions of word-processing programs facilitate this process, but should be done with care to ensure that thought units are associated with specific interviews. Adding margin remarks is closely related to the coding process and occurs both before and after materials have been rearranged according to their codes. Margin remarks are usually written into the margin of the log after it has been typed and can include RAP team reactions, questions about the meaning of statements, and notes about connections between parts of the data, etc.

Data Display

The second aspect of analysis is data display. Development of data displays should begin during the coding and continued throughout the research process. Data displays are often drawn on large sheets of paper such as flip charts, and include matrices, graphs, words, and drawings of objects linked by lines, suggesting relationships. "Rich pictures" can be used as data displays.

Drawing Conclusions

The third element in the data analysis process is conclusion drawing and verification. Miles and Huberman suggest that people can make sense of the most chaotic events, but that the critical question is whether the meanings they find in qualitative data are valid. They identify 13 tactics for generating meaning. The first six of these are descriptive and provide a good beginning point for rapid assessment. The first three identify connections and concern (1) patterns and themes, (2) seeking plausibility, and (3) clustering. The next three sharpen understanding and include (4) metaphor making, (5) counting, and (6) making contrasts and comparisons.

Member Checking

Before conclusions are final, the RAP team should share them with the people who provided the information and check for their agreement. This can be done either formally or informally. The local people can provide corrections to facts and their own interpretation of the situation. This process is often called "member checking."

Preparation of the Report by the Entire RAP Team

The joint preparation of the rapid assessment report by the entire team continues the intensive team interaction. The preparation of the report should start while there is still time for additional data collection. The involvement of the entire team, including the local members, provides the report with more depth than is often found in reports prepared by a single individual.

The RAP Sheet

One of the major challenges for rapid assessment is promoting flexibility and creativity without diminishing rigor. Rigor can be enhanced by agreement on basic principles, and then documentation, as part of the rapid assessment report, of the specific techniques used. The documentation on what was actually done can be summarized in a checklist, called a "RAP sheet." The RAP sheet is attached to the report and can include information about team members, hours of data collecting, hours of team interaction discussing the data, types of information collected by direct observations, the number of individuals interviewed, methods of selection of respondents, places of

interviews, and specific information about the diversity of the respondents interviewed. The RAP sheet allows the reader of the report to judge the quality of the work.

How Much is Enough?

Every RAP team will face the question of how much information is needed, how much time should be spent in the field, and how many interviews are sufficient. The data is sufficient when themes begin to repeat. Experience shows that when interviews have been relaxed and sufficient time has been spent by the team making sense out of the data that are collected, themes often began to repeat after 8 or 10 interviews. The actual number of interviews needed can be significantly more than this, but is usually not less. It is usually easier for the team to recognize when data is insufficient than when enough data has been collected.

Ethics, Participation, and Bogus Empowerment

There is widespread consensus on the value of participation. As discussed earlier, rapid assessment is a participatory research approach. However, ethical issues related to participation are often ignored. These issues are almost always aggravated by an inappropriate belief that problems must be identified and solved at the local level, without the involvement of outsiders. Unlike some forms of participatory action research, RAP does not assume that the study population can unilaterally solve its own problems. Decision makers and authorities higher in the system who control resources often need to be members of the RAP team in order to feel committed to the outcome of the research and as stakeholders are more inclined to take action. Even when outside decision makers are not part of the research effort, it is critical that the research effort be designed with sufficient rigor to allow outsiders to have confidence in the results.

The most serious negative consequence of an excessive focus on participation is shifting responsibility for change onto the poor and outsiders relinquishing their responsibilities. Even when such a shift does not occur, discussing problems can raise unrealistic expectations in local communities that the problems will be addressed. Closely related to raising unrealistic expectations is what Joanne Ciulla has called "bogus empowerment." Several aspects of bogus empowerment are especially relevant to rapid assessment. Both the RAP team and those responsible for bringing in the RAP team must keep their promises. The best way to do this is to make promises that can be kept. Everyone involved in rapid assessment needs to be clear about the often limited power of the local community and must avoid the temptation of engaging in hyperbole about the democratic nature of the situation. Rapid assessment also can be an accessory to bogus empowerment by encouraging people to believe falsely that actions will be taken in response to their input.

Learning to RAP

For individuals who have had limited experience with qualitative techniques, there is a need to provide a strong rationale for, and an introduction to, qualitative research. For individuals with a background in qualitative research, there is a need to enhance their understanding of ways in which rapid assessment differs from traditional approaches. There is general consensus among practitioners that rapid assessment is best learned while participating as a team member with someone with experience. However, because rapid research methods are "organized common sense," they can be self-taught. From reading reports by others, it is possible to learn about the methodology and to acquire realistic expectations. Copies of reports, information on the availability of workshops about rapid assessment, additional references, and useful tools are available on the rapid assessment website (www.rapidassessment.net). Rapid assessment uses many of the techniques of traditional qualitative research, and reference books on qualitative research can be very useful. To begin experimenting with rapid assessment, all that the RAP team needs to remember is that the goal is to talk with people and to get them to tell their stories, as opposed to answering the questions of the team.

A combination of brief introductions to qualitative research techniques and a willingness to listen intently and genuine respect for others can help a RAP team get started. Everyone on the RAP team needs to recognize that (1) they do not know enough in advance to even know what questions to ask and (2) they do not know enough to provide answers, but (3) they do know enough to want to empower others to solve their own problems.

Plausible Directions for the Future

The accelerating rate of change in the world and a lessening of financial support for long-term research have driven the increasing interest in rapid assessment. Limited resources often are better used to provide needed services rather than to support traditional research. There is growing recognition of the relationship between leadership and the creation of organizations based on participation. Rapid assessment can be an important tool for leaders because of its ability to promote participation and to help create learning

organizations. Rapid assessment has tremendous potential. The major challenges to rapid assessment include overselling it, confusing "rapid" with "rushed," and failing to implement it rigorously. If these challenges can be met, rapid assessment may be an idea whose time has come.

See Also the Following Articles

Coding Variables • Data Collection, Primary vs. Secondary • Ethical Issues, Overview • Ethnography • Field Experimentation

Further Reading

Beebe, J. (1995). Basic concepts and techniques of rapid appraisal. *Human Organiz.* **54**(1), 42–51.

Beebe, J. (2001). *Rapid Assessment Process: An Introduction.* AltaMira, Walnut Creek, CA.

Bernard, H. R. (1995). *Research Methods in Anthropology: Qualitative and Quantitative Approaches.* 2nd Ed. AltaMira, Walnut Creek, CA.

Chambers, R. (1991). Shortcut and participatory methods for gaining social information for projects. In *Putting People First: Sociological Variables in Rural Development,* 2nd Ed. (M. M. Cernea, ed.), pp. 515–537. Oxford University Press, World Bank, Washington, D.C.

Chambers, R. (1996). *Introduction to Participatory Approaches and Methodologies.* Available on the Internet at www.ids.ac.uk

Ciulla, J. B. (1998). Leadership and the problem of bogus empowerment. In *Ethics: The Heart of Leadership* (J. B. Ciulla, ed.), pp. 63–86. Quorum Books, Westport, CT.

Creswell, J. W. (1998). *Qualitative Inquiry and Research Design: Choosing Among Five Traditions.* Sage, Thousand Oaks, CA.

Ely, M., Anzul, M., Friedman, T., Garner, D., and McCormack, A. S. (1991). *Doing Qualitative Research: Circles within Circles.* Falmer Press, Bristol, PA.

Fetterman, D. M. (1998). *Ethnography: Step by Step.* 2nd Ed. Sage, Thousand Oaks, CA.

Hildebrand, P. E. (1979). *Summary of the Sondeo Methodology Used by ICTA.* Paper presented at the Rapid Rural Appraisal Conference at the Institute of Development Studies. University of Sussex, Brighton, England.

Khon Kaen University. (1987). *Rural Systems Research and Framing Systems Research Projects.* Proceeding of the 1985 International Conference on Rapid Rural Appraisal. Khon Kaen, Thailand.

Kumar, K. (1993). *Rapid Appraisal Methods.* World Bank, Washington, D.C.

Scrimshaw, N., and Gleason, G. R. (1992). *Rapid Assessment Procedures: Qualitative Methodologies for Planning and Evaluation of Health Related Programmes.* International Nutrition Foundation for Developing Countries, Boston, MA.

Van Willigen, J., and Finan, T. L. (1991). *Soundings: Rapid and Reliable Research Methods for Practicing Anthropologists.* American Anthropological Association, Washington, D.C.

Wolcott, H. F. (1995). *The Art of Fieldwork.* AltaMira, Walnut Creek, CA.

Rare Events Research

Will Lowe
Harvard University, Cambridge, Massachusetts, USA

Glossary

relative risk The proportional increase in the probability of an event between two values of an independent variable.
response-based sampling A biased sampling scheme in which data are selected on the dependent variable.
risk difference The difference in the probability of an event between two values of an independent variable.

State collapse, the outbreak of war, and cases of rare diseases are events that occur very infrequently in a population, but are of considerable interest when they do. Rare events cause particular problems for statistical models, such as logistic regression, that are used to understand and predict them: a small random sample is unlikely to contain enough instances of the rare event to make reliable inferences, but a sample large enough to ensure a reasonable number of rare events may be prohibitively expensive to collect. Schemes that ensure samples with a balance of event types are easily found, but except in special circumstances, applying standard estimators to data generated this way leads to inconsistency. In addition to sampling scheme issues, maximum likelihood estimators are well known to be biased in samples of less than 200 observations. Less well known is that small sample problems can occur when one value of the dependent variable is rare, even in very large data sets. Bias and inconsistency in parameter estimation lead to unsound estimates of the quantities of substantive interest: estimates of conditional probability, relative risk, and risk difference. Rare events research methods are designed to resolve these closely related issues by showing how to fit models while maintaining consistency under biased sampling schemes and by generating appropriate small sample corrections.

Introduction

We focus on the following simple system. $P(Y, X)$ is a joint distribution over event types and covariate values. $P(Y)$ is a discrete distribution over M possible event types, any one of which may be rare. In the case of predicting the outbreak of war, $M = 2$ and the probability of war $P(Y = 1)$ is much smaller than the probability of peace $P(Y = 0)$. $P(X)$ is a distribution of possibly real-valued independent variables thought to be useful for predicting or explaining Y. Although $P(X)$ is typically unknown, we are willing to entertain a parameterized model of the relationship between Y and X, $P(Y \mid X; \beta)$. Because this model very often takes the form of a multinomial logit, we shall concentrate on results for this model class.

Substantive interest centers on the fitted model's estimates of $\pi = P(Y \mid X)$ because these quantify the effect of covariate changes on Y, both by themselves and as part of relative risk and risk difference statistics. We are interested in consistent, and preferably efficient and unbiased, estimates of β primarily as a means to this end.

In the next section the choice of biased sampling schemes available for rare events researchers is reviewed. The following section considers a range of estimators for β that are consistent under biased sampling schemes. In addition to sampling scheme issues, rare events data suffer from finite sample bias in maximum likelihood estimation. We consider two debiasing parameter estimators to address this problem. Separate issues arise in the estimation of π itself; three estimators for this task are considered. With a fitted model finally in hand, the final section shows how risk quantities can be estimated under varying amounts of information about $P(Y)$.

Sampling Schemes for Rare Events

In the following, the distribution of a sample under some sampling scheme is called H(Y, X). In the simplest case of simple random sampling from the population, H(Y, X) = P(Y, X). Alternatively, it may be more practical to define blocks of observations based on values of X and take subsamples within them. In this case, H(Y, X) = P($Y \mid X$) H(X) $\neq$ P(Y, X). Both of these schemes are exogenous sampling schemes. They share the property that maximizing the joint likelihood function P(Y, X; β) with respect to β is equivalent to maximizing only the conditional part, leading to

$$\text{argmax}(\beta): \sum_{i=1}^{N} \ln P(y_i \mid x_i; \beta). \tag{1}$$

This estimator yields maximum likelihood estimates with entirely classical properties. The equivalence holds when there is selection on X because although H(X) $\neq$ P(X), β does not depend on P(X), allowing P(X) to drop out.

Unfortunately exogenous sampling schemes can be highly inefficient for studying rare events. In small sample sizes, the rare events may not appear at all, or not sufficiently often to provide reliable parameter estimates. Indeed, the rarer the event, the more information is provided by its occurrence. In logistic regression, for example, the covariance for β is

$$\text{Var}(\hat{\beta}) = 1 \Big/ \sum_{i=1}^{N} \pi_i(1-\pi_i)\mathbf{x}\mathbf{x}'. \tag{2}$$

In a model with well-chosen covariates, π_i, and therefore $\pi_i(1-\pi_i)$, will be larger for rare events than for common ones. Thus, the variance shrinks more for rare observations, and does so in proportion to how rare they are. Sample sizes large enough to guarantee a reasonable number of rare events may also be prohibitively expensive to collect and will be unbalanced, exacerbating finite sample bias.

An alternative endogenous sampling scheme is to divide Y into blocks and sample within each block. Because Y is a decision, outcome, or response variable, sampling schemes that select on Y are referred to as response-based sampling. For response-based sampling, H(Y, X) = P($X \mid Y$)H(Y). This makes standard maximum likelihood approaches inconsistent. To see this, note that for a sample of size D,

$$P_D(Y \mid X) = P_D(X \mid Y)\frac{P_D(Y)}{P_D(X)} \tag{3}$$

converges in distribution to

$$P(X \mid Y)\frac{P(Y)}{P(X)} = P(Y \mid X). \tag{4}$$

The corresponding sample version

$$H_D(Y \mid X) = H_D(X \mid Y)\frac{H_D(Y)}{H_D(X)} \tag{5}$$

does not converge to

$$P(X \mid Y)\frac{P(Y)}{P(X)} = P(Y \mid X). \tag{6}$$

This is because H(Y) $\neq$ P(Y) by design and will not converge to it, irrespective of sample size. Consequently, P($Y \mid X$) cannot be consistently estimated from response-sampled data without a correction that makes use of information about P(Y) from outside the sample.

Estimators for Response-Based Samples

A wide range of estimators are available for maximum likelihood estimation in response-based samples. However, in the special but substantively important case of multiplicative intercept models such as logistic regression, these choices reduce to two: weighting and intercept correction.

The Weighting Estimator

The weighted exogenous sampling maximum likelihood estimator (WESML) replaces Eq. (1) with

$$\text{argmax}(\beta): \sum_{i=1}^{N} w_i \ln P(y_i \mid x_i; \beta), \tag{7}$$

where w_i = P(y_i)/H(y_i). With a slight abuse of notation, P(y_i) refers to the probability of seeing an observation with the same Y-value as that of the ith sample value. H(y_i) is the corresponding probability under a response-based sampling scheme. Thus, when $M = 2$, and one-quarter of the sample data are expected to have $Y = 1$, P(y_i) = 0.25 whenever $y_i = 1$ and 0.75 otherwise.

The WESML estimator is consistent and asymptotically normal. Replacing the expected proportions from the response-based sampling scheme H(y_i) with their sample values makes the estimator more efficient. Although WESML is essentially a weighted version of the regular maximum likelihood approach, regular standard errors should be computed using White's heteroskedasticity-consistent variance matrix.

The Intercept Correction Estimator

Three other estimators are available for response-based samples: Manski and McFadden have provided a conditional maximum likelihood, and described a more efficient full information approach. A generalized method of moments estimator is also available. All these methods are consistent and asymptotically efficient. Rather than pursue the detail of each estimators, it is most useful to note that for multiplicative intercept models, the three estimators coincide. Morever, they may be computed by fitting the model to response-based sample data as if they were exogenously sampled, and then adjusting the model's intercept parameter. In the case of $M = 2$, the new intercept takes a particularly simple form:

$$\hat{\beta}_0 - \ln\left[\left(\frac{\mathrm{P}(Y=0)}{\mathrm{P}(Y=1)}\right)\left(\frac{\mathrm{H}(Y=1)}{\mathrm{H}(Y=0)}\right)\right]. \qquad (8)$$

In fact, in this class of models, non-intercept coefficients are consistently estimated under both exogenous and response-based sampling schemes. The intercept correction estimator is consistent and asymptotically efficient. By comparison, WESML is slightly less efficient but less sensitive to model mispecification. For this reason it may be preferable for social scientific applications.

Although both WESML and intercept correction assume perfect knowledge of P(Y), their statistical properties are the same when an independent sample is available from which P(Y) may be consistently estimated.

Response-Based Sample Design

Researchers using response-based sampling methods for rare events data should normally choose H(Y) to be an equal division among Y categories. For, although an optimal splitting value for rare and common response categories will always exist, it is essentially impossible to determine in advance. Luckily, Monte Carlo studies suggest that a sampling procedure giving each value of Y an equal number of cases in the sample is seldom far from optimal across a range of parameter values and estimators, including WESML. A balanced sample has the added advantage of minimizing finite sample problems in maximum likelihood estimation.

Finite Sample Corrections

Standard maximum likelihood estimators for β in multiplicative intercept models yield consistent and asymptotically efficient estimates. However, in finite samples they are biased. Two related methods for correcting finite sample bias in generalized linear models are available.

McCullagh and Nelder provided a somewhat complex but easily calculated direct correction to the maximum likelihood estimate of β by using a second regression. This correction has the advantage of reducing variance as well as bias. One disadvantage of this correction is that because it is applied after estimation has finished, it provides no protection from the infinite parameter values that arise from perfectly separable data, a particular risk for small samples and rare events.

Approximately the same debiasing effect can be obtained by fitting a model using Jeffreys' non-informative prior and taking the maximum *a posteriori* value of β. A non-Bayesian interpretation of this method is as the minimization of the penalized likelihood

$$\text{argmax}(\beta)\colon \ln \mathrm{P}(Y \mid X;\, \beta) + \frac{1}{2} \mid I(\beta) \mid, \qquad (9)$$

where $I(\beta)$ is the Fisher information matrix. This method guarantees finite parameter estimates even when data are perfectly separable.

Bias, like inconsistency in response-based sampling, primarily affects the intercept term. For example, in a model with a single fixed covariate coefficient and intercept, the intercept bias is approximately, $(\bar{\pi} - 0.5)/N\bar{\pi}(1-\bar{\pi})$, where $\bar{\pi}$ is the average estimated conditional probability in the sample. For events of probability less than one-half, this bias is negative. Consequently, the intercept, and therefore the estimated marginal probability of $Y = 1$, will be too small.

Whichever approach is taken, debiasing corrections should always be used in place of the standard maximum likelihood estimate.

Estimating P($Y \mid X$; β)

The preceding corrections affect parameter estimates. However, the aim of parameter estimation is to generate reasonable estimates of π conditioned on covariates. But there is a separate choice to make when estimating π. In a non-linear model such as logistic regression, simply inserting a debiased estimate of β into a logistic regression model does not generate a similarly unbiased estimate of π. This is a quite general problem with unbiased estimators; an unbiased estimator may not exist, and if it does, functions of unbiased estimator may not be unbiased themselves.

An Approximately Unbiased Estimator

In order to reduce bias in π_i it is necessary to subtract

$$C_i = (0.5 - \pi_i)\pi_i(1-\pi_i)\mathbf{x}'_i\mathrm{Var}(\beta)\mathbf{x}_i \qquad (10)$$

from each estimated conditional probability. However, considerably better estimators of π, in a mean square error sense, are available.

A Semi-Bayesian Estimator

An easily computable alternative is King and Zeng's semi-Bayesian estimator. Logistic regression assumes a linear predictor determines the mean value of an unobserved logistic distribution $p(Y^*)$. When this distribution generates a value greater than zero, a one is observed, otherwise a zero. π is then equal to $p(Y^* > 0)$. In a Bayesian treatment, uncertainty about β, as expressed in its posterior covariance, must be integrated out of estimates of π as

$$\pi = \int \mathrm{P}(y \mid x, \beta)\mathrm{P}(\beta)\, d\beta; \tag{11}$$

where P(β) is assumed to be normally distributed with the mean and variance of the distribution of $\hat{\beta}$. Identifying the sampling distribution of $\hat{\beta}$ with a posterior distribution leads to the same numerical results as when a non-informative prior is used, for example, using Firth's estimator. Of course, the interpretation is quite different, but here the estimator's performance is evaluated in terms of its sampling properties.

For distributions of β with significant variance, this integration leads to a wider distribution of Y^*, so a larger proportion of $p(Y^*)$'s support will be greater than zero than if a single value of β had been used. Consequently, the semi-Bayesian estimator increases the estimated value of π. Surprisingly, the amount of probability increase is approximately C_i. Intuitively, C_i grows with sampling distribution, and its direction is provided by the first term. Thus, for rare events, the debiasing estimator decreases π_i and the semi-Bayesian estimator increases it, in about equal degrees.

Of these estimators, the semi-Bayesian estimator, which trades bias for substantially decreased variance, is to be preferred on mean square error grounds. The unbiased estimator is preferred only if there is a compelling reason to demand unbiased estimation. Finally, Monte Carlo studies suggest that the maximum likelihood estimator, that is, uncorrected π, is superior to the unbiased estimator, inferior to the semi-Bayesian estimator in terms of mean square error, and provides a compromise between the two with respect to bias.

Risk Statistics

When a model of $\mathrm{P}(Y \mid X; \beta)$ is finally in hand, it can be used to compute quantities of substantive interest. The three most important quantities are the following:

- absolute risk: $\mathrm{P}(Y \mid X = b)$
- relative risk (RR): $\mathrm{P}(Y \mid X = b) / \mathrm{P}(Y \mid X = a)$
- risk difference (RD): $\mathrm{P}(Y \mid X = b) - \mathrm{P}(Y \mid X = a)$.

An important auxiliary quantity is the odds ratio (OR):

$$\mathrm{OR} = \frac{\mathrm{P}(Y = 1 \mid X = b)/\mathrm{P}(Y = 0 \mid X = b)}{\mathrm{P}(Y = 1 \mid X = a)/\mathrm{P}(Y = 0 \mid X = a)} \tag{12}$$

$$= \frac{\mathrm{P}(X = b \mid Y = 1)\mathrm{P}(X = a \mid Y = 0)}{\mathrm{P}(X = a \mid Y = 1)\mathrm{P}(X = b \mid Y = 0)}, \tag{13}$$

where the second line follows by Bayes theorem.

The Odds Ratio

Unlike the risk statistics, the odds ratio is not straightforward to interpret substantively, although this is often attempted. Also unlike the risk statistics, the odds ratio has the advantage of being calculable directly from response-based samples. The preceding reformulations are possible because it is invariant to addition to or rescaling of either marginal. Consequently, it is unaffected by sample distributions of Y that do not match their population. The odds ratio is also easy to extract from a fitted logistic regression model: $\mathrm{OR} = e^{(X_b - X_a)\beta}$. Moreover, this extraction is always safe to perform: even an uncorrected and therefore inconsistently estimated model fitted under response-based sampling restricts its inconsistency to the intercept term, which is not used in the extraction process.

Given the straightforward corrections for response-based sampling designs and finite sample corrections described previously, it is no longer necessary to work with the odds ratio in place of risk statistics. However, it remains an important intermediate factor in risk computations when P(Y) is not known with certainty.

Uncertainty about P(Y)

If P(Y) is known, the weighting or intercept correction estimators should be used to fit a model, and risk quantities computed directly. In rare cases, P(Y) is in fact known with certainty, for example, when a population census is available, or when the response in question is generated by a computer algorithm, e.g., a statistical estimator, which can be run on all available data at low cost. The majority of applications, however, do not enjoy this situation. In this majority of cases, P(Y) is not known for certain, and four cases can be distinguished.

P(Y) Can Be Estimated from Auxiliary Data

In this case, an independent sample is available from which P(Y) can be consistently estimated. This does not affect the statistical properties of any of the estimators

described previously. Risk statistics may be computed directly from the model.

P(Y) Is Vanishingly Rare

In second case, $Y=1$ is treated as vanishingly rare. Rearranging the definition of OR, it can be seen that

$$\mathrm{OR} = \mathrm{RR}\left[\frac{1-\mathrm{P}(Y=1\,|\,X=a)}{1-\mathrm{P}(Y=1\,|\,X=b)}\right]. \qquad (14)$$

Taking the limit as $\mathrm{P}(Y=1)\to 0$, $\mathrm{OR}\to\mathrm{RR}$. A traditional approximation suggests using the odds ratio in place of the relative risk when $Y=1$ is very rare. But this can no longer be recommended; not only does the odds ratio overestimate the relative risk, but $\mathrm{P}(Y=1)$ is known not to be 0, and better methods are available. This approach is more clearly a problem with the risk difference, where the same limit necessarily leads to an estimate of 0 (no effect).

P(Y) Is Completely Unknown

In a third case, P(Y) is assumed to be completely unknown. Manski showed that if P(Y) can lie anywhere on [0, 1], the relative risk is bounded by

$$[\min(1,\mathrm{OR}),\max(1,\mathrm{OR})], \qquad (15)$$

where the odds ratio is used as an effect size bound. Under the same conditions, the risk difference is bounded by

$$\left[\min\left(0,\frac{\sqrt{\mathrm{OR}}-1}{\sqrt{\mathrm{OR}}+1}\right),\max\left(0,\frac{\sqrt{\mathrm{OR}}-1}{\sqrt{\mathrm{OR}}+1}\right)\right] \qquad (16)$$

Unsurprisingly, due to the extreme level of uncertainty, estimates of relative risk and risk difference overlap 1 and 0, respectively (no effect).

P(Y) Is Bounded

In the fourth case, P(Y) is known only to lie on the interval $[\pi_a,\pi_b]$. In this case, a tighter set of bounds than the previous case is possible. The relative risk lies on

$$[\min(\mathrm{RR}_a,\mathrm{RR}_b),\max(\mathrm{RR}_a,\mathrm{RR}_b)], \qquad (17)$$

where RR_a is the risk ratio evaluated at $\mathrm{P}(Y=1)=\pi_a$.

Computing the risk difference is slightly more involved. Let π_K be the value of P(Y) that would generate the risk difference $K=\left(\sqrt{\mathrm{OR}}-1\right)/\left(\sqrt{\mathrm{OR}}+1\right)$. When π_a and π_b are either both greater than or both less than π_K, then the risk difference lies on

$$[\min(\mathrm{RD}_a,\mathrm{RD}_b),\max(\mathrm{RD}_a,\mathrm{RD}_b)]; \qquad (18)$$

otherwise, it lies on

$$[\min(\mathrm{RD}_a,\mathrm{RD}_b,K),\max(\mathrm{RD}_a,\mathrm{RD}_b,K)]. \qquad (19)$$

Conservative error bars can be placed on all these bounds by using the sampling distribution of β to generate outer bounds from the underlying model via simulation.

See Also the Following Articles

Maximum Likelihood Estimation • Sample Size

Further Reading

Cosslett, S. R. (1983). Efficient estimation of discrete choice models. In *Structural Analysis of Discrete Data with Econometric Applications* (C. F. Manski and D. McFadden, eds.), pp. 2–49. MIT Press, Cambridge, MA.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**(1), 27–38.

Hsieh, D., Manski, C. F., and McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Am. Statist. Assn.* **80**(391), 651–662.

Imbens, G. (1992). An efficient method of moments estimator for discrete choice models with choice-based-sampling. *Econometrica* **60**(5), 1187–1214.

King, G., and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis* **9,** 137–163.

King, G., and Zeng, L. (2002). Estimating risk and rate levels, ratios, and differences in case-control studies. *Statist. Med.* **22,** 1409–1427.

Manski, C. F. (1995). *Identification Problems in the Social Sciences.* Harvard University Press, Cambridge, MA.

Manski, C. F., and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data with Econometric Applications* (C. F. Manski and D. McFadden, eds.), pp. 2–49. MIT Press, Cambridge, MA.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd Ed. Chapman and Hall, New York, NY.

Rasch, Georg

David Andrich
Murdoch University, Murdoch, Western Australia

Glossary

deterministic model A mathematical model that gives the exact value for an outcome as a function of parameters.
frame of reference Contains sets of objects and agents that can be brought into contact with each other to produce a reaction or a response.
parameter A variable whose values govern or determine other values in mathematical models.
probabilistic model A mathematical expression that gives the probability of an outcome from a defined set of possible outcomes as a function of parameters.
relative invariance A comparison between two objects that is independent of the agents, and vice versa, within a frame of reference.
specific objectivity The property of invariance of comparisons of objects and of agents within a specified frame of reference.
sufficient statistic A function of data in a probabilistic model is a sufficient statistic with respect to a parameter if, given this statistic, the resultant distribution does not depend on that parameter.

Georg Rasch had a distinctive impact on social measurement. A mathematician who turned to statistics in the 1930s to earn a living, Rasch worked for 50 years during an exciting era for statistics. A member of the International Statistics Institute and charter member of the Biometrics Society, he knew all of the most influential statisticians of that era. His main works were published in 1960 and 1961, with a retrospective summary in 1997. His 1960 book, *Probabilistic Models for Some Intelligence and Attainment Tests*, summarized his original empirical work; his 1961 paper from the IVth Berkeley Symposium on Mathematical Statistics and Probability, "On General Laws and the Meaning of Measurement in Psychology," abstracted a class of probabilistic models from his empirical work that satisfied requirements for measurement; and his 1997 paper "On Specific Objectivity: An Attempt at Formalising the Request for Generality and Validity of Scientific Statements" in the *Danish Yearbook of Philosophy* articulated a general framework for invariant scientific reference.

The author recorded an interview with Rasch in June 1979 that is the basis for this article (Rasch died in 1980). Unreferenced quotes are from that interview.

Development as a Mathematician

Elementary and Secondary Schooling

Rasch was born in 1901 in Denmark. His mother, who died early in her life, seemed to have little impact on him. His main influence was his religious father, who had taught mathematics in a nautical school. When Rasch was ready to enter secondary school, three factors converged to lead him into mathematics. First, his father concluded that Rasch should continue his schooling, apparently because he was not practical; second, he came across his father's trigonometry books and was intrigued by them; third, he had an excellent teacher in arithmetic and algebra who persuaded Rasch's father to go to substantial extra expense to send him to a cathedral high school in Odense specializing in mathematics rather than a closer one specializing in languages.

After three years in high school, in September 1919, Rasch became a student of mathematics at the University of Copenhagen. There, Rasch immediately began working with his professors, and while still an undergraduate published a joint paper with Professor Niels Nielsen.

University Studies

Rasch then worked with Professor Nørland, an association that lasted for some 20 years. His first task was to study the preserved library of another Danish mathematician, J. L. W. V. Jensen, who claimed he had proved a theorem proposed by the German mathematician Rieman that would have given a great deal of information about how prime numbers are distributed. Though Rasch never found the proof, he did publish papers based on this search. He completed his masters degree in 1925 and his doctor of science degree in 1930.

> *My old teacher, Lehn, was quite right when he declared that the son of Mr. Rasch was a born mathematician. Not the best one in the world, by no means, but the interest in mathematics and the need for making research in mathematics has followed me from very early days. Although I have been known as a statistician, my original training and my original gift is in mathematics.*

Rasch published an elegant proof of Wishart's theorem concerned with the distribution of variances and covariances in multi-dimensions, and from his doctoral dissertation, *Matrix Calculus and Its Application to Difference Equations and Differential Equations*, published "The Theory and Application of the Product Integral" in a German journal. This paper was acknowledged in work in atomic and group theory research.

On completing his doctorate, Rasch was qualified to be appointed as a full professor in mathematics. He applied unsuccessfully for two such positions. Rasch believed these rejections were due to his association with Nørland, because the professorships were given to students of Harold Bohr, a mathematician and brother of the nuclear physicist Niels Bohr.

Start in Statistics

Analysis of a Data Set

If Rasch had been able to get a professorship in mathematics, his career would have taken a very different path, and his contribution to social measurement may never have appeared. His mathematical ability and training, however, were central to the contribution that he did make. Rasch did not become a full professor until 1960, when he was appointed as a professor of statistics in the faculty of social sciences at the University of Copenhagen on the basis of the work he had done by taking this very different career path.

Not having a faculty position in the 1930s meant times were difficult for Rasch, as they were for many people in that period. Rasch's ability in mathematics and networks with fellow students lead to improved personal circumstances and to his hesitant start in analyzing very different kinds of data. Again, a number of events converged. First, two medical acquaintances asked him to look at their data on reabsorption of cerebrospinal fluid. Although he did not know even the method of least squares, by trial and error he fitted an exponential curve to the data. He displayed his results in a way that anticipated his way of doing so for the rest of his career—revealing regularities in the data, if they existed, as straight lines.

> *When I just had the curvatures determined and then calculated the points corresponding to their position, the plots of the observed points through these calculated points gave the nicest straight lines I could ever wish. That was to them sensational. Out of this came a paper by Stűrup, Fog and Rasch. That was my first experimental paper.*

Second, appreciating that mathematics was relevant to them, Fog and Stűrup invited Rasch to teach them and some of their colleagues further mathematics. The same group persuaded him that he could read a statistics book more readily than they could and that he could teach them statistics. Through this group, he also became a consultant at the Hygienic Institute in Copenhagen.

Third, and again through personal contacts in which Rasch's mathematical ability was invoked to critique a doctoral dissertation, Rasch was invited by the head of the State Serum Institute, Dr. T. Madsen, to join the institute as a consultant.

Fourth, Nørland and Madsen, who knew each other, decided that for Rasch to make a worthwhile contribution in statistics, he needed proper training. They obtained a Carlsberg scholarship for Rasch to study with Ragner Frisch, an econometrician in Oslo, for 3 months and a Rockefeller Foundation scholarship to study with R. A. Fisher in London for a year, beginning in September 1935. Fisher's work had a major impact on Rasch.

Studies in Statistics

Although the direction Rasch took seems to have occurred through a series of coincidences, in each case he excelled in using his mathematics ability, and even though he did not have direct training in the work he was doing, he greatly impressed with what he could do with data, culminating in his opportunity to study with the most significant statistician of the time (perhaps of all time). Rasch believed that Fisher's realization of the concept of sufficiency was the high mark of Fisher's many contributions.

> *Many may consider it just a mathematical trick, but I think it's much more than that .... Now in my language today, the sufficiency concept is very remarkable. It plays an important role in, as we shall see, the probabilistic theory of specific objectivity.*

Return to Denmark: The Study of Individuals

On his return to Denmark, the range of data Rasch analyzed expanded. He analyzed psychological test data in the same distinctive way he had analyzed biological data. He critiqued the population studies of the biologist Julian Huxley.

Fairly early I got around to the problem of dealing with individuals. I had tried to do that for the growth of children already before I came to London. But meeting Julian Huxley showed me that this was really an important line of my research. I continued to stick, as far as I could, to the study of individuals ever since. It meant quite a lot to me to realize the meaning and importance of dealing with individuals and not with demography. Later on I realized that test psychologists were not dealing with the testing of individuals, but what they were studying was how traits, such as intelligence, were distributed in a population. From the data on children's growth, which was individual, and through my connection with Huxley, and then the move to psychological testing, has a continuous line.

The Multiplicative Poisson Model

Design for Invariant Comparisons

Rasch's next big opportunity occurred when, following preliminary analysis of errors in reading words in texts, a special design was used to collect new data. The design held that the texts students read should not be so difficult that they became frustrated and distracted, nor so easy that they were not challenged. The number of errors by each student should be on the order of 5 to 10%. This meant that the same text could not be given to all students. The design of the data collection took the form shown in Table I with students in higher grades given more difficult texts than those in lower grades. This design led Rasch to his models for measurement.

Table I Design of Reading Experiments—Taken from Rasch (1960, p. 5)

	Grade					
Test	*2*	*3*	*4*	*5*	*6*	*7*
ORF	+	+				
ORU		+		+		
ORS			+	+	+	
OR5			+	+	+	+
OR6					+	+

With the probability of an error being relatively low and the opportunity to make one relatively large, Rasch hypothesized that the Poisson distribution

$$\Pr\{X = x;\, \lambda\} = e^{-\lambda}\frac{\lambda^x}{x!}, \tag{1}$$

where X is the random variable for the number of errors, in which $E[X] = \lambda$ is the mean number of errors, would characterize the data. However, rather than seeing this as the characterization of a population, he resolved the parameter λ into two components, one for the person and one for the text: $\lambda_{vi} = \delta_i/\xi_v$ where δ_i is the difficulty of text i and ξ_v=0 is the ability of person v giving

$$\Pr\{X_{vi} = x;\, \delta_i, \xi_v\} = e^{-\delta/\xi_v}\frac{(\delta_i/\xi_v)^x}{x!}, \tag{2}$$

in which $E[X_{vi}] = \delta_i/\xi_v$.

Rasch called Eq. (2) the multiplicative Poisson model (MPM). It is evident that the mean number of errors is proportional to the text's difficulty, δ_i, and inversely proportional to the person's ability, ξ_v. This means, for example, that if the difficulty of text j is given by $\delta_j = k\delta_i$, then text j is k times more difficult than text i. This statement is analogous to the kinds of statements that are made in measurements in physics, to which Rasch related his work.

The Poisson distribution has the property that the probability of the sum of two Poisson distributions (e.g., $X_{v(i+j)} = X_{vi} + X_{vj}$) is governed by a parameter that is the sum of the parameters of the individual distributions (e.g., $\lambda_{v(i+j)} = \lambda_{vi} + \lambda_{vj}$). In the case of a person reading two texts, the multiplicative structure has the property that the difficulty of the combined texts is simply the sum of the difficulties of the individual texts: $\lambda_{v(i+j)} = \delta_{(i+j)}/\xi_v = (\delta_i + \delta_j)/\xi_v$, giving

$$\Pr\{X_{v(i+j)} = x;\, \delta_i, \delta_j, \xi_v\} = e^{-(\delta_i+\delta_j)/\xi_v}\frac{((\delta_i + \delta_j)/\xi_v)^x}{x!}, \tag{3}$$

in which $E[X_{v(i+j)}] = (\delta_i + \delta_j)/\xi_v$.

Rasch noticed that this structure, too, had an analogy to physical measurement—just like two masses in the laws of classical physics, two texts could be concatenated with the same response law prevailing. However, rather than taking concatenation as his starting point for measurement, as is done in representational measurement theory, he took a different route.

Eliminating the Person Parameter

The MPM has the property that the probability of the score x_{vi} by person v on text i, conditional on the person's total score on the two texts i and j being $r_v = x_{vi} + x_{vj}$, is given by the binomial distribution

$$\Pr\{X_{vi} = x;\, \delta_i, \delta_j, \xi_v \mid r_v\} = (r_v x_{vi})(\pi_{ij})^x(1 - \pi_{ij})^{r_v - x_{vi}}, \tag{4}$$

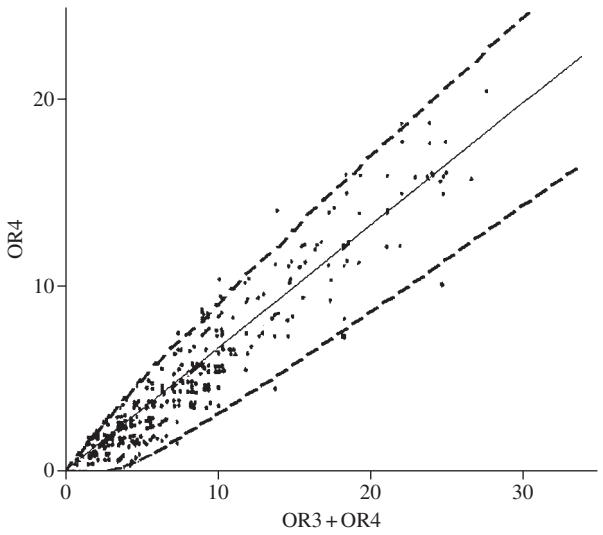


Figure 1 Graph showing the hypothesis line, the confidence interval, and observed points for the score of persons on one text relative to the total score on two texts. Adapted from Rasch (1960).

in which $\pi_{ij} = \delta_i/(\delta_i + \delta_j)$ is the probability of observing an error on text i. Equation (4) has no person parameter. The statistic r_v is sufficient for the person parameter ξ_v—that is, all the information that the data can provide about the parameter ξ_v is contained in r_v. Thus, r_v can be used directly to estimate ξ_v. However, for Rasch this was not the main issue. Instead, he saw two other points to the model. First, it meant that the relative difficulties of the texts could be estimated independently of the abilities of the persons. Second, it provided immediately an opportunity to check the fit of the model using a hypothesized straight line with confidence limits around it. Taken from Rasch's 1960 book, Fig. 1 shows such a check for scores on one text given the total score on two texts.

Rasch showed that with the MPM, relevant comparisons could be made even if all of the persons did not read all of the texts. More than that, Rasch turned this solution to a practical problem into a requirement for measurement.

The Dichotomous Model

Construction of the Model

Rasch's mathematical orientation was central to his formulation of his model for dichotomous responses. Rasch appreciated that there was a dichotomous response for each word in the count of the number of errors for a text as a whole:

> *The discovery of the model actually was an achievement in connection with the reading tests and the study of the multiplicative Poisson models. I chose the multiplicative Poisson because it seemed a good idea mathematically if it would work. It turned out that it did work. Then I wanted to have some good motivation for using it, and not only the excuse that statistically it worked perfectly. I wanted to have a good reason for trying that after I had used it.*

Rasch derived the model for dichotomous responses in which an error ($x = 1$) was counted, that produced the MPM for texts as a whole. If the correct response is counted ($x = 1$) rather than an incorrect one as in the case of reading, then the relationship $\lambda_{vi} = \delta_i/\xi_v$ is written inversely as $\lambda_{vi} = \xi_v/\delta_i$, $\delta_i > 0$, giving the dichotomous model

$$\Pr\{X_{vi} = x; \delta_i, \xi_v\} = \frac{(\xi_v/\delta_i)^x}{1 + \xi_v/\delta_i}, \qquad (5)$$

where $X_{vi} = x \in \{0,1\}$.

Rasch had earlier worked with intelligence tests, so when he had the dichotomous model, he analyzed existing data from two tests out of curiosity.

> *The first thing I did was, in fact, to analyze the Raven's tests. They worked almost perfectly according to the multiplicative model for dichotomous items. That was my first really nice example using the newly discovered model. Now I compared the Raven's test and the results of an analysis of the military tests which had been taken over as a routine intelligence test for recruits. The intelligence tests did not conform and I showed it to the head of the military psychologists group.*

Constructing Items to Conform to the Model

Rasch showed from the deviations of the data from the model that the 72 items fell into seven groups. The head of the military psychology group immediately understood Rasch's point, and instigated the construction of tests in which the items in each group were intended to conform to Rasch's new model.

This was a remarkable instruction. Rasch had tried his model on only two sets of data at the time: in one set, the Raven's progressive matrices, which is a non-verbal intelligence test, the model worked; and in the other set, which was the military's own test, it did not. Rather than abandoning the model at this point, and complicating it to better account for the data of the latter test, new items were to be constructed with the explicit intention that they conform to the model. Rasch's analysis of the data and conclusions with the model must have been extremely compelling. After the construction of these tests and analysis of their responses, Rasch wrote,

> *It is tempting, therefore, in the case with deviations of one sort or other to ask* whether it is the model or the test that

has gone wrong. *In one sense this of course turns the question upside down, but in another sense the question is meaningful. For one thing, it is not easy to believe that several cases of accordance between model and observations should be isolated occurrences. Furthermore the application of the model must have something to do with the construction of the test; at least, if a pair of tests showed results in accordance with our theory, this relationship could easily be destroyed by adding alien items to the tests. Anyhow, it may be worth while to know of conditions for the applicability of such relatively simple principles for evaluating test results [emphasis in original].* (Rasch, 1960/1978, p. 51)

This model is now more conventionally written and studied in the form

$$\Pr\{X_{vi} = x; \sigma_i, \beta_v\} = \frac{\exp x(\beta_v - \sigma_i)}{1 + \exp(\beta_v - \sigma_i)}, \qquad (6)$$

where $\beta_v = \ln \xi_v$ and $\sigma_i = \ln \delta_i$. Although it is the simplest possible model for a dichotomous response, it has generated extraordinary literature, both theoretical and empirical.

The Model for Any Number of Response Categories

Invariance as a Property of a Model—Not Just Data

The requirement that the comparisons of items should not depend on specific persons, and vice versa, was not new. Thurstone articulated the same requirement in a number of important papers in the 1920s. However, for Thurstone, the requirement of invariance was left as a requirement of data. Invariance was important for Rasch as well, but Rasch made it a property of a mathematical model. His ability and training in mathematics were important in first setting up the opportunity to do so, and then in identifying a class of models with this property. Equally importantly, this property made it possible to carry out mathematical derivations to understand other implications of the requirement of invariance.

The Insight into the Implications of Invariance

I saw the importance of finding an answer to the following question: which class of probabilistic models has the property in common with the multiplicative Poisson model, that one set of parameters can be eliminated by means of conditional probabilities while attention is concentrated on the other set, and vice versa.

In identifying a class of models for the probabilistic case that gives invariance, two related components are involved. First, the probability distribution itself needs to have the property of sufficiency for its parameters. The family of distributions that have sufficient statistics, which following Fisher's work were well defined, are known as the exponential family. Second, the structure of the parameters must give the possibility of separating the person parameters from the item parameters.

The General Response Model

In 1961, Rasch presented a very general class of distributions with the property of sufficiency in which the dichotomous and MPM models are special cases, which took the form

$$\Pr\{X_{vi} = x; \sigma_i, \beta_v\} = \frac{\exp(\varphi_x \beta_v + \psi_x(-\sigma_i) + \chi_x \beta_v \sigma_i + \kappa_x)}{\sum_{x=0}^{m} \exp(\varphi_x \beta_v + \psi_x \sigma_i + \chi_x \beta_v \sigma_i + \kappa_x)}. \qquad (7)$$

where φ_x and ψ_x were scoring functions of the categories, κ_x were category coefficients, and $X_{vi} = x$ was a response in any one of $m + 1$ categories to an item. To give measurements, which required the parameters to be scalar, Rasch quickly specialized this general model to the form

$$\Pr\{X_{vi} = x; \sigma_i, \beta_v\} = \frac{\exp(\varphi_x(\beta_v - \sigma_i) + \kappa_x)}{\sum_{x=0}^{m} \exp(\varphi_x(\beta_v - \sigma_i) + \kappa_x)}. \qquad (8)$$

There has been substantial development of this model for rating and performance assessments in items with formats with ordered response categories. Andersen (1977) proved that the sufficiency condition actually imposed the constraint

$$\varphi_x - \varphi_{x-1} = \varphi_{x+1} - \varphi_x, \qquad (9)$$

and Andrich (1978) interpreted the scoring functions φ_x and the category coefficients κ_x in terms of familiar concepts in psychometrics, thresholds τ_x, $x = 1, \ldots, m$ that partitioned the continuum into categories, and discriminations α_x, $x = 1, 2, 3, \ldots, m$ at the thresholds. Specifically,

$$\varphi_x = \alpha_1 + \alpha_2 + \alpha_3 + \cdots + \alpha_x \quad \text{and}$$

$$\kappa_x = -(\alpha_1\tau_1 + \alpha_2\tau_2 + \alpha_3\tau_3 + \cdots + \alpha_x\tau_x).$$

Without loss of generality, let $\alpha_x \equiv 1$, $x = 1, 2, \ldots, m$, giving $\varphi_x = x$ and $\kappa = -(\tau_1 + \tau_2 + \tau_3 + \cdots + \tau_x)$, which reduces Eq. (8) to

$$\Pr\{X_{vi} = x; \sigma_i, \beta_v, \tau_x\} = \frac{\exp(x(\beta_v - \sigma_i) - \sum_{k=0}^{x} \tau_k)}{\sum_{x=0}^{m} \exp(x'(\beta_v - \sigma_i) - \sum_{k=0}^{x'} \tau_k)}, \quad (10)$$

where $\tau_0 \equiv 0$. This model has been generalized further to the case in which different items may have different thresholds:

$$\Pr\{X_{vi} = x; \sigma_i, \beta_v, \tau_x\} = \frac{\exp(x(\beta_v - \sigma_i) - \sum_{k=0}^{x} \tau_{ki})}{\sum_{x=0}^{m} \exp(x(\beta_v - \sigma_i) - \sum_{k=0}^{x} \tau_{ki})}. \quad (11)$$

Interestingly, in this model the successive categories for an item are scored with successive integers, just as is commonly done in elementary analyses of ordered category data as in Likert-style questionnaires and in performance assessment. However, this scoring is not a result of equal distances between thresholds defining the categories, but of equal discriminations at the thresholds. This is the same condition that makes the total score in the dichotomous Rasch model the relevant statistic from which the person parameters can be eliminated when estimating the item parameters. Data with ordered categories are ubiquitous in the social sciences. They are used often by analogy to measurement in the physical sciences. Equations (10) and (11) make practical statistically advanced analyses of such data.

The Two-Way Frame of Reference

The Probabilistic Case

To formalize his mathematical reasoning in measurement, Rasch made explicit a two-way frame of reference as follows. Suppose that there is a set of objects O_v, $v = 1, 2, 3, \ldots \in \mathrm{O}$ and a set of agents A_i, $i = 1, 2, 3, \ldots, \in \mathrm{A}$, which may come into contact with each other in pairs to give responses $X_{vi} = x \in \mathrm{X}$ within some frame of reference $\mathrm{F} \equiv [\mathrm{O}, \mathrm{A}, \mathrm{X}]$, summarized in Table II. The random variable X_{vi} can be a vector. Although Rasch wrote the two-way frame of reference in general, in the case of measurement, the objects and agents are characterized by scalar parameters.

The formulation from which he derived Eq. (9), in which two agents A_i and A_j are brought into contact with object O_v to provide the response $X_{vi} = x$, Rasch specified as

$$\Pr\{(x_{vi}, x_{vj}), \delta_i, \delta_j, \xi_v | f(x_{vi}, x_{vj})\} = \vartheta(x_{vi}, x_{vj}, \delta_i, \delta_j) \quad (12)$$

Table II Rasch's Two-way Frame of Reference of Objects, Agents and Responses

		Agents				
Reactions	X_{vi}	A_1	A_2	$\cdot$	A_i	$\cdot$	A_I
Objects	O_1	x_{11}	x_{12}	$\cdot$	x_{1i}	$\cdot$	x_{1I}
	O_2	x_{21}	x_{22}	$\cdot$	x_{2i}	$\cdot$	x_{2I}
	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
	O_v	x_{v1}	x_{v2}	$\cdot$	x_{vi}	$\cdot$	x_{vI}
	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
	O_V	x_{V1}	x_{V2}	$\cdot$	x_{Vi}	$\cdot$	x_{VI}

so that the probability statement should not depend on the person parameter ξ_v. For example, in the case of the dichotomous model of Eq. (6), Eq. (12) specializes to

$$\Pr\{(x_{vi}, x_{vj}); \delta_i, \delta_j, \xi_v \,|\, r_v = x_{vi} + x_{vj}\} = \frac{\exp(-x_{vi}\delta_i - x_{vj}\delta_j)}{\exp(-\delta_i) + \exp(\delta_j)}. \quad (13)$$

Equation (4) with respect to the MPM has the same structure.

Rasch articulated the requirement of invariance between parameters in the two-way frame of reference in terms of comparisons:

> *The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which other stimuli within the considered class were or might also have been compared.*
>
> *Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion.* (Rasch, 1961, p. 332)

The Determinate Case

Rasch extrapolated the requirement of invariant comparisons from the probabilistic case to the determinate case and to the connection with measurement in physics. As indicated previously, Rasch already saw these connections from the multiplicative Poisson model.

In the determinate case, Rasch replaced the response space X of the probabilistic case with a reaction ρ_{vi}, $v = 1, 2, 3, \ldots; i = 1, 2, 3, \ldots, \in \mathrm{R}$, giving the frame of reference $\mathrm{F} \equiv [\mathrm{O}, \mathrm{A}, \mathrm{R}]$. Suppose again that the objects, agents, and now reactions, are characterized by scalar parameters ξ_v, α_i and ρ_{vi}, where $\rho_{vi} = \rho(\xi_v, \alpha_i)$ is a function of the object and agent parameters. The equation that reflects a local

comparison of agents 1 and 2 and depends on object v in the reaction is given by

$$u(\rho_{v1}, \rho_{v2}) = u(\rho(\xi_v, \alpha_1),\ \rho(\xi_v, \alpha_2)) = \vartheta(\alpha_1, \alpha_2 \mid \xi_v). \tag{14}$$

The equation that then reflects a global comparison of agents 1 and 2 and that is invariant with respect to the object v is given by

$$u(\rho_{v1}, \rho_{v2}) = u(\rho(\xi_v, \alpha_1),\ \rho(\xi_2, \alpha_k)) = \vartheta(\alpha_1, \alpha_2). \tag{15}$$

With usual conditions of continuity and differentiability on the functions u, ρ, and ϑ, Rasch established that the existence of strictly monotonic functions $\xi' = \varphi(\xi)$; $\alpha' = \psi(\alpha)$; $\rho' = \chi(\rho)$, which transform the reaction function $\rho(\xi_v, \alpha_i)$ into an additive relation of the form

$$\rho'_{vi} = \xi'_v + \alpha'_i \tag{16}$$

is a necessary and sufficient condition for Eq. (15) to hold. The same equation arises for the invariant comparison of objects relative to agents. The exponential transformation is strictly monotonic, and therefore, after transformations $\rho^*_{vi} = \exp \rho'_{vi}$, $\xi^*_v = \exp \xi'_v$, and $\alpha^* = \exp \alpha'_i$ of the variables in Eq. (16),

$$\rho^*_{vi} = \xi^*_v \alpha^*_i. \tag{17}$$

The comparison between agents 1 and 2 through the interaction with an object in the frame of reference F is then obtained as the ratio

$$\frac{\rho^*_{v1}}{\rho^*_{v2}} = \frac{\xi^*_v \alpha^*_1}{\xi^*_v \alpha^*_2} = \frac{\alpha^*_1}{\alpha^*_2}. \tag{18}$$

Rasch referred to a frame of reference F that provides invariant comparisons of agents with respect to objects, and vice versa, as specifically objective: objective because the comparisons were invariant, and specifically objective because they depended on the specified frame of reference.

Rasch showed the connection of his derivations to measurement in physics. In particular, in 1977 he dissected the general gas equation in the form

$$p = \frac{r}{v}(t + \gamma) = \frac{r}{v} T, \tag{19}$$

in which r and γ are constants, p (pressure) and v (volume) are positive real numbers, t is the temperature measured in centigrade degrees and T is the temperature measured on the absolute Kelvin scale to show its conformity to Eq. (17). Rasch was aware that such laws prevailed in physics.

Fundamental Measurement and the Laws of Physics

According to Krantz *et al*. (1971), the results of Eqs. (16) and (17) are entirely compatible with the axiomatic treatments of representational or fundamental measurement, which in general terms is also known as additive conjoint measurement. In addition to being compatible, it seems to answer a key question in physical measurement observed by Ramsay (1975) in a review of the Krantz *et al*. book (Rasch had made the same observation):

> *Also somewhat outside the concerns of the rest of the book, this chapter is unique in considering the representation of relations between measurable structures explicitly. It deals with the fact that virtually all the laws of physics can be expressed numerically as multiplications or divisions of measurements. Although this rule has been known for a long time and forms the basis of the techniques of dimensional analysis widely used in engineering and physics, it remains a phenomenon for which no satisfactory explanation has been forthcoming* (Ramsay, 1975, p. 58).

Rasch's conclusion—that for invariant comparisons to be possible in the case of determinate relationships among variables, it is necessary to have structures that are additive or, equivalently, multiplicative, within acceptable transformations—seems to go a long way toward explaining Ramsay's observation that the laws of physics are multiplicative.

Rasch and Fundamental Measurement

Rasch began his formulations with invariant comparison, and concluded them with the requirement of an additive or, equivalently, a multiplicative structure. Rasch speculated as to whether the comparison is the most elemental basis of knowledge. He believed that the comparison of objects must involve some reactions to some agent (and vice versa) and that the identification of a specifically objective frame of reference was essential for scientific inference.

Although Rasch strove for necessity and sufficiency in the relationships he formulated, he did not publish rigorous proofs of his results. He carried out proofs to his satisfaction, shared them with students and colleagues, and presented them in more or less detail in informal papers. Early in his work he did not seem to be aware of existing results in group theory that could have made his proofs simpler. Toward the end of his life, he was made aware of them, and showed great interest in them. In addition, it seems Rasch was not aware of the axiomatic approach to representational measurement theory or fundamental measurement.

Implications of Rasch's Insights

Implications for Practice

Axiomatic treatment of deterministic fundamental measurement is important in the understanding of

measurement. However, it has had relatively little impact on the practice of measurement in the social sciences, because data in the social sciences are generally not deterministic. Rasch's models provide an opportunity to test data generated in the social sciences against criteria of fundamental measurement, with the data in the measurement of reading and intelligence being exemplary. His models are now applied in a range of social science settings, including sociology, health care, social medicine, and marketing, as well as in education and psychology, from where they originated. When the criterion of fundamental measurement is the motivation for the application of one of his models, items are constructed with a view that they conform to the chosen model, and the emphasis on the data fitting the model rather than the other way around. This continues the tradition set by Rasch. Software programs are now available that make the analysis of data using Rasch models relatively routine.

Expansion of Rasch's Work

Rasch had a reputation for reading relatively little; he also wrote relatively little. However, the material he wrote was careful and compelling in giving a view that was different from the traditional. He was also a passionate advocate of the concept of invariant comparisons within a specified frame of reference and of the classes of models that had this property in lectures and personal interactions. This advocacy inspired others to take up his work. In Denmark, his approach to psychometrics is now the standard approach. He had a direct impact on a generation of his students, including his successor Erling B. Andersen, who advanced both the theory and practice of Rasch models. Outside of Denmark, he inspired Benjamin D. Wright at the University of Chicago and Gerhard Fischer at the University of Vienna to work on the models. They introduced many of their colleagues and students to Rasch's work and together they have further developed this class of models. These developments have included different ways of deriving the models, developing algorithms for estimating parameters in the models, and establishing means for detecting deviations of data from the models.

See Also the Following Articles

Contingency Tables and Log-Linear Models • Duncan, Otis Dudley • Guttman Scaling • Partial Credit Model • Rating Scale Model

Further Reading

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42,** 69–81.

Andersen, E. B., and Wolk Olsen, L. (2000). The life of Georg Rasch as a mathematician and as a statistician. In *Essays in Item Response Theory*, (A. Boomsma, M. A. J. van Duijn, and T. A. B. Snijders, eds.), pp. 3–4. Springer, New York.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* **43,** 561–574.

Duncan, O. D. (1984). *Notes on Social Measurement.* Russell Sage Foundation, New York.

Fischer, G. H., and Molenaar, I. W. (eds.) (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, New York.

Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A. (1971). . *Foundations of Measurement,* Vol. 1. Academic Press, New York.

Ramsay, J. O. (1975). Review of foundations of measurement Vol. I, by D. H. Krantz, R. D. Luce, P. Suppes, A. Tverskey. *Psychometrika* **40,** 257–262.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* (J. Neyman, ed.), Vol. IV, pp. 321–334. University of California Press, Berkeley CA.

Rasch, G. (1977). On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. *Dan. Yearb. Philos.* **14,** 58–94.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests,* expanded ed. The University of Chicago Press, Chicago, IL.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychol. Rev.* **34,** 278–286.

Wright, B. D. (1998). A history of social science measurement. *Educ. Meas. Issues Pract.* **16,** 33–45.

Rating Scale Model

Erling B. Andersen
University of Copenhagen, Copenhagen, Denmark

Glossary

conditional maximum likelihood estimates Estimates that are based on maximizing a conditional likelihood function.
goodness-of-fit test A test for judging the closeness of the observed data to those predicted by the model.
likelihood ratio test A test that is derived by the ratio of the likelihood under the model and the likelihood without assuming a model.
marginal maximum likelihood estimates Estimates that are based on maximizing a marginal likelihood function.
residual Difference between observed quantities and their expected values under the model, divided by the standard error of this difference.
symptom checklist for discomfort A set of items used by psychiatrists to measure the degree of discomfort/depression of their patients.

The rating scale model is a latent structure model for polytomous responses to a set of test items. The basic structure of the model is an extension of the Rasch model for dichotomous responses, suggested by Georg Rasch in 1961. It is called a rating scale model because the response categories are scored such that the total score for all items constitutes a rating of the respondents on a latent scale. It is assumed that the category scores are equally spaced, giving the highest score to the first or the last category. Thus, the phrasing of the response categories must reflect a scaling of the responses, such as "very good," "good," " not so good," and "bad."

Presentation of the Model

The rating scale model is a special case of the polytomous model, first presented by Rasch in 1961. It was reconstructed as a rating scale model by Andrich in 1978. A similar model was presented by Andersen in 1977. The main assumption for the rating scale model, apart from being a polytomous Rasch model, is that the scoring of the response categories must be equidistant (i.e., their values must increase by a constant). Both Andersen and Andrich presented convincing arguments for specifying this condition. In 1983, Andersen discussed a somewhat more general model, which is still a polytomous Rasch model but with less restrictive scoring of the response categories.

Consider n polytomous test items, $i = 1, \ldots, n$. The response U_i on test item i can take the values $h = 1, \ldots, m$. The response function for item i, category h, is

$$P_{ih} = \text{Prob}(U_i = h).$$

The response pattern for an individual is defined as

$$\mathbf{U} = (U_1, \ldots, U_n).$$

It is sometimes convenient to introduce the selection vectors

$$(u_{i1}, \ldots, u_{im}) = (0, \ldots, 1, 0, \ldots, 0),$$

where $u_{ih} = 1$ if response h is chosen on item i, and 0 otherwise.

The response function is assumed to depend for each individual on the value of an ability parameter θ, describing the individual. Accordingly, we write

$$P_{ih} = P_{ih}(\theta).$$

Under the rating scale model, the response functions have the form

$$P_{ih}(\theta) = \frac{e^{w_h\theta - a_{ih}}}{\sum_{h=1}^{m} e^{w_h\theta - a_{ih}}}, \qquad (1)$$

where $w_1, \ldots, w_m$ are the category scores, which prescribe how the m response categories are scored,

and a_{ih} are item parameters connected with the items and categories. Figure 1 shows the four response functions for a typical item with four possible responses. The category scores are chosen to be 3, 2, 1, and 0, and the item parameters are chosen to be $(a_{i1}, \ldots, a_{i4}) = (1.0, 0.75, 0.25, 0.0)$.

As can be seen in Fig. 1 and verified from the analytical form of the model in Eq. (1), the response function for the category with the highest category score tends to 1 as $\theta \to \infty$ and to 0 as $\theta \to -\infty$, whereas the response function tends to 0 as $\theta \to \infty$ and to 1 as $\theta \to -\infty$ for the category with the lowest category score. It follows that individuals with high abilities will, with very high probability, choose category 1, whereas individuals with very low abilities most likely will choose category 4. For middle values of the ability, there are moderate probabilities for all four categories to be chosen. The role of the item parameters is to shift the response function to the left for small values of a_{ih} and to the right for larger values of a_{ih}. Thus, the higher the value of a_{ih}, the lower the probability of choosing category h for the category with the largest category score and the higher the probability of choosing category h for the item with the lowest category score. For the category with middle category scores, the probability of choosing category h will increase with a_{ih} for high θ values and decrease for low θ values with high probability for large θ values and with low probability for small θ values.

The rating scale model is based on the assumption that the category scores $w_1, \ldots, w_h$ are equidistant; that is, the differences

$$d_h = w_h - w_{h-1}, \quad h = 2, \ldots, m,$$

are all equal. This assumption was first suggested in Andersen in 1977 based on certain properties connected with the sufficient statistic for θ, where θ is regarded as an unknown parameter to be estimated. When Andrich introduced the rating scale model in 1978, he showed that $w_h = h$ could be interpreted as the number of the categories in which the response occurs or as thresholds being exceeded. Any linear transformation

$$w'_h = c_0 + c_1 w_h$$

is, of course, equivalent to w_h since this only entails a redefinition of the scale of the θ axes.

In order to derive the probability of observing the response $\mathbf{u} = (u_1, \ldots, u_m)$, the assumption of local independence is needed—namely that given the value of θ, the responses on the n items will be independent. Thus, the required probability is

$$\begin{aligned} f(\mathbf{u} \mid \theta) &= \text{Prob}(\mathbf{U} = \mathbf{u} \mid \theta) \\ &= \prod_{i=1}^{n} \text{Prob}(U_i = u_i \mid \theta) \\ &= \frac{\prod_{i=1}^{n} \exp\left(\theta \sum_{h=1}^{m} w_h u_{ih} - \sum_{h=1}^{m} a_{ih} u_{ih}\right)}{C_i(\theta, \mathbf{a})}, \end{aligned} \tag{2}$$

where

$$C_i(\theta, \mathbf{a}) = \sum_{h=1}^{m} \exp(\theta w_h - a_{ih}).$$

The selection vector notation for u_{ih} is used in Eq. (2) to express which w_h and a_{ih} come into play when category h is chosen by the individual. A rating scale model can thus be described as Eq. (1) combined with local independence.

Equation (2) can be rewritten as

$$\begin{aligned} f(u \mid \theta) = \exp\left(\theta \sum_{i=1}^{n} \sum_{h=1}^{m} w_h u_{ih} - \sum_{i=1}^{n} \sum_{h=1}^{m} a_{ih} u_{ih}\right) \\ - C^{-1}(\theta, \mathbf{a}), \end{aligned} \tag{3}$$

where

$$C(\theta, \mathbf{a}) = \prod_{i=1}^{n} C_i(\theta, \mathbf{a}).$$

Since, obviously, $C(\theta, \mathbf{a})$ is just a normalization factor independent of the responses $(u_1, \ldots, u_u)$, Eq. (3) shows that the score

$$t = \sum_{i=1}^{n} \sum_{h=1}^{m} w_h u_{ih} = \sum_{i=1}^{n} w_{ih}$$

is sufficient for θ, where w_{ih} is the score of the category h the individual chooses on item i. This result is clear from ordinary exponential family theory. It can also be seen directly from Eq. (3), which is the likelihood function pertaining to θ and t. Note that the total score for a given individual is simply the sum of the category scores of the chosen response categories. The category

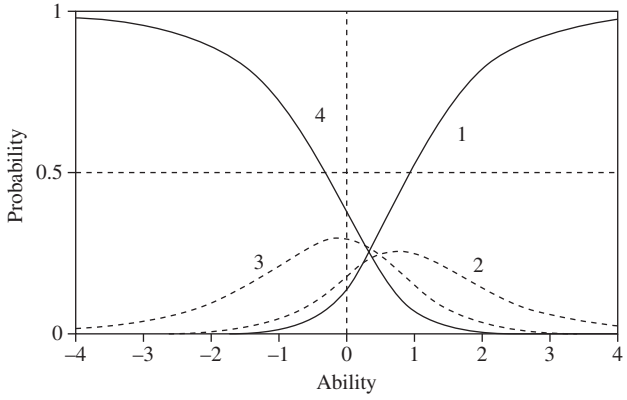


Figure 1 Response functions for four categories of an item (with parameter values 1.00, 0.75, 0.25, and 0.0). Reproduced with permission from W. van der Linden and R. K. Hambleton (eds.) (1997). "Handbook of Modern Item Response Theory," pp. 67–82, © Springer-Verlag.

scores accordingly can also be regarded as a scoring of the n responses into the total score. In 1977, Andersen showed that equidistant scoring was the only one allowing for a smooth variation over t values when a response was changed by changing just one category on one item.

The model is rather general with regard to the item parameters. For inference purposes, there is no real advantage in assuming a more specified structure of the item parameters (i.e., the a's), but for many applications it is natural, as well as in accordance with the theory behind the model, to assume a certain structure. One may assume, for example, that

$$a_{ih} = w_h a_i - \sum_{\ell=1}^{h} d_\ell,$$

where $d_1, \ldots, d_{m-1}$ are threshold parameters, making the item parameters a_{ih} smaller (or larger for negative d_ℓ), depending on the thresholds

$$d_h^* = \sum_{\ell=1}^{h} d_\ell.$$

In 1961, Rasch, in his first formulation of the model for polytomous items, argued that the item parameters a_{ih} should have the same multiplicative form

$$a_{ih} = v_h d_i^*$$

as the term θw_h, and even that the factor v_h corresponding to category h should be the same as the category scores, although he assumed that w_h was a parameter to be estimated.

The description of the model is completed by introducing the ability density $\varphi(\theta)$, which describes the variation of θ in the given population. The marginal distribution of any response pattern u is thus given by

$$f(\mathbf{u}) = \int f(\mathbf{u} \mid \theta)\phi(\theta)\, d\theta,$$

or, according to Eq. (3),

$$f(u) = \exp\Big(-\sum_i \sum_h a_{ih}u_{ih}\Big) \int \exp\Big(\theta \sum_{i=1}^{n} \sum_{h=1}^{m} w_h u_{ih}\Big) \times C^{-1}(\theta, \mathbf{a})\phi(\theta)\, d\theta. \quad (4)$$

Extensions of the rating scale model are provided by Fischer and Parzer and also by Glas.

Parameter Estimation

The item parameters a_{ih} can be estimated by conditional maximum likelihood (CML) or by marginal maximum likelihood (MML). From Eq. (1), it follows that only $n(m-1)-1$ parameters are unconstrained since $\sum_h P_{ih}(\theta) = 1$ and one a_{ih} can be changed by changing the θ scale.

CML Estimation

From Eq. (3), it follows that the marginal distribution of t (i.e., the probability of observing score t) is equal to

$$f(t \mid \theta) = e^{t\theta} C^{-1}(\theta, \mathbf{a}) \sum_{(t)} e^{-a_{ih}u_{ih}}, \quad (5)$$

where the summation (t) is overall response vectors with

$$t = \sum_i \sum_h w_h u_{ih}.$$

The last factor in Eq. (5) is usually denoted $\gamma_t(\mathbf{a})$ so that using

$$\gamma_t(\mathbf{a}) = \sum_{(t)} e^{-a_{ih}u_{ih}},$$

the distribution of t can be written as

$$f(t \mid \theta) = e^{t\theta} C^{-1}(\theta, \mathbf{a})\gamma_t(\mathbf{a}). \quad (6)$$

Hence, the conditional probability of observing response pattern u, given t, is

$$f(\mathbf{u} \mid t) = \frac{\exp\big(-\sum_i \sum_h a_{ih}u_{ih}\big)}{\gamma_t(\mathbf{a})}.$$

Now consider N individuals, who respond independently on the n items. If $(u_{ij1}, \ldots, u_{ijm})$ is the selection vector for individual j's response item i, the joint conditional distribution L_C is

$$L_C = \prod_{j=1}^{N} f(u_j \mid t_j) = \frac{\exp\Big(-\sum_i \sum_h a_{ih} \sum_{j=1}^{N} u_{ihj}\Big)}{\prod_{j=1}^{N} \gamma_{t_j}(\mathbf{a})} = \frac{\exp\Big(-\sum_i \sum_h a_{ih} y_{ih}\Big)}{\prod_t [\gamma_t(\mathbf{a})]^N t}, \quad (7)$$

where N_t is the number of individuals with score t, t_j is the score for individual j, and y_{ih} is the total number of responses h on item i.

Differentiation of the logarithm, $\ln L_C$, of Eq. (7) with respect to a_{ih} gives the conditional likelihood equations

$$\frac{\partial \ln L_C}{\partial a_{ih}} = \frac{y_{ih} - \sum_t N_t \partial \ln \gamma_t(a)}{\partial a_{ih}}.$$

Hence, the CML estimates are obtained as solutions to

$$y_{ih} = \frac{-\sum_t N_t \partial \ln \gamma_t(a)}{\partial a_{ih}}. \quad (8)$$

These likelihood equations have a unique set of solutions except for extreme sets of response patterns. Since Eq. (7) is an exponential family, it follows from a result by Barndorff-Nielsen that there is a unique solution unless the observed set of y's is a point on the

convex hull of the set of y's with positive probability given the observed frequencies over score values. Unfortunately, it is not possible to describe the set of extreme response patterns in close form, as is the case for the dichotomous Rasch model.

The conditional likelihood equations are usually solved by a Newton–Raphson procedure, which also provides standard errors for the parameters. The CML method was suggested by Rasch in 1961 and developed by Andersen in 1972. The recursive procedure needed for the calculation of γ functions is also described by Andersen.

MML Estimation

For MML estimation of the a's, the marginal likelihood L is needed. This likelihood is defined as the marginal probability of the responses u_j for $j = 1, \ldots, N$ (i.e., for the N respondents).

From Eq. (4), assuming independence,

$$L = \prod_{j=1}^{N} f(u_j) = \exp\Big(-\sum_i \sum_h a_{ih} y_{ih}\Big) \prod_t \Big[\int e^{\theta t} C^{-1}(\theta, \mathbf{a}) \phi(\theta)\, d\theta\Big]^{N_t}, \tag{9}$$

where again y_{ih}, $i = 1, \ldots, n$, $h = 1, \ldots, m$ are the item totals, and N_t is the size of score group t. In order to maximize Eq. (9), assumptions concerning the form of $\phi(\theta)$ are needed. It may thus be assumed that $\phi(\theta)$ belongs to a parametric family with two parameters b_1 and b_2. The log-likelihood function then becomes

$$\ln L = -\sum_i \sum_i a_{ih} y_{ih} + \sum_t N_t \ln \int e^{\theta_t} C^{-1}(\theta, \mathbf{a}) \phi(\theta \mid b_1 b_2)\, d\theta. \tag{10}$$

From this likelihood simultaneous estimates of the a_{ih}'s, b_1, and b_2 can be obtained.

The maximization of Eq. (10) requires numerical integration of functions such as

$$\frac{e^{\theta t} C^{-1}(\theta, \mathbf{a}) M \phi(\theta \mid b_1, b_2)}{M b_j}.$$

In case of a normal latent density, integration does not seem to be a serious numerical problem.

If one tries to maximize Eq. (10) for an unspecified latent density, only a set of discrete values for the ϕ function can be identified. In fact, nonparametric estimation of ϕ can be defined using the maximum likelihood estimate

$$\hat{\pi}_t = \frac{N_t}{N} \tag{11}$$

for the marginal probability of obtaining score t. The likelihood then becomes

$$\ln L = \ln L_C + \sum_t N_t \ln\left(\frac{N_t}{N}\right). \tag{12}$$

For this nonparametric estimation of ϕ, it was first shown by Tjur that the results from MML and CML estimation coincide.

Goodness of Fit

Multinomial Tests

The most direct goodness-of-fit test is based on the multinomial distribution of response patterns, given the model holds true. Given n items with m response categories for each item, there are n^m possible response patterns ranging from $(l, \ldots, l)$ to $(m, \ldots, m)$. Let u be a typical response pattern; then the joint distribution of all response patterns has the same likelihood as the multinomial distribution

$$\{n_u\} \sim \text{Mult}(N, \{\pi_u\}), \tag{13}$$

where n_u is the observed number of response pattern u, N is the total number of respondents, and π_u is the probability of observing response pattern u given by Eq. (3).

For estimated cell probabilities, π_u, the fit of the multinomial model (Eq. 13) is tested by the likelihood ratio test quantity

$$Z = 2 \sum_u n_u [\ln n_u - \ln(N \hat{\pi}_u)]. \tag{14}$$

Under suitable regularity conditions, Z is approximately χ^2 distributed with

$$\text{df} = n^m - 1 - q \tag{15}$$

degrees of freedom, where q is the number of estimated parameters in the model. An equivalent test is the Pearson test statistic

$$Q = \sum_u \frac{(n_u - N\hat{\pi}_u)^2}{N\hat{\pi}_u}, \tag{16}$$

which is also approximately χ^2 distributed with a number of degrees of freedom given by Eq. (15). For the rating scale model in Eq. (4), there are $n(m-1) - 1 + s$ parameters to be estimated, where s is the number of parameters in $\phi(\theta)$. Hence, the number of degrees of freedom is

$$\text{df} = n^m - n(m-1) - 1 - s.$$

A critical condition for the approximation to the χ^2 distribution is that the expected number $N\pi_u$ for the response patterns are not too close to zero. For many practical applications, the approximation is rather safe if we require that $N\pi_u > 3$ for all u. On the other hand, it

is clear that even for moderate values of n, especially if $m > 2$, a requirement such as $N\pi_u > 3$ is difficult to meet.

If the number of possible response patterns is very large—for example, $4^{20} = 1.0995 \cdot 10^{12}$, which is the number of possible patterns for $n = 20$ and $m = 4$—it is not likely that any expected number will satisfy the requirement, and we have to search for alternative tests. If the number of response patterns is moderately large, such as 243 (for $n = 5$ and $m = 3$), most response patterns will have low observed counts, but some response patterns may meet the requirement $N\pi_u > 3$. In this case, one possibility is to group the response patterns with low counts into groups with similar response patterns and let grouped observed and expected numbers appear in Eq. (14) or Eq. (16) as a single item.

The degrees of freedom for Z and Q will then be equal to

$$\mathrm{df} = N_G - 1 - q,$$

where N_G is the total number of terms corresponding to grouped or single response patterns appearing in the test quantities.

For a small set of items, where the number of response patterns is very limited, such as for $n = 5$ and $m = 2$, grouping can be avoided; however, in most cases a grouping is necessary.

A Likelihood Ratio Test

If a goodness-of-fit test cannot be based on the multinomial distribution over response patterns, one may use the statistical test suggested by Andersen in 1973.

Let $y_{ih}^{(t)}$ be the item totals for all respondents with score t; that is, $y_{ih}^{(t)}$ is the number of respondents with score t who respond h on item i. Then, the conditional likelihood (Eq. 7) can be factored as

$$\begin{aligned} L_C(\mathbf{a}) &= \prod_t L_C^{(t)}(\mathbf{a}) \\ &= \frac{\prod_t \exp\left(-\sum_i \sum_h a_{ih} y_{ih}^{(t)}\right)}{[\gamma_t(\mathbf{a})]^{N_t}}, \end{aligned} \tag{17}$$

where $L_C^{(t)}$ is the likelihood of all responses belonging to individuals with score t.

The CML estimates $\hat{a}_{ih}$ are the values of a_{ih} that maximize L_C given by Eq. (17). These estimates are the overall CML estimates. It is possible, however, to maximize the individual factors $L_C^{(t)}$ in L_C. This will result in a number of different sets of estimates called score-group CML estimates.

For practical reasons, one would often like to group the scores in interval groups such as $t_{g-1} < t < t_g$, but such a grouping does not affect the following argument.

If the model holds true, the score group estimates should only differ randomly from the overall estimates. If the model does not hold, the factors L_C in Eq. (17) would depend on the distribution of the abilities in the score groups. In 1973, Andersen suggested the use of the likelihood ratio test

$$Z_C = 2 \ln L_C(\hat{\mathbf{a}}) + 2 \sum_t \ln L_C^{(t)}(\hat{\mathbf{a}}^{(t)}). \tag{18}$$

In this test statistic, the maximum of $\ln L_C$, with the overall estimates $\hat{a}_{ih}$ inserted, is compared with $\ln L_C$ with the score group estimates $\hat{a}_{ih}^{(t)}$ inserted. Clearly, Z_C is larger than zero, and the larger the value of Z_C, the less likely it is that the model fits the data. Under suitable regularity conditions,

$$Z_C \sim \Pi^2(\mathrm{df}),$$

where

$$\mathrm{df} = [n(m-1) - 1](T - 1),$$

where T is the number of score groups.

One has to be particularly careful with the approximation to the limiting Π^2 distribution in this case since many score group totals $y_{ih}^{(t)}$ are likely to be small. Hence, a grouping of scores into score intervals is necessary except for the case in which there is a very small number of items.

A goodness-of-fit test based on Z_C given by Eq. (18) is less sensitive to model deviations than a test based on the multinomial tests Z or Q. However, as mentioned previously, it is often the only practical possibility. The types of model deviations that are likely to be detected by the test statistic Z_C were discussed by Andersen in 1973 and Glas in 1989. Glas suggested the use of Wald-type tests based on the score group total $y_{ih}^{(t)}$. The likelihood ratio test statistic Z_C checks only the part of the rating scale model contained in Eq. (2) since the conditional likelihood is independent of Eq. (θ) and, hence, of the form of the latent density (Eq. 8). However, since the probability $f(\mathbf{u} \mid \theta)$ of response pattern U, given by Eq. (4), can be factored as

$$f(\mathbf{u} \mid \theta) = f(\mathbf{u} \mid t) \cdot f(t \mid \theta),$$

where $f(\mathbf{u} \mid t)$ and $f(t \mid \theta)$ are given by Eq. (6) and the formula before Eq. (7), the total likelihood L factors into the conditional likelihood L_C and a term containing the marginal probabilities $f(t) = \pi_t$ of the scores, where π_t is given by

$$\pi_t = \gamma_t(\mathbf{a}) \int e^{\theta t} C^{-1}(\theta, \mathbf{a}) \phi(\theta)\, d\theta.$$

Hence, a two-stage procedure can be adopted: in the first stage the item parameters are estimated based on L_C and the model fit is checked based on Z_C, whereas in the second stage the form of $\phi(\theta)$ is checked based on

the multinomial distribution of the score t over its range. The relevant goodness-of-fit test statistic in stage two would then be

$$Z_T = \sum_t N_t[\ln N_t - \ln(N\hat{\pi}_t)], \tag{19}$$

where $\hat{\pi}_t$ is π_t with its parameters estimated. It is safe, although not theoretically optimal in stage two, to use the CML estimates from stage one as estimates for the a_{ih}'s.

If a goodness-of-fit test has an observed level of significance so low that the model must be rejected, it is important to be able to identify data points contributing significantly to the lack of fit. Residuals are the appropriate tools for this purpose.

Residuals

For the multinomial tests Z, given by Eqs. (14) and (16), the residuals are defined as

$$r_u = \frac{(n_u - N\hat{\pi}_u)}{\text{s.e.}\{n_u - N\hat{\pi}_u\}}. \tag{20}$$

The standard error, s.e.$\{n_u - N\hat{\pi}_u\}$, is the square root of

$$\text{var}[n_u - N\hat{\pi}_u].$$

The precise form of this variance was derived by Rao as

$$\text{var}[n_u - N\hat{\pi}_u] = N\pi_u(1-\pi_u)(1-h_u),$$

where h_u, with $0 < h_u < 1$, is a correction term that depends (on matrix form) on the response probabilities and their derivatives. Note that the variance is smaller than the multinomial variance $N\pi_u(1-\pi_u)$. The correction term h_u can be large, even close to 1, especially if a substantial percentage of the respondents choose response pattern u.

For the test statistic Z_C, there are two possibilities for residuals. The first possibility is to standardize the differences

$$y_{ih}^{(t)} - E[Y_{ih}^{(t)}], \tag{21}$$

where the mean values are estimated using the overall estimates. Since Eq. (21) set equal to 0, for all i and h, represents the likelihood equations for the score group CML estimates, large values of these residuals would point to model deviations. The variance of the expression in Eq. (21) can, in principle, be derived from the response pattern variances and covariances, but actual computations are time-consuming. As a second possibility, the overall CML estimates and the score group CML estimates can be compared directly in the differences

$$\hat{a}_{ih}^{(t)} - \hat{a}_{ih},$$

for all i, h, and t.

It was proved by Andersen in 1995 that approximately

$$\text{var}[\hat{a}_{ih}^{(t)} - \hat{a}_{ih}] = \text{var}[\hat{a}_{ih}^{(t)}] - \text{var}[\hat{a}_{ih}].$$

Thus, it is very easy to obtain the residuals

$$r_{ih}^{(t)} = \frac{(\hat{a}_{ih}^{(t)} - \hat{a}_{ih})}{\text{s.e.}\{\hat{a}_{ih}^{(t)} - \hat{a}_{ih}\}}. \tag{22}$$

Example

In order to illustrate the use of the rating scale model, 14 items from the Symptoms Check List for Discomfort scale were analyzed by Bech *et al*. in 1992. To obtain the scaling, psychiatric patients were presented with a number of items corresponding to problems or complaints that people experience. For each item, the patients were asked to describe how much that particular problem had bothered or distressed them during the last week. The response categories were as follows: (1) extremely, (2) quite a bit, (3) moderately, (4) a little bit, and (5) not at all. For present purposes, categories 1–3 were merged.

The following 14 items were selected for illustrative purposes:

1. Blaming yourself for things
2. Feeling critical of others
3. Your feelings being easily hurt
4. Feeling hopeless about the future
5. Feeling blue
6. Feeling lonely
7. Thoughts of ending your life
8. Having to do things very slowly
9. Difficulty making decisions
10. Trouble concentrating
11. Your mind going blank
12. Lack of sexual interest
13. Trouble falling asleep
14. Feeling low in energy

If a rating scale model fits the patients' responses to these 14 items, each patient's total score would represent a degree of discomfort and thus implicitly provide a scaling of the patient's discomfort/depression. Table I shows the item totals and score group totals for the equidistant scores

$$w_h = m - h, \quad h = 1, \ldots, m.$$

The minimum and maximum obtainable scores across the 14 items with this scoring of response categories were 0 and 28 (i.e., 14×2).

The minimum score corresponded to no symptoms, whereas a high score reflected a high degree of discomfort

or a high indication of psychiatric problems. The CML estimates are given in Table II.

The CML estimates show, as do the item totals, that items 2, 4, 5, 9, and 12 contributed most to a high discomfort score, whereas items 6–8, 13, and 14 were less often used as indicators of discomfort. Thus, in a rating scale model the item parameters reflected how strongly a given item tended to provoke a response contributing to a high score.

For $n = 14$ items with $m = 3$ response categories, there are approximately 5 million possible response patterns. For this case, a goodness-of-fit test based on the multinomial distribution over response patterns was not possible. However, it was possible to use the test statistic Z_C given by Eq. (18) based on score groups. To ensure that the required approximations could be expected to hold, the following interval grouping of the score t was selected: 0–7, 8, 9–11, 12–13, 14, and 15–28. Figure 2 shows the score group CML estimates plotted against the overall estimates. The main structure was satisfactory, but there were obvious deviations from the ideal identity line. The test statistic Z_C had observed value

$$Z_C = 174.45, \quad df = 135,$$

with a level of significance of approximately 0.1%. It was questionable, therefore, to accept the model and, consequently, a rating scale based on all 14 items.

To inspect the lack of fit more closely, the residuals in Eq. (22) were plotted against the item number. An inspection of this plot (Fig. 3) reveals that the significant residuals were concentrated in items 5, 13, and 14. Hence, a set of items without these three items should be expected to fit the model better. When the goodness-of-fit test was repeated for this set and the residuals were plotted again, the level of significance for the Z_C test became 0.06, and the residual plot looked satisfactory. If item 8 was also excluded from the set, the goodness-of-fit test Z_C had the observed value $Z_C = 124.50$, with 114 degrees of freedom. This corresponded to a level of significance of approximately 25% and was judged to be a very satisfactory fit. It seemed that the rating scale model fit the data well with 11 of the original 14 items included in the scoring of the responses and very well with 10 items included. Table III provides a summary of the goodness-of-fit tests.

To illustrate the use of a normal latent density, consider the test statistic Z_T given by Eq. (19). In Table IV, the observed score group counts N_t and their expected values

Table I Item Totals, y_{ih}, and Score Group Counts N_t

y_{ih}	$h=1$	2	3	t	N_t	t	N_t
$i=1$	117	155	526	0	151	15	12
2	154	235	409	1	70	16	10
3	142	149	507	2	54	17	10
4	161	226	411	3	48	18	12
5	187	237	374	4	54	19	7
6	72	126	600	5	40	20	12
7	67	101	630	6	32	21	8
8	76	135	587	7	40	22	7
9	147	206	445	8	37	23	4
10	91	177	530	9	40	24	5
11	120	164	514	10	36	25	6
12	173	213	412	11	24	26	5
13	27	53	718	12	19	27	6
14	76	116	606	13	22	28	4
				14	23		

Table II CML Estimates[a]

$\hat{a}_{ih}$	$h=1$	2
$i=1$	−0.032 (0.133)	0.138 (0.101)
2	1.048 (0.125)	1.120 (0.097)
3	0.366 (0.125)	0.205 (0.103)
4	1.104 (0.124)	1.079 (0.097)
5	1.578 (0.122)	1.351 (0.100)
6	−1.153 (0.159)	−0.416 (0.106)
7	−1.424 (0.165)	−0.756 (0.114)
8	−1.003 (0.155)	−0.291 (0.104)
9	0.761 (0.125)	0.806 (0.097)
10	−0.432 (0.144)	0.222 (0.097)
11	0.070 (0.132)	0.247 (0.100)
12	1.211 (0.122)	1.028 (0.099)
13	−3.371 (0.256)	−1.839 (0.149)
14	−1.104 (0.156)	−0.514 (0.109)

[a] With Normalizations $\hat{a}_{i3} = 0$. Standard errors are in parentheses.

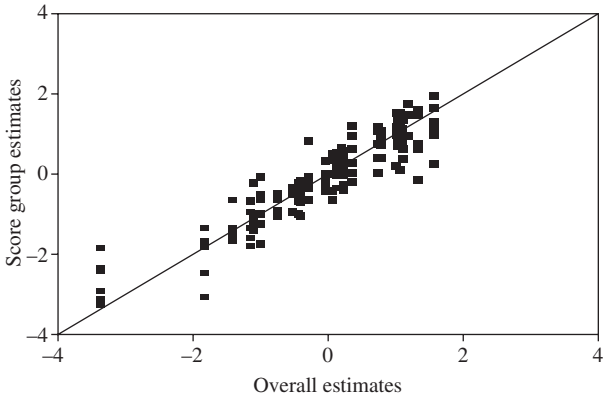


Figure 2 Score group CML estimates plotted against overall CML estimates. Reproduced with permission from W. van der Linden and R. K. Hambleton (eds.) (1997) "Handbook of Modern Item Response Theory," pp. 67–82, © Springer-Verlag.

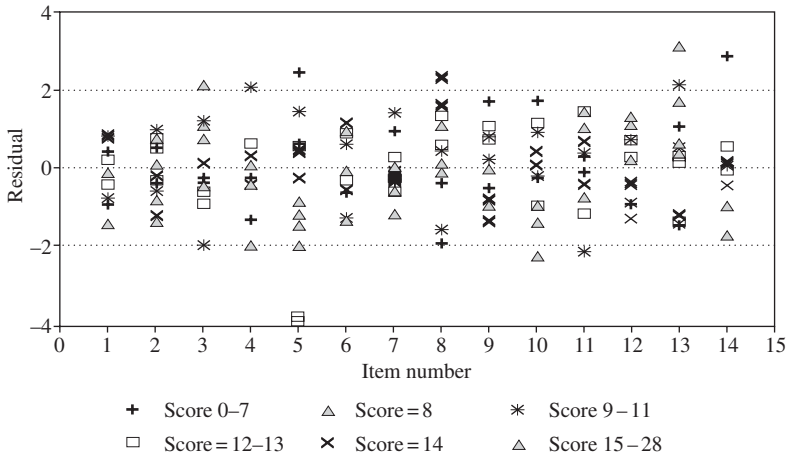


Figure 3 Residuals plotted against item number (14 items). Reproduced with permission from W. van der Linden and R. K. Hambleton (eds.) (1997) "Handbook of Modern Item Response Theory," pp. 67–82, © Springer-Verlag.

Table III Goodness-of-Fit Tests for Various Selections of Items

No. of items	*Items included*	*Items excluded*	*Goodness-of-fit test*	*Degrees of freedom*	*Level of significance*
14	1–14	None	174.45	135	0.013
11	1–4, 6–12	5, 13, 14	129.13	105	0.055
10	1–4, 6–7, 9–12	5, 8, 13, 14	124.50	114	0.236

Table IV Observed and Expected Score Group Counts for 10 Items

Score	*Observed count*	*Expected count*	*Standardized residuals*
$t = 0$	176	149.88	4.744
1	78	111.30	−3.486
2	61	82.22	−2.552
3	57	63.76	−0.912
4	57	51.55	0.808
5	48	43.03	0.800
6	45	36.82	1.414
7	48	32.12	2.920
8	47	28.45	3.605
9	20	25.53	−1.131
10	24	23.15	0.183
11	25	21.16	0.858
12	16	19.49	−0.813
13	19	18.04	0.233
14	16	16.77	−0.194
15	14	15.63	−0.428
16	11	14.58	−0.979
17	5	13.54	−2.455
18	9	12.39	−1.040
19	12	10.80	0.408
20	10	7.78	0.925

$N_t\hat{\pi}_t$ for a normal latent density are given. The estimated mean and variance were found to be

$$\hat{\mu} = -1.35 \quad \text{and} \quad \hat{\sigma}^2 = 2.96$$

The observed value of Z_T was $Z_T = 54.63$, with 18 degrees of freedom. This result was significant at 0.1%. Hence, a latent structure model seemed to fit the data but a normal latent density did not. The residuals given in Table IV (i.e., the differences between observed and expected values divided by their standard errors) show that the lack of fit is primarily due to a clear overrepresentation of score 0 as compared to the model. Thus, the model to a large extent described the shape of the score distribution but was not able, it seemed, to describe the frequency of the (not very interesting) respondents with score 0.

See Also the Following Articles

Maximum Likelihood Estimation • Rasch, Georg

Further Reading

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42,** 69–81.

Andersen, E. B. (1983). A general latent structure model for contingency table data. In *Principles of Modern*

Psychological Measurement (P. Pichot, ed.), pp. 117–139. Erlbaum, Hillsdale, NJ.

Andersen, E. B. (1991). *The Statistical Analysis of Categorical Data,* 2nd Ed. Springer-Verlag, Heidelberg.

Andersen, E. B. (1995). Residual analysis in the polytomous Rasch model. *Psychometrika* **60,** 375–393.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika* **43,** 561–573.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Appl. Psychol. Measurement* **2,** 581–594.

Andrich, D. (1982). An extension of the Rasch model to ratings providing both location and dispersion parameters. *Psychometrika* **47,** 105–113.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

Bech, P. (1990). Methodological problems in assessing quality of life as outcome in psychopharmacology: A multiaxial approach. In *Methodology of the Evaluation of Psychotropic Drugs* (O. Benkert, W. Maier, and K. Rickels, eds.), pp. 79–110. Springer-Verlag, Berlin.

Bech, P., Allerup, P., Maier, W., Alhus, M., Lavori, P., and Ayuso, J. L. (1992). The Hamilton scale and the Hopkins Symptom Checklist (SCL-90): A cross-national validity study in patients with panic disorders. *Br. J. Psych.* **160,** 206–211.

Cressie, N., and Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psyckometrika* **48,** 129–141.

de Leeuw, J., and Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *J. Educational Statistics* **11,** 183–196.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., and Covi, L. (1974). The Hopkins Symptom Checklist (HSCL). In *Psychological Measurement in Psychopharmacology* (P. Pichot, ed.), pp. 79–110. Karger, Basel.

Fischer, G. H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika* **46,** 59–77.

Fischer, G. H., and Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika* **56,** 637–651.

Glas, C. A. W. (1988a). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika* **53,** 525–546.

Glas, C. A. W. (1988b). The Rasch model and multistage testing. *J. Educational Statistics* **13,** 45–52.

Glas, C. A. W. (1989). Contributions to estimating and testing Rasch models. Doctoral dissertation, University of Twente, Enschede, The Netherlands.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika* **49,** 359–381.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd Ed. Wiley, New York.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Statistics* **9,** 23–30.

Regional Input–Output Analysis

Eveline S. van Leeuwen
Free University, Amsterdam, The Netherlands

Peter Nijkamp
Free University, Amsterdam, The Netherlands

Piet Rietveld
Free University, Amsterdam, The Netherlands

Glossary

biregional input–output table An input–output table that describes two regions; usually, the first region is a particular area such as a county, and the second region is "the rest of the country."
final demand The sum of the value of consumption by households, purchases by the government, purchases as investments, mostly by companies, and exports.
intermediary deliveries The delivery of goods from a certain industry to other industries or to the same industry in another region.
internal deliveries The delivery of goods from a certain industry within the same industry.
multiplier A value that shows the effect of a certain impulse as a result of the estimated recirculation of spending within the area concerned.
primary input Sum of labor costs, capital costs, payments to the government (e.g., taxes), and import costs.
technical coefficients The proportion in which a specific input is required to produce one unit of output.
transactions matrix A representation of the monetary value of the flows of goods between industries, the final demand, and the primary input within an economy.

Regional input–output analysis depicts the flows of goods (and services) between industries in the economy of a region. In the early history of input–output analysis, only national input–output tables were used; however, as interest increased in regional economic analysis, regional input–output tables were developed. In dealing with a regional input–output table, regional multipliers can also be derived. Depending on the geographic subdivision pertaining to the table, they show the effects of a certain impulse on the region, on other regions, or on the national economy. A great advantage of input–output models is their internal consistency. All effects of any given change in final demand can be recorded. This article gives an overview of the potential of this method as well as a numerical illustration.

Introduction

Input–Output Analysis

Input–output analysis is an established technique in quantitative economic research. It belongs to the family of impact assessment methods and aims to map the direct and indirect consequences of an initial impulse into an economic system across all economic sectors. It is essentially a method that depicts the systemwide effects of an exogenous change in a relevant economic system.

Input–output models are based on the idea that any output requires a corresponding input. Such input may comprise raw materials and services from other industries but also labor from households or certain amenities provided by the government. The output consists of a sectoral variety of products and services. A conventional input–output table is based on double-entry

bookkeeping: the totals of the columns equal the totals of the rows. Input–output tables may relate to the global economic system, the national economy, and also regional systems.

With the help of regional input–output tables, interdependencies and linkages between industries, households, and the government in and between regions can be examined. This method has found many applications worldwide and is one of the foundation stones of policy impact analysis.

Examples

One example of the use of input–output modeling can be found in a 1998 Australian study by McKay in which the impact of foreign students on the local economy of the city of Wollongong was estimated. Since large amounts of money are often invested in the image of a university and its home city, it is interesting to know what the expected or calculated effects of the expenditures to attract (foreign) students are on the local economy. In this study, it appeared that the effects on the local economy were significant because $1 of investments appeared to create $1.8 of household income and a large number of jobs were created.

Another example is the impact of a certain industry, such as forestry, on other industries. Because the forestry industry uses inputs from other industries and labor from the households in the area, and because it delivers products to several industries, an input–output analysis is a proper analytical instrument. With the use of an input–output table, the impact of a growing forestry industry and also a decreasing industry can be estimated. McGregor and McNicoll performed this exercise in the United Kingdom. In a simulation experiment, they assumed a reduction in output of the forestry industry to zero. Of course, the largest impact is seen in the forestry industry, but the households and the banking, finance, and insurance sectors, as well as the energy and water industries, also encounter major negative effects.

Input–output analysis has become a dominant analytical method in applied economic research. Its strength is its overall consistency at a systemwide level and its power to estimate all direct and indirect implications of an initial stimulus. This potential has been broadly recognized in the history of input–output modeling.

Historical Background

The input–output theory was developed by Wassily Leontief in the late 1920s. He was born in 1906 in Saint Petersburg, and in 1932 he developed the first input–output table of the U.S. economy. Until World War II, the theory was not in demand because many researchers thought it was too mathematical and too data demanding. During World War II, however, the value of the theory was more appreciated as it became very useful to identify bottlenecks in military production chains, such as determining when additional workers were needed.

At that time, input–output tables were used at a national level. Later, regional and international tables were developed. In 1973, Leontief won the Nobel Prize in economics for this groundbreaking and influential work in the assessment of input–output transaction tables. This methodology has become one of the standard tools for economists.

Input–Output Model

The basic information dealt with in the input–output model concerns the flows of products from each industrial industry considered a producer (output) to each of the industries considered a user (input). These flows are often physical or material in nature, but they are usually expressed in monetary terms and described in an interindustry transactions table. The rows of the table describe the distribution of a producer's output throughout the economy. The columns describe the inputs required by a particular industry to produce its output. The "sales to final markets" is included in an additional "final demand" column. Other inputs for the production, such as labor, are included in an additional "primary input" row. The table can describe national flows but also international or regional flows of products.

The usual sources of data for input–output tables are the national (or regional) economic accounts. These accounts are often collected on a regular (yearly) basis by means of surveys among individual firms. For national input–output tables, for example, national income and products accounts and interindustry or input–output accounts are used. The data requirements for building a comprehensive input–output model are formidable.

Input–output theory can be applied to modeling experiments and used for descriptive analyses. Modeling experiments use economic multipliers to map all relevant effects. Descriptive analyses describe, for example, the spatial relationships between regions or the structure of the industries in a certain region.

The Structure of an Input–Output Table

Because the input–output table describes the flows of products between and within industries for a certain region, the table can be divided into four quadrants (Fig. 1):

1. The upper left quadrant of the table contains the internal and intermediary delivery of goods from the industries to their own or to other industries.

From	To	Industry 1	Industry 2	Industry 3	Industry 4	Households	Government	Investments	Export	Total (X)
Industry	1	z_{11}	z_{12}	z_{13}	z_{14}	c_1	g_1	i_1	e_1	X_1
	2	z_{21}	z_{22}	z_{23}	z_{24}	c_2	g_2	i_2	e_2	X_2
	3	z_{31}	z_{32}	z_{33}	z_{34}	c_3	g_3	i_3	e_3	X_3
	4	z_{41}	z_{42}	z_{43}	z_{44}	c_4	g_4	i_4	e_4	X_4
Primary input factors	Labor	l_1	l_2	l_3	l_4					L
	Capital	k_1	k_2	k_3	k_4					K
	Government	o_1	o_2	o_3	o_4					O
	Import	m_1	m_2	m_3	m_4					M
Total (Z)		Z_1	Z_2	Z_3	Z_4	C	G	I	E	

Figure 1 Elements of an input–output table.

2. The bottom left quadrant shows the primary costs, such as import and labor costs.
3. The upper right quadrant contains the final demand consisting of the demand of households and government, investments, and export.
4. The bottom right quadrant shows the portion of the primary costs that directly apply to the final sales.

As previously mentioned, the values of the cells are often expressed in monetary terms. Because the goods, produced by the different industries, are very heterogeneous, monetary terms are the best way to compare the flows. In some cases, such as when a table describes flows of energy, prices are not interesting. In these cases, energy terms are used.

The notation used in Fig. 1 is as follows:

Internal and intermediary deliveries

z_{ij} is the value (in monetary terms) of the delivery from industry i to industry j in a certain period for a certain economic system.

X_i is the total value of the goods produced by industry i.

Z_j is the total value of all inputs required by industry j.

Final demand

c_i is the value of consumption by households of goods from industry i.

g_i is purchases by the government of goods from industry i.

i_i is purchases as investment, mostly by companies, from industry i.

e_i is exports by industry i.

Primary input factors

l_j is labor costs of industry j.

k_j is capital costs of industry j.

o_j is payments to the government by industry j.

m_j is import costs of industry j.

Summarizing, the rows of the input–output table contain the internal deliveries (e.g., the agricultural industry that delivers products to other companies within the agricultural industry) as well as intermediary deliveries (e.g., the agricultural industry delivering to the food industry). Furthermore, the rows contain deliveries to the final demand, such as the consumers, the export sector, or the government. Each row sums up to the total output of an industry.

The columns contain deliveries from other industries as well as from the own industry to the pertaining industry (e.g., from the building industry to the agricultural industry). The columns also contain the rest of the necessary inputs, such as labor, imports, and indirect taxes.

Regional Input–Output Tables

As mentioned previously, initially only input–output tables concerning the national economy were used.

However, decades after the first national table was constructed interest arose in regional economic analysis. Therefore, regional input–output tables had to be envisaged and developed. According to Miller and Blair, there are two basic features of a regional economy that affect the characteristics of a regional input–output table.

First, the technical coefficients (describing the proportion in which a different input is required to produce one unit of output), which at the national level are composed of average data of individual producers, have to be adapted to regional specificities. Although at the national level the average technical coefficients of all regions are given, the production process of similar goods in distinct regions can differ substantially. For example, energy can be produced using water power, wind power, or coal. The required input differs very much; therefore, the national input–output table should be adapted using region-specific data. Information such as household income and the output of industries at the regional level is used to modify the national data.

Second, smaller economic areas are more dependent on trade with other areas both for the sales of outputs (export) and for the purchase of inputs (import). One can imagine that a city cannot produce all the inputs that are needed for its production, and that it cannot sell all the goods that it has produced. On the other hand, if we take the world as an economic area, it becomes clear that no import or export will exist at a global level.

When examining the spatial level of detail within input–output tables, we can distinguish between national, intraregional, and interregional tables. The intraregional table describes the relations, the flows of products, between firms in one region. The transactions with industries in other regions of a country are seen as export or import to a foreign country. The interregional table discerns different industries in different regions. It divides the transactions according to the industries concerned. When an interregional table describes two regions, it is called a biregional table. The construction is easier because the second region is "the rest of the country," which means that the two regions sum to the total national economy. Biregional tables give detailed information concerning the relations between firms within the region but also of the relations between firms within the region and firms within the rest of the country. Clearly, the step to multiregional tables is straightforward but, of course, more demanding from a data perspective.

Multipliers

Direct and Indirect Effects

An important characteristic of input–output models is that they provide a detailed industry-by-industry breakdown of the predicted effects of changes in demand. It is sometimes useful, however, to provide a summary statement of these forecasts by examining direct and indirect effects. This can be done by constructing multipliers based on the estimated recirculation of spending within the region; recipients use some of their income for consumption spending, which results in further income and employment. This generated effect appears at three levels. First, the direct effect of production changes. For example, an increase in tourists staying in a hotel will directly increase the output of the hotel industry. Indirect effects result from various rounds of respending of, for example, tourism receipts in linked industries. If more hotel rooms are rented, then more breakfast products or cleaning services are needed. This will have an indirect effect on these industries. The third level of effects consists of the induced effects. These effects only occur in a closed input–output model. In this case, the household industry is converted into an endogenous industry, which means that it responds to a change in income. The induced effects also include changes in economic activity resulting from household spending of income earned directly or indirectly as a result of tourism spending, for example. These households may be employees of restaurants, who spend their income in the local economy.

The three most frequently used types of multipliers estimate the effects on (i) outputs of the industries, (ii) income earned by households due to new outputs, and (iii) employment expected to be generated because of the new outputs.

An output multiplier for the tourism industry can be defined as the total value of production in all industries of the economy that is necessary to satisfy $1's worth of final demand for the output of the industry. Income multipliers describe the impacts of final demand changes into changes in income received by households. Finally, the employment multiplier describes the number of jobs created because of one new job.

Multipliers can be derived as follows:

Output multiplier: (direct effect + indirect effect)/direct effect

Households income multiplier (type I): (direct effect + indirect effect)/direct effect

Households income multiplier (type II): (direct effect + indirect effect + induced effects)/direct effect

Employment multiplier: (direct jobs created + indirect jobs created)/direct jobs created

The size of the multiplier depends on several factors. First, it depends on the overall size and economic diversity of the region's economy. Regions with large, diversified economies that produce many goods and services will have high multipliers because households and businesses can find most of the goods and services they need in their own region. Also, the geographic scale of the region and

its role within the broader region play a role. Regions of a large geographic coverage will have higher multipliers, compared to similar small areas, because transportation costs will tend to inhibit imports (imports can be considered as leakage and have a negative effect on a multiplier). Regions that serve as central places for the surrounding area will also have higher multipliers than more isolated areas. Furthermore, the nature of the specific industries can have a significant effect. Multipliers vary across different industries of the economy based on the mix of labor and other inputs and the tendency of each industry to buy goods and services from within the region (less leakage to other regions). Tourism-related businesses tend to be labor-intensive. Therefore, they often have larger induced effects because of household spending rather than indirect effects. Finally, the year of the development of the input–output table should be taken into account. A multiplier represents the characteristics of the economy at a single point in time. Multipliers for a given region may change over time in response to changes in the economic structure as well as price changes. When comparing the sizes of the multipliers, it is important to distinguish the different effects that are taken into account.

Regional Multipliers

When working with a regional input–output table, regional multipliers can also be derived. Depending on the kind of table, relevant economic effects on the region, on other regions, or on the national economy can be computed. These effects are interesting, for example, when the government has to decide upon a new location for a military base or for a new main post office. If the government wants to use this relocation to stimulate a certain region, the regional effects on employment or income are relevant.

The effect of a change in final demand within a region (region 1) on the region, the regional (output) multiplier, can be derived as follows:

$$M_a = \frac{(X_1^1 - X_1^0)}{\text{impulse}},$$

where M_a is the regional multiplier, X_1^1 is the total output of region 1 after the impulse, and X_1^0 is the total output of region 1 before the impulse. The essence of an interregional input–output model is that it includes impacts in one region (region 2) that are caused by changes in another region; this effect is measured by the spillover multiplier. Later, we explain in more detail how the multipliers can be computed.

Advantages and Disadvantages of Regional Input–Output

The major advantage of input–output models is their internal consistency. All effects of any given change in final demand can be recorded. Important, and sometimes restrictive, assumptions made in the input–output model are that all firms in a given industry employ the same production technology (usually assumed to be the national average for that industry) and produce identical products. Because the tables are produced for a certain period, the model can become irrelevant as a forecasting tool when production techniques change. Other disadvantages are that the model assumes that there are no economies or diseconomies of scale in production or factor substitution, and that they do not incorporate the existence of supply constraints. Finally, input–output models are essentially based on a linear production technology; doubling the level of agricultural production will double the inputs, the number of jobs, etc. This reveals something of the inflexibility of the model. Thus, the model is entirely demand driven, implying that bottlenecks in the supply of inputs are largely ignored.

There are also some practical problems in (regional) input–output theory. The development of a new input–output table is very labor-intensive and expensive. This is mainly due to the fact that most information is gathered using microsurvey questionnaires. Another problem of this method is that interviewees, firms or households, are not able to give perfect answers. Sometimes, they do not understand the question, or they do not want to tell the truth and therefore the data are not always perfect.

Another problem is that the data are expressed in monetary terms. This is done because it is impossible to compare physical units, but monetary values may increase and decrease due to price changes. Still, input–output analysis is seen as a very clear and important method, which has its limitations but is often embedded as a module in more extensive models.

Numerical Example

Here, a simple illustration of input–output analysis is presented. This example presents an input–output table (Fig. 2) describing an economy with two regions that both have two industries, industry A and industry B. From Fig. 2, we can read, concerning the output, that industry A from region 1 delivers 10 units to its own industry, the internal deliveries. The industry also delivers 5 units to industry B in region 1, 6 units to industry A in region 2, and 2 units to industry B in region 2. Final demand, which goes directly to consumers, investments, and the government, consists of 15 units. Finally, 4 units are exported.

The input consists of 10 units of internal deliveries and a total of 14 units of intermediary deliveries. Furthermore, 6 units are imported and 12 units go to the value-added account, the labor payments, taxes, and profit.

From the transactions matrix, we move to the technical coefficients matrix (Fig. 3). The technical coefficients

		Region 1		Region 2				
		Industry A	Industry B	Industry A	Industry B	F	E	X
	Industry A	10	5	6	2	15	4	42
Region 1	Industry B	6	7	4	5	10	15	47
Region 1	Industry A	4	2	12	6	8	4	36
Region 2	Industry B	4	2	8	10	12	10	46
Region 2	M	6	4	2	2			
	V.A.	12	27	4	21			
	X	42	47	36	46			249

F = Final demand; E = Export; X = Total input or output; M = Import; V.A. = Value added

Figure 2 Transactions matrix.

$$\begin{vmatrix} 10/42 & 5/47 & 6/36 & 2/46 \\ 6/42 & 7/47 & 4/36 & 5/46 \\ 4/42 & 2/47 & 12/36 & 6/46 \\ 4/42 & 2/47 & 8/36 & 10/46 \end{vmatrix}$$

Figure 3 Matrix with technical coefficients (A).

matrix A can be obtained by dividing the row of internal and intermediary deliveries by the column of total output (X).

The first element of the matrix indicates that 10/42 units of input from industry A from region 1 are necessary per unit output. If industry A is the agricultural industry and industry B can be seen as the service industry, it means that the agricultural industry in region 1 needs 10/42 units of input from its own industry to produce 1 unit of output. The second element of the table indicates that the agricultural industry also needs 5/47 units of input from the service industry in region 1, 6/36 units of input from the agricultural industry in region 2, and 2/46 units of input from the service industry from region 2.

Then the (inverse) Leontief equation can be applied: $X = [I - A]^{-1} \times (F + E)$, where X is the total output column, I is the identity matrix, A is the technical coefficients matrix, F is the final demand vector, and E is the export vector.

For this example, the total output can be calculated as follows:

$$X = \left\| \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{vmatrix} - A \right|^{-1} \times \begin{vmatrix} 19 \\ 25 \\ 12 \\ 22 \end{vmatrix}$$

$$= \begin{vmatrix} 0.76 & -0.11 & -0.17 & -0.04 \\ -0.14 & 0.85 & -0.11 & -0.11 \\ -0.10 & 0.04 & 0.67 & -0.13 \\ -0.10 & -0.04 & -0.22 & 0.78 \end{vmatrix}^{-1} \times \begin{vmatrix} 19 \\ 25 \\ 12 \\ 22 \end{vmatrix}$$

$$= \begin{vmatrix} 1.43 & 0.21 & 0.45 & 0.18 \\ 0.31 & 1.24 & 0.37 & 0.25 \\ 0.28 & 0.14 & 1.70 & 0.32 \\ 0.27 & 0.13 & 0.56 & 1.40 \end{vmatrix} \times \begin{vmatrix} 19 \\ 25 \\ 12 \\ 22 \end{vmatrix} = \begin{vmatrix} 42 \\ 47 \\ 36 \\ 46 \end{vmatrix}$$

Here, the computed X is identical to the last column of the transactions matrix, as it should be. Now, we can simulate an impulse in this small economy and compute the multipliers.

The multipliers we are searching for are the elements of the inverse Leontief matrix. For example, the value 0.31 in the second row means that a unit increase of demand in industry A of region 1 leads to an increase of 0.31 in industry B in this region. The overall effect of a unit increase in this sector is the sum of the values in the first column of this matrix: $(1.43 + 0.31 + 0.28 + 0.27) = 2.29$. This is the value of the output multiplier for sector A in region 1. Thus, an increase in final demand of 10 units for sector A in region 1 would have a total effect of 22.9 on the whole economy, 17.4 of which would materialize in region 1 and the rest in region 2.

See Also the Following Articles

Data Envelopment Analysis (DEA) • Regional Science • Regionalization and Classification

Further Reading

Isard, W., Azis, I. J., Drennan, M. P., Miller, R. E., Saltzman, S., and Thorbecke, E. (1998). *Methods of Interregional and Regional Analysis.*

McGregor, P. G., and McNicoll, I. (1992). The impact of forestry on output in the UK and its member countries. *Regional Studies* **26,** 69–80.

Miller, R. E., and Blair, P. D. (1985). *Input–Output Analysis: Foundations and Extensions.* Prentice Hall, Englewood Cliffs, NJ.

Paelinck, J. H. P., and Nijkamp, P. (1975). *Operational Theory and Method in Regional Economics.* Saxon House, Westmead, UK.

Stynes, D. J. (1997). Economic impacts of tourism. A handbook for tourism professionals. Illinois Bureau of Tourism. [Available at http://www.tourism.uiuc.edu/itn/etools/eguides/econimpacts.pdf]

Regional Science

Darla K. Munroe
University of North Carolina, Charlotte, North Carolina, USA

James J. Biles
Western Michigan University, Kalamazoo, Michigan, USA

Glossary

entropy maximization Extension of the traditional gravity model, used to model spatial interaction as a stochastic, rather than deterministic, process.
geographic information system Computerized environment for collecting, processing, analyzing, and displaying spatial data.
GIScience Theory and science behind the tools and technologies of geographic information systems (GIS) and the application of GIS in scientific research.
linkages Interdependencies between sectors that comprise an economy.
location–allocation models Operations research models designed to identify the optimal location of a facility given the location of markets, suppliers, and raw materials; the goal of such models is to determine the location that will minimize costs or maximize profits.
multiplier Direct and indirect impacts of changes in final demand.
space economy The spatial structure of regional economies; geographic distribution of inputs and outputs, spatial variation in prices and costs, and geographic patterns resulting from economic processes.
spatial autocorrelation Tendency for observations of social and economic variables to be correlated according to their geographic location; empirical regularity in which the value of one geographic observation is influenced by the value of other (usually nearby) observations.
spatial interaction Movement of goods, people, information, and capital; geographic patterns resulting from economic processes of supply and demand, competition, and distance.

In this introductory review of methods commonly used in the field of regional science, the discussion focuses on those fields and subareas within which regional science has made and continues to make the greatest impact, in theory, in practice and in policy. Economic and spatial methods and their potential applications are considered, and some basic planning applications are discussed. Because of the traditional spatial focus of regional science, the tools and technologies of geographic information science have been quick to take hold.

Introduction

The advent of regional science in the 1950s sought to forge a dynamic framework for regional analysis. Unlike its neighbors in geography and economics, regional science explicitly recognizes that all economic processes exist in space as well as in time. An important point of departure is the attempt to link broad-scale, or regional, patterns to underlying social and economic processes. Traditionally, regions have been defined on the basis of some social, economic, or physical characteristic, drawing on the concept of nodality—functional relationships between cities and their hinterlands—or based on administrative or political boundaries. Notwithstanding these apparently clear-cut definitions, the regional concept becomes murky because it has been applied to a variety of spatial scales, from the very local to the international. In addition, regional boundaries are not static; rather, they may depend on the research problem in question and they may change over time. This discussion can be framed on the basis of a recent assessment, by Sergio J. Rey and Luc Anselin, of publication patterns within the discipline, and elaborated by looking at those fields and

subareas within which regional science has most greatly impacted practice and policy.

Economic Methods

Shift-and-Share Analysis

Shift-and-share analysis is a simple analytical framework used to identify patterns of regional economic change over time. Typically, the method makes use of employment data at the state or county level for two time periods to decompose change in economic structure into three components; national shift, industry mix, and regional competitiveness. The national-shift component corresponds to the change in regional employment that may be attributed to macro-level forces beyond the borders of the study area. The industry-mix effect expresses the expected change in employment had each sector of the regional economy followed its corresponding national growth rate; it represents a "proportional shift" due to differences between national and regional economic structures. The regional-competitiveness component comprises a "differential shift" between regional- and national-level sectoral growth rates that is the result of natural resource endowments, comparative advantage, and the effects of regional policy.

The sum of national, industry-mix and regional-competitiveness components equals overall change in employment between two time periods. The basic shift-and-share model may be expressed as follows:

$$E^R_{i(t+1)} = E^R_{i(t)} \cdot \left[(E^N_{t+1}/E^N_t) + (E^N_{i(t+1)}/E^N_{i(t)} - E^N_{t+1}/E^N_t) + (E^R_{i(t+1)}/E^R_{i(t)} - E^N_{i(t+1)}/E^N_{i(t)})\right], \quad (1)$$

where E represents employment, R identifies the region of interest, N refers to the nation, i indicates a given sector of the economy, and t and $t+1$ identify the two time periods. In general, shift-and-share analysis ascribes regional economic change to some combination of national, regional, and sectoral forces. If the national-shift component accounts for most employment change, national trends can be used to make regional forecasts. If the industrial-mix and/or regional-competitiveness components of the model display substantial shifts, the distribution of employment in specific sectors of the economy must be considered. For instance, if a large share of regional employment is found mainly in fast-growing activities, the regional economy has a "favorable" industrial mix and its growth rate will exceed the national average. The competitiveness component may reveal that some regions attract a greater share of employment in a particular industry because they have better access than other regions to important markets or inputs.

Although shift-and-share analysis is a purely descriptive technique, it remains a useful method for decomposing regional change. A recent analysis, for example, has employed a shift-and-share framework to identify the components of urban population change in the United States between 1950 and 2000. Though the shift-and-share technique identifies the relative importance of different components of regional change, it does not attempt to explain the processes that bring about these shifts. Notwithstanding its utility, regional scientists have identified a number of limitations with the technique. In spite of attempts to model industry-mix and regional-competitiveness components as a function of locational characteristics, the method lacks a theoretical foundation. As with many regional economic models, the results of shift-and-share analysis are highly dependent on the level of temporal, geographic, and sectoral aggregation. The technique is of limited utility as forecasting tool and ignores linkages between industries and spatial interaction effects.

In response to these criticisms, several researchers have attempted to extend the basic shift-and-share technique. Because the traditional model confounds some of the regional-competitiveness effects with industrial mix, regional competitiveness was divided, in one study, into a traditional regional shift component and an allocation effect. Another research team developed a "dynamic" shift-and-share method to obtain a more accurate indication of regional economic change in New England. Finally, several scholars have been at the forefront of attempts to reformulate traditional shift-and-share analysis econometrically in order to provide a better understanding of regional dynamics.

Economic-Base Model

The economic-base model is a useful tool for estimating the overall impacts of regional export activity. The debate surrounding the significance of the export base in regional economic growth dates back to the seminal contributions of Douglass North and Charles Tiebout in the 1950s and 1960s. North introduced the export theory of growth, in which regional economic growth is attributed to external demand for a region's goods and services, to describe the process by which regional economic development takes place. In the simple export-base model, local markets are too small to generate scale economies and other positive externalities. As a consequence, only significant external demand can generate the forward and backward linkages and increased division of labor necessary to promote development.

The traditional economic-base model distinguishes between two kinds of economic activity: basic and nonbasic. As indicated in Eq. (2), total regional economic activity is merely the sum of basic and nonbasic components:

$$E_T = E_B + E_{NB}, \quad (2)$$

where E_T refers to total economic activity, E_B indicates basic activity, and E_{NB} represents nonbasic activity. Basic (or export) activities serve demands beyond the boundaries of the region. These activities are derived from a combination of locational factors, comparative advantage, and historical accident. Nonbasic (or local) activity depends on the level of basic activities and serves demands within regional boundaries.

During the past several decades, many regional scientists have addressed the crucial issue of identifying basic economic activity. Due to the lack of reliable economic information at the subnational level, employment data are typically used. In general, three basic techniques exist to estimate the level of basic economic activity at the regional level: (1) the assumption method, in which some portion of regional employment is presumed to comprise the export sector, (2) the minimum-requirements method, in which the share of employment in each sector of the region is compared with the "minimum-required" share in similar locations, and (3) the location quotient, based on the ratio of shares of employment in each sector of the regional economy to employment in each sector of the state or national economy. As mentioned previously, the economic-base model is premised on the fundamental assumption that nonbasic economic activity depends on basic activities. Furthermore, nonbasic activity presumably comprises a constant share (k) of the total economy. Therefore, the model may be reformulated as follows:

$$\begin{aligned} E_T &= (k)E_T + E_B, \\ E_T - (k)E_T &= E_B, \\ E_T &= E_B/(1-k), \\ E_T &= (1-k)^{-1}E_B. \end{aligned} \tag{3}$$

As these equations show, total economic activity is a function of basic employment. The relationship between basic economic activity and total activity is specified by the economic-base multiplier $(1-k)^{-1}$, which reveals the overall impacts within the regional economy of a change in the basic sector.

Because the economic-base model is easy to estimate and makes use of readily available data, it remains an invaluable tool for regional impact analysis. However, the model suffers from several important limitations: it combines all export sectors into one multiplier, does not include linkages between industries, and ignores the role of household consumption. Furthermore, the traditional model is static and appropriate only for short-run analysis and relatively small regions. Like the other methods discussed here, the base model also fails to account for spatial interaction and spatial structure. In response to these shortcomings, regional scientists have attempted to expand and enhance the traditional economic-base model. Improvements include attempts to establish an empirical and theoretical basis for the model econometrically, development of "demo-metric" and other dynamic models, and attempts to incorporate space explicitly into the traditional model.

An econometric approach to the economic base initially proposed in the 1970s stated that basic economic activity at the regional scale is a function of economic activity in the "rest of the world." During the past two decades, econometric methods have been used to account for differences in the magnitude of economic-base multipliers and to confirm that nonbasic employment is driven not only by basic employment, but also by nonbasic employment in all other sectors of the local economy.

Demo-metric models link population changes with regional economic growth; these techniques were developed in response to concerns first voiced by Tiebout that population growth and migration may generate an endogenous growth process over and above the export base. Seminal contributions by others have incorporated population change into the economic-base framework, resulting in more accurate analysis of economic impacts and estimation of multipliers that vary over time. Recently, economic-base and regional-adjustment models have been fused to analyze relationships between population and employment change in the western United States. Other scholars have presented time-series applications in order to rectify the shortcomings of the static traditional model.

Notwithstanding the need to incorporate space explicitly into economic impact analysis, only a handful of scholars have attempted to move from traditional sectoral multipliers to more dynamic geographic multipliers. Recent examples involve using basic spatial econometric techniques to demonstrate that economic activity generates impacts not only locally, but also among neighboring communities.

Input–Output Analysis

Input–output (IO) analysis is a modeling technique that divides the economy into final demand and production and accounts for the direct and indirect interdependencies among different sectors. Several researchers have demonstrated empirically the link between economic-base and input–output models. The technique was introduced by Wassily Leontief in the 1930s and adapted for the purposes of regional analysis by Walter Isard in the 1950s.

Input–output analysis requires regional accounts that capture the transactions among the different sectors of the economy for a given period of time (typically 1 year). Table I offers an example of a regional input–output table. The regional IO table is essentially a double-entry accounting system. Rows of the table correspond to sales from a given sector to all sectors of the regional economy; columns represent purchases of intermediate inputs, labor, etc. made by each sector.

Table I Regional Input–Output Table for Yucatán, Mexico (1993)[a]

	Sector purchases					*Final demand*		
Sector sales	*Agriculture and mining*	*Food and textiles*	*Other manufacturing*	*Commerce, hotels, and restaurants*	*Services*	*Household consumption*	*Other final demand*	*Total output*
Agriculture and mining	**175,000**	**243,777**	**42,267**	**0**	**2108**	390,375	379,561	1,233,089
Food and textiles	**47,298**	**33,422**	**156**	**867**	**5063**	509,499	566,641	1,162,947
Other manufacturing	**7742**	**6920**	**210,051**	**3692**	**12,141**	157,201	1,310,810	1,708,557
Commerce, hotels, and restaurants	**71,297**	**128,386**	**451,502**	**60,033**	**114,749**	1,205,139	1,347,968	3,379,074
Services	**48,838**	**26,778**	**59,433**	**621,324**	**1,011,809**	5,497,833	662,960	7,928,975
Salaries	464,264	131,054	270,915	638,641	1,587,520			
Other value added	263,776	286,541	355,101	1,598,486	2,532,475			
Imports	154,875	306,068	319,131	456,030	2,663,109			
Total outlays	1,233,089	1,162,947	1,708,557	3,379,074	7,928,975			

[a] Values in thousands of Mexican pesos.

Table II Technical coefficients for Yucatán, Mexico (1993)

	Agriculture and mining	*Food and textiles*	*Other manufacturing*	*Commerce, hotels, and restaurants*	*Services*
Agriculture and mining	0.1419	0.2096	0.0247	0.0000	0.0003
Food and textiles	0.0384	0.0287	0.0001	0.0003	0.0006
Other manufacturing	0.0063	0.0060	0.1229	0.0011	0.0015
Commerce, hotels, and restaurants	0.0578	0.1104	0.2643	0.0178	0.0145
Services	0.0396	0.0230	0.0348	0.1839	0.1276

The portion of the table with values in boldface type identifies interindustry transactions; these are linkages among firms in the different industries that make up the regional economy. Final demand is composed of household consumption, government spending, investment, and exports. The row sum of interindustry sales and final demand comprises an industry's total output; the column sum of intermediate inputs, salaries, imports, and other value-added components equals total outlays. For the economy as a whole, total output must equal total outlays.

Table I is used for input–output analysis by making three critical assumptions. First, IO analysis assumes that each sector of the economy consumes inputs in fixed proportions (demand for intermediate inputs is a linear function of output). A second assumption is constant returns to scale, which precludes increasing returns in different industries. Finally, the model assumes no substitution between different inputs. Once these assumptions have been made, the fundamental input–output relationship may be expressed mathematically:

$$\sum X_{ij} + f_i = X_i, \tag{4}$$

which states that total output of a given sector (X_i) is equal to interindustry sales to all other sectors ($\sum X_{ij}$) plus sales to final demand (f_i). Subsequently, interindustry transactions and total outlays (X_j) may be used to estimate technical coefficients (a_{ij}) for each sector of the economy:

$$a_{ij} = X_{ij}/X_j. \tag{5}$$

Technical coefficients reveal the total direct input requirements for each industry per unit of output. In regional input–output analysis, these values correspond to inputs purchased within the region exclusively. The technical coefficients corresponding to the IO table for Yucatán are shown in Table II. Once technical coefficients have been estimated, the input–output relationship may be reformulated as follows:

$$\sum a_{ij}X_j + f_i = X_i, \tag{6}$$

which decomposes interindustry transactions into two components: the product of technical coefficients (a_{ij}) and total outlays (X_j). If technical coefficients are thought of as forming a square matrix (A) and final

Table III Leontief Inverse and Output Multipliers for Yucatán, Mexico (1993)

	Agriculture and mining	*Food and textiles*	*Other manufacturing*	*Commerce, hotels, and restaurants*	*Services*
Agriculture and mining	1.1770	0.2543	0.0333	0.0002	0.0006
Food and textiles	0.0466	1.0397	0.0016	0.0004	0.0008
Other manufacturing	0.0090	0.0092	1.1410	0.0017	0.0020
Commerce, hotels, and restaurants	0.0780	0.1353	0.3108	1.0218	0.0176
Services	0.0715	0.0679	0.1126	0.2154	1.1501
Output multiplier	1.38	1.51	1.60	1.24	1.17

demand (f) and total output (X) as column vectors, the input–output model may be derived in matrix terms as shown in Eq. (7):

$$\begin{aligned} AX + f &= X, \\ f &= X - AX, \\ f &= (I - A)X, \\ (I - A)^{-1} f &= X. \end{aligned} \tag{7}$$

As these equations indicate, the input–output model ascribes total output in the regional economy to final demand and a multiplier process captured by the "Leontief inverse" $(I - A)^{-1}$. Unlike the economic-base model, IO analysis provides multipliers for each sector of the economy. Three basic input–output multipliers may be identified: output multipliers, which quantify direct and indirect impacts on production of goods and services; income multipliers, which facilitate assessment of effects in terms of returns to labor (wages); and employment multipliers, which quantify job creation effects. The Leontief inverse and output multipliers for Yucatán, Mexico are shown in Table III.

The IO model is typically used by simulating exogenous changes to final demand and using multipliers to estimate resulting impacts. The regional input–output framework can be extended to a multiregional context in order to account for linkages between different locations, as well as different sectors of the economy. A multiregional input–output (MRIO) table partitions transactions into intraregional and interregional economic activity. In addition to the basic data required for regional input–output analysis, some estimate of the interregional distribution of interindustry transactions and final demand must be obtained.

Notwithstanding its utility, the input–output model has some shortcomings. At the regional level, salaries and consumer spending, rather than interindustry linkages, may have the greatest impact on the economy. The basic IO framework, however, fails to incorporate household consumption. In response, an alternative, "Type II" multiplier, which endogenizes consumption effects, has been developed. In addition, the income and consumption portions of the model have been partitioned to estimate interrelational income multipliers that account for the distribution of income among different groups. Social accounting models (discussed later) further decompose economic interdependencies among institutions, factors of production, and activities.

The three previously mentioned assumptions (proportionality, constant returns, and no substitution), are also limitations because they violate basic economic theory. Other assumptions, including perfectly elastic supply and price-inelastic final demands, result in overestimated multiplier effects. During the past decade, several regional scientists have employed econometric methods to address some of these issues. The resulting "integrated" IO/econometric models have been used to forecast changes in technical coefficients, to estimate changes in final demand, to allow for increasing returns, and to incorporate the impact of changes in supply and demand on prices and substitution.

Social Accounting and Computable General Equilibrium Models

Social accounting models (SAMs) are an extension of input–output analysis and provide far greater detail about the relationships between sectors of the economy, social groups, and economic agents. Because they capture information about factors of production and institutions such as households and government, SAMs are especially useful in assessing the social welfare implications of economic policy decisions. The basic social accounting framework, established by Richard Stone in 1961, has since been extended by several other scholars. Whereas IO analysis concentrates primarily on interindustry transactions, social accounting models divide the socioeconomic system into three equally important components: factors of production, institutions, and activities.

The structure of a social accounting matrix facilitates identification of the flows of resources throughout the regional economy. For example, as shown in Fig. 1, production activities transfer "value added" to factors of

SOCIAL ACCOUNTING MATRIX		EXPENDITURES				
		Production activity	*Factors of production*	*Institutions*	*Other (exogenous)*	TOTAL
RECEIPTS	*Production activities*	Interindustry transactions		Consumer expenditures	Exports	
	Factors of production	Value added			Net factor income from abroad	
	Institutions		Wages and profits	Transfers and taxes	Transfers (remittances)	
	Other (exogenous)	Imports		Imported consumer goods		
	TOTAL					

Figure 1 Stylized social accounting matrix.

production (land, labor, and capital). Owners of the factors of production (workers, for example) pass on their income to institutions (such as households). Institutions subsequently devote a portion of their income to the consumption of needed goods and services (provided by production activities and imports). Like their IO counterparts, SAM multipliers are completely demand driven. Policy impacts are typically estimated by simulating a change in exogenous accounts, which propagates impacts throughout endogenous components of the model (institutions, factors of production, activities, and other final demand). Social accounting models have been used in a variety of contexts at a number of spatial scales, particularly in developing countries. Seminal examples include the application to a large-scale development project in Malaysia, the use of social accounts at the village level in Mexico, and development of a multiregion model to estimate the impact of taxes on counties in Iowa. As an extension of the IO framework, SAMs are the fundamental component of computable general equilibrium (CGE) models. CGE models are built on the social accounting framework and use neoclassical economic theory to provide a more realistic representation of behavioral and structural relationships among economic agents. Consequently, these models eliminate many of the restrictive assumptions associated with the traditional input–output model.

The most common "Walrasian" (after Léon Walras, 1834–1910) CGE model is based on assumptions of perfect markets and optimal behavior among producers and consumers; the model simulates economic behavior in which prices and output (supply and demand) adjust to clear markets. Consequently, CGE multipliers tend to be somewhat smaller than their IO counterparts (though they converge to IO multipliers in the long-run). Like IO and social accounting models, CGE models have been used in a wide variety of applications. Examples include the Global Trade Analysis Project, which assesses distributional consequences of trade liberalization for individual nations and regional blocs; a regional CGE model for Scotland (a Macro–Micro Model of Scotland, or AMOS); and the three-region Brazilian Multisectoral and Regional/Interregional Analysis Model (B-MARIA), focusing on regional inequalities and structural change in Brazil.

Although CGE models offer many advantages over the traditional IO and SAM models, they require massive amounts of data and depend greatly on the calibration of parameters that are frequently unknown or unavailable. In addition, they are much more complex than other models are, making it difficult, if not impossible, to trace impacts throughout the economic system. Furthermore, CGE models may not provide significantly better results compared to those obtained from traditional input–output or social accounting models.

Spatial Methods

Space and spatial relationships have always been central to the discipline of regional science, beginning with Isard's seminal work on the space economy. More recently, these ideas have been integrated with more sophisticated mathematical analyses by Masahisa Fujita, Paul Krugman, and Anthony Venables. In a broader perspective, space provides an organizing principle: different processes and actors are linked through their location. A suite of spatial methods commonly used in regional science research is reviewed in the following

discussions. Because regional science has always been concerned with the rigorous quantitative analysis of regional processes, it has provided fertile ground for the development, application, and innovation of spatial social science.

Geographic Information Science

The analysis of space and place in social science has become increasingly important in recent years, and the foundations of spatially integrated social science have been discussed. Interestingly, advances in geographic information science (GIScience) have, in large part, been driven by increasingly available spatial data and by developments in software and hardware that facilitate spatial modeling. Precisely because regional science has been at the forefront of thinking about space and spatial relationships, the integration of GIScience into the discipline has been prolific.

Some researchers have discussed a particularly synergistic aspect of GIScience and its potential contribution to the discipline. Visualization has long played a crucial role in shedding insights about patterns and processes, stimulating new theory and applications. The tools and technologies within GIScience have provided further impetus for new model development and testing. Particular emphasis has been placed on the fruits of quantitative spatial thinking and on geocomputation (quantitative analysis of spatial data in which computing plays a pivotal role). Increases in both computing power and available spatial data have greatly enhanced inferential capabilities of spatial models, in that rigid statistical assumptions (i.e., normality and independence) can be relaxed with permutation or bootstrapping approaches. Spatial modeling techniques have been employed by regional scientists in a host of applications, including urban and regional planning, business location analysis, transportation, studies of epidemiology and crime, community development, and environmental and natural resource management. All of the methods used in the models predate computerized GIS tools, but geographic information systems and GIScience have been pivotal in the expansion and application of these tools.

Spatial Interaction Models

Many relevant social science questions boil down to identifying the driving forces of interaction through spaces. Therefore, techniques have emerged within regional science as a means of measuring and estimating the scope, nature, and volume of flows. The initial application of spatial interaction was borrowed from social physics. These deterministic models were replaced with entropy maximization, which views spatial interaction as a stochastic process. Spatial interaction models were expanded to formulate a theory of movements, in which flows of goods and people over time collectively form the development of a region. Relevant flows shaping regional development can be movements of people (migration), but can also be other varied processes, such as international or interregional trade, shopping trips, or commuting patterns.

Spatial interaction models may be applied to two basic questions. The first type of model would involve the explanation or prediction of spatial choice, i.e., allocation of flows to a particular destination. Many of the techniques for spatial choice modeling have emerged from transportation and demography, wherein flows are explicitly part of any study. The formulation of these models is generally as follows:

$$I_{ij} = \frac{E_i A_j}{d_{ij}^b}, \tag{8}$$

where I represents the interaction between locations i and j, Edescribes the volume of flows originating at point i, A describes the opportunities to be found at point j, d is the distance between them, and b represents the degree of distance decay, or how quickly interaction drops off with increasing distance.

There are also competing destinations models, which look at the choice of a particular destination, given several options. Continuing with the retail example, assume the following conditions:

$$P(I_{ij}) = \frac{(A_j/d_{ij}^b)}{\sum_j (A_j/d_{ij}^b)}, \tag{9}$$

where now there is some probability (P) of interaction between locations i and site j based on the characteristics of j compared to all possible retail locations. This has been further refined into the intervening-opportunities model, asserting that, in an urban area, the consumer will shop at the first service area that meets his/her minimum expectations.

A larger suite of location–allocation models has been developed based on spatial interaction principles. These models were developed to provide recommendations regarding the best locations of public facilities, such as health care services and schools, and to address private sector issues such as location of retail stores, warehouses, or factories. Location–allocation models attempt to optimize aggregate behavior (flows of interaction), satisfying multiple criteria:

$$\min Z = \sum_{i=1}^{n} \sum_{j=1}^{m} w_i d_{ij} a_{ij}, \tag{10}$$

where Z is some objective function to be minimized (such as total travel cost), n is the number of origin

points, m is the number of potential destinations, w is some weight for each origin, and d is distance. The choice of allocation of source i to destination j is flagged by a; if the site is chosen, its value is 1; otherwise it is 0. More advanced models specify hierarchies of centers. The most important implication of this work is the assertion that observable regional patterns emerge as a result of individual decision-making processes, which are subject to a great deal of variation and uncertainty. There have been notable extensions of this basic approach. This approach was linked in one study to an underlying behavioral model (based on cost minimization) that is congruent with economic theory to the larger macro model suggested by gravity or entropy-maximizing approaches. Other researchers have provided a dynamic extension of these models, specifying the mathematical relationships in which the models could more flexibly be used to represent evolving and changing interactions among units.

Location Analysis

Traditional location analysis includes the central place theory of Walter Christaller, the refinements by August Lösch, Ricardian (David Ricardo) ideas of land rent, and Johann von Thünen's insights on market accessibility. At root, these models state that the location of firms and households relates to transportation costs (both in the movement of goods and people) and the accessibility of inputs to production processes, and back to the resulting output markets. The relationship between space or distance and firm profit can be denoted as follows:

$$\pi_i^{m_i} = P_i Q_i - A_C^{m_i} - t_i m Q_i - R^m, \qquad (11)$$

where π represents the profits of firms in industry i at distance m miles from the center of the city, P and Q represent price and output, A_C is the average production cost, t is the per unit output transport cost, and R denotes land rent at location m. Therefore, both per unit costs and average costs are affected directly by transportation costs, leading to discernible spatial patterns in economic activity even at points of equilibrium in a competitive market.

Location theory has been applied to explain urban density, labor migration, and land use patterns. Regional scientists learned long ago that strategic location behavior by firms can lead to industrial agglomeration, economies of scope and of scale, relating very clearly to Marshallian (Alfred Marshall) external economies. Classical location theory assumes that markets are competitive and consumers are homogeneous, but factors of production may be differentially spatially distributed, and firms will choose location strategies to minimize costs. As firms attempt to minimize transportation costs, for example, they may choose to locate in an area where they can exploit economies of scope and scale. Paul Krugman has pointed out, for instance, that firms may choose to locate in an area with a greater pool of highly skilled workers.

The practical applications of location analysis are numerous. First and foremost, location analyses have been employed in applied business questions, such as retail and market area analysis. Second, a major thrust of location analysis within regional science has been on examining and explaining urban–rural disparities and providing guidance to policymakers. These theories relate to long-standing rural underdevelopment, but also serve as a guide to policymakers regarding best-practice policy instruments to stimulate rural economic growth and development.

Spatial Statistics and Spatial Econometrics

Econometric analysis of areal units poses a particular set of problems and challenges for statistical estimation and inference. Isard stressed the importance of thinking about space as continuous phenomenon, but, in practice, data are often assigned to irregular units (such as administrative boundaries) due to data collection techniques, confidentiality requirements, and other logistical concerns. Therefore, the assignment of individual behavior to a possibly arbitrary spatial unit can induce spatial dependence or heterogeneity that may mask or obscure individual decision-making processes. Spatial autocorrelation necessarily induces heteroscedasticity as well as other violations of ideal conditions for econometric analysis, such as normality or independence among observations. The term "spatial econometrics" was coined by Jean Paelinck and Paul Klaassen in 1979, based on five principles: spatial interdependence among units, asymmetry in this dependence, importance of explanatory factors located in other areas, difference between prior and subsequent interaction, and the explicit modeling of spatial relationships.

Interaction effects are expressed in terms of an N by N spatial weighting functions and a spatial parameter. Theoretically, the spatial relationship between two observations in a cross-sectional data set would be given by the covariance matrix; e.g., σ_{ij} represents the covariance between observations i and j. Because estimation of the full covariance matrix is not possible, the researcher must impose a structure on that model. The number of interactions among N observations can be as many as N^2, which cannot be identified (this is referred to as the incidental parameter problem). Instead, the most common way of specifying the structure of spatial dependence when the units of analysis are discrete objects (e.g., plots of land) is by specifying a spatial weighting function such that a matrix W defines a nonzero w_{ij} as

the neighborhood j around observation i. The most commonly used spatial weights can represent connectivity, contiguity, adjacency or association, or the distance decay of a particular process. Weights that define the magnitude of interaction or potential (the so-called spatial economic multiplier) can also be defined. The specification of the interaction effect in a spatial lag model amounts to a weighted average of neighbors' choices, weighted by w_{ij}. Writing out the spatial lag for observation i in scalar terms makes this readily apparent:

$$[Wy]_i = \sum_j w_{ij} y_{ji}. \tag{12}$$

There are two categories of spatial pattern: spatial dependence, or autocorrelation, and spatial heterogeneity. The most widely used statistic to characterize spatial autocorrelation for continuous variables is Moran's I:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{13}$$

where x is the variable in question, and the weights matrix w is the assumed structure of interactions between observation i and j. This statistic is somewhat analogous to a correlation coefficient, except that only the deviations between each observation and its neighbors, defined by the weights matrix, are considered. There are local extensions to this global statistic, such as the local indicator of spatial association:

$$I_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \sum_j w_{ij}(x_j - \bar{x}), \tag{14}$$

where an I value is assigned to each observation i, which indicates the standardized difference between that observation and the mean relative to a weighted average of its neighbors. In this way, researchers can identify hot spots or spatial outliers within the data set, obtaining a more complete understanding of how individual observations are contributing to overall patterns within the data.

Spatial econometric techniques have been employed in a host of applications. Such approaches have been used in studies of regional growth and convergence in the United States and Europe and in myriad land applications, such as analyzing land use and land value processes, in assessing local interdependencies that arise from land use externalities, and in studying the impact of open space on land values. Other noteworthy applications include studies of public finance and taxation, firm spatial competition for market area, and knowledge spillovers. A current challenge for regional science is the consideration of spatiotemporal patterns and processes. Storage and manipulation of repeated observations on a single object are nontrivial; and developing systems that can separately identify spatial and temporal interactions remains an onerous task, because such systems are analytically intractable due to multiple equilibria and multidimensional interactions.

Planning Applications

Population and Migration Models

Regional scientists are in a unique position to evaluate both the issue of space and the concept of "region," thus adding an important perspective to population and migration studies. Particular examples have included regional population forecasting. There is a need for general equilibrium approaches in studying migration flows and other demographic changes so that the researcher can effectively model the interplay between urban and rural, across different household types, and between local and nonlocal goods.

Regional models of migration flows are often based on assumptions that two types of factors determine an individual's propensity to migrate. The first factor includes characteristics of the individual, such as wealth, education, age, employment history, and ethnicity. The second factor includes characteristics of the points of origin and destination (e.g., relative wages and unemployment rates in each location, so that the migrant makes a decision regarding how the move will change his/her wealth and employment status). The regional element is important both in terms of specifying the migration question and in estimating its impact. Thus, the first objective might be to understand and characterize regional drivers of population change: what are the forces that make some regions more attractive destinations than others, and what is the relative intensity of these driving forces? There are many factors beyond simple wage rates, such as amenities, that influence a potential migrant's utility in a way that often is spatially complex. National and regional economies also change over time, and correspondingly, economic opportunities change. For instance, over the past several decades, much of the population has moved from employment in agriculture, to industry, to service sector activities, and these broad-scale changes have profound impacts on the demographic makeup of regions, as migrants continually seek new opportunities. A second objective is understanding the nontrivial feedback effects of population movements on the regional economy in terms of labor pool, tax base, and related impacts.

Other work in regional science has concerned creating population projections for regions. Understanding of multiregional population trends involves characterizing two aspects of the demographic makeup: the state of the population at a given point of time and the propensity of that state to change, as well as the intensity of the change.

Thus, most techniques for forecasting population changes are based on Markov transition probability matrices. Markov transition probabilities can be specified as follows:

$$p_{ij} = P\left(\frac{X_{n+1}=j}{X_n=i}\right), \tag{15}$$

where X is a random variable representing the state or value of the system at time n, and p_{ij} represents the probability that X will be in state j at time $n+1$, given its value at time n. Thus, any change in population is a function of its value in past periods. Extensions on this basic approach have been to relax the restrictive assumption of stationarity, in that each X state is independent of past values of X. One approach is to include lagged values in the relationship to estimate explicitly the effect of prior periods on the current transition probabilities. Another general set of approaches is to decompose changes in population into various effects such as age, time period, and cohort effects. The shift-and-share model can be adapted to decompose these various components of population change. Maximum-likelihood principles can be used to model origin, destination, and time effects on migration flows via an expectation conditional maximization (ECM) algorithm.

Regional science demographic studies have been concerned with linking changes in urban and regional economies to population trends. Trends in national and regional urban systems related to postindustrialism and the impacts on population flows have been considered. At a national level, there may be overall trends in population movement (e.g., from north to south). At a regional level, population movements may relate much more directly to employment patterns. To identify the extent to which total movement within an area is leading to population change, the demographic efficiency can be defined as follows:

$$E_j = 100\frac{N_j}{T_j}, \tag{16}$$

where T refers to total migration and N refers to net migration. In this manner, the researcher can identify the impact of regular movements among employment centers (such as job turnover, etc.) vs. systemic changes, and where economic functions within regions are changing (e.g., as employment opportunities move from north to south in response to a changing economy). Regional science approaches to population modeling have also responded to the changing nature of regions. The rarely studied but increasingly important phenomenon of "micropolitan" areas, corresponding to changing regional economies, and counterurbanization trends as employment in traditional sectoral activities declines, have been examined.

Multiobjective Decision Making and Operations Research

Because of its inherently interdisciplinary and applied orientation, applied decision-making has long been an important part of regional science techniques. In a decision-support system, a computer or set of computers is used to integrate and manage data for problem solving, planning, and decision making within a particular topic domain. Such a system integrates scientific knowledge with data in an overarching framework of modeling, analysis, or assessment to facilitate planners or decision makers in their plan-making or decision-making efforts. Examples include demography, facility location and allocation, watershed planning and management, air quality assessment, urban growth management, and project impact assessment.

One well-known linear programming technique is the transportation model. In this model, something is "shipped" from a set of discrete sources to a set of discrete destinations. This type of model applies to transportation problems, but it can also be used in any problem when a set of independent sources is supplying some set of independent demands. The amount to be shipped from each source to each of the destinations, while minimizing the total shipping cost and satisfying all supply limits and demand requirements, can be specified as follows:

$$\min z = \sum_i \sum_j c_{ij}x_{ij}, \tag{17}$$

subject to

$$\sum_j x_{ij} = a_i; \tag{18}$$

output from each source equals supply,

$$\sum_i x_{ij} = b_j, \tag{19}$$

input to each destination must equal demand, and $x_{ij} \geq 0$, $\forall i$, $\forall j$. The transshipment model is a variation of the basic transportation problem in which supply locations transport products to centers of demand via intermediate nodes. A variation of the transshipment model, the Beckmann–Marschak problem, includes production/processing activities at each transshipment location. It has been described as a two-commodity spatial allocation process that examines the pattern of production, manufacturing and trade of a single commodity for a system of n suppliers, transportation centers, and m markets.

Regional scientists were among the first scholars to extend traditional linear programming techniques to a multiregional context. Isard applied linear programming to a set of regions given a set of limited resources, production technologies, factor prices, and commodities that were interrelated through trade. However, perhaps due to the limited computing technology available at the time, his model did not consider change over time explicitly.

The first comprehensive source on these methods within regional science can be found in the work of Peter Nijkamp and A. van Delft in 1977. GIScience has provided a new thrust to multicriteria decision support in that it provides a means for linking disparate processes and decision makers through spatial identifiers, and to thus create a shared database of relevant information. Advances in computing power, storage, and processing speed have also facilitated quantum leaps in development and application of these approaches.

See Also the Following Articles

Geographic Information Systems • Modeling Migration • Regional Input–Output Analysis • Regionalization and Classification • Spatial Autocorrelation • Spatial Discounting • Spatial Econometrics

Further Reading

Alonso, W. (1964). *Location and Land Use.* Harvard University Press, Cambridge.

Anselin, L., and Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In *Handbook of Applied Economic Statistics* (A. Ullah and D. E. A. Giles, eds.), pp. 237–289. Marcel Dekker, New York.

Bailey, T. C., and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis.* Longmann Scientific & Technical Publ., Essex.

Fotheringham, A. S., and O'Kelly, M. E. (1989). *Spatial Interaction Models: Formulations and Applications.* Kluwer Academic Press, Amsterdam.

Getis, A., and Fischer, M. M. (eds.) (1997). *Recent Developments in Spatial Analysis: Spatial Statistics, Behavioural Modelling and Neurocomputing.* Springer, Heidelberg.

Ghosh, A., and Rushton, G. (1991). *Spatial Analysis and Location–Allocation Models.* Van Nostrand Reinhold, New York.

Hewings, G. J. D. (1985). *Regional Input–Output Analysis.* Sage Publ., Hollywood, CA.

Isard, W., Asis, I., Drennan, M., Miller, R., Saltzman, S., and Thorbecke, E. (1998). *Methods of Interregional and Regional Analysis.* Ashgate Publ., Aldershot.

Isserman, A. (1980). Estimating export activity in a regional economy: A theoretical and empirical analysis of alternative methods. *Int. Region. Sci. Rev.* **5,** 155–184.

Jackson, R. (ed.) (2003). *The Web Book of Regional Science.* Regional Research Institute–West Virginia University Available on the Internet at www.rri.wvu.edu.

Rey, S. (2000). Integrated regional econometric + input–output modeling: Issues and opportunities. *Pap. Region. Sci.* **79,** 271–292.

van Delft, A., and Nijkamp, P. (1977). *Multi-Criteria Analysis and Regional Decision-Making.* Kroese, Leiden.

Regionalization and Classification

Ron Johnston
University of Bristol, Bristol, United Kingdom

Glossary

aggregation Inductive procedures for grouping like individuals on predetermined criteria.
association analysis A technique for dividing a population into groups/classes using presence/absence data.
classification The grouping of individuals into classes (categories) according to predefined criteria.
factorial ecology The classification of areas according to the socioeconomic and other characteristics of their populations.
geodemographics A particular example of factorial ecologies, with the area classifications used for targeted marketing campaigns.
modifiable areal unit problem (MAUP) A major problem for statistical analyses using regionalized data because of the large number of different ways in which the same spatial data set can be classified on predetermined criteria including contiguity constraints.
redistricting A particular example of the regionalization problem, involving the creation of continuous territories to be used in legislative elections, to which the MAUP applies.
regionalization The creation of classes in which the member individuals are spatially contiguous.

This article discusses methods for either dividing a population or grouping a set of individuals into mutually exclusive classes; in regionalization, these classes are identified by their spatial properties.

Introduction

Classification

Classification involves allocating individuals to classes (or types, categories, groups, sets, etc.) on the basis of predefined criteria. It is the foundation for much scientific activity, placing individuals into classes with like individuals about which generalizations can be made. In many cases, the classification is straightforward, based on readily-applicable rules, often associated with binary decision making: a person is either male or female, for example. Many other situations have no such easy rules, however, and there are no a priori clear distinctions between classes either because of the absence of theory/empirical work that would establish their nature and boundaries or because the criteria on which the classification is to be made (e.g., the height and weight of humans) are continuous, not binary. In the latter situation, it is assumed that for the purposes of analysis, and perhaps subsequent action, there are groups of individuals who are similar enough on the relevant criteria to be treated as members of a class. The questions that then arise are how should those classes be determined and what are the class boundaries? These questions have been addressed in several disciplines, although many of the procedures employed derive from work in either psychology or ecology (often termed numerical taxonomy).

Regionalization

Regions are a particular case of this general category of classes. The region is a core concept in geographical scholarship, and it is not infrequently used in other social sciences. It is usually defined as an area possessing a degree of unity on certain organizing principles or criteria, which distinguish it from surrounding areas. Thus, most large areas can be divided into a mosaic of component regions, separate parts that differ from their neighbors in some ways. Each of these regions may be distinct, sharing no characteristics with any of the others, or some may be similar in all (or the majority of) respects but are not contiguous. This leads to the separate identification of

(i) regions, which are contiguous blocks of territory that are homogeneous on predefined criteria, and (ii) regional types, which are two or more separate contiguous blocks of territory that are homogeneous on the predefined criteria but differ on them from neighboring territories. Scale is crucial in the process of defining regions: an homogeneous area defined at one scale may be heterogeneous at another, implying a hierarchy of regions at different scales, with homogeneous regions at one scale nesting into larger regions at another. A river valley may be a region at one scale, for example, but at another it may be divided into the flood plain, the valley slopes, and a watershed plateau dividing it from the next valley.

In its original formulation, the concept of a region was applied to the physical environmental characteristics of areas plus their occupation and manipulation by humans. Thus, the areal differentiation of the earth's surface—the mosaic of regions—was a function of environmental differences and human engagement with them. However, with the emergence of spatial analysis within geography and related disciplines, such as regional science, an alternative definition was introduced in which regional homogeneity resulted from patterns of spatial organization reflecting patterns of movement focused on regional cores (which is the foundation of major texts). This led to a twofold categorization of (i) formal regions, which are areas with a common unity based on local characteristics, and (ii) functional regions, which are areas whose unity was based on them being tributary to the same organizing node, such as a port and its hinterland. The traditional definition covered formal regions; many of the developments in spatial analysis were associated with functional regions. Scale was important in the latter also: the functional region associated with a small town, for example, might nest within the larger tributary area of a large city.

The delineation of both types of regions was long undertaken through map comparison and analysis. Maps showing the distribution of the predetermined criteria were overlain and boundaries identified, usually with some difficulties in deciding exactly where one region ends and another begins. For functional regions, these maps showed flow patterns; again, the goal was to identify areas with common directionality to their flows—almost certainly focused on a central node.

During the past four decades, these "subjective" methods have been superceded by more "objective" procedures using computer algorithms. The pioneer work was done in rural sociology. Berry extended and promoted the approach using a "geographical field theory" that was applied in a variety of contexts. For the definition of regional types (mostly comprising sets of formal regions, although comparable sets of functional regions are feasible—city hinterlands need not be composed of contiguous areas), there is no difference from the general process of classification: regionalization is just a particular case of the larger procedure, in which the observation units to be classified are spatially defined blocks of territory. However, for the definition of regions *sensu stricto* an additional criterion is introduced to the classification procedure—contiguity.

Most regionalizations involve the use of individuals whose nature is predetermined outside the research enterprise. Many formal regionalizations, for example, involve the classification of administrative areas whose boundaries are predefined. In such cases, the characteristics of the populations of these areas are used as the criteria for regionalizing. The areas may be internally heterogeneous, with several separate components that differ among themselves. These separate areas cannot be identified, however, and the regionalization has to use the areal units provided: they are "forced on" the analyst rather than defined "naturally" from the lowest possible scale—in many cases, individual persons or households.

The Approach

Procedures for classification and regionalization were topics of substantial interest two to three decades ago when quantitative procedures for the social sciences were being explored. Today, most of the classifications and regionalizations undertaken in these disciplines employ standard routines in statistical software packages (such as SPSS) and there is little innovation. Thus, the methods set out in the following section are standard procedures. However, workers in cognate disciplines use the massive computing power now available to explore other forms of classification, notably in the analysis of remotely sensed data. Relatively little of this work has infiltrated the social sciences. Such procedures are not readily illustrated using small, simple examples as undertaken here for the established methods. They are introduced here, however, as are methods of regionalizing using point data, techniques that are becoming increasingly popular in the analysis of clusters of events such as disease outbreaks. (There is a much larger literature on various aspects of spatial pattern analysis.)

One of the major findings of research on regionalization during the 1970s was its relationship to an issue in the analysis of much spatial data known as the modifiable areal unit problem. The nature of this problem is set out here as one that has to be faced whenever classifications and regionalizations are undertaken and one that is linked to a larger problem within quantitative social science, the ecological inference problem.

Inductive Classification and Regionalization: Aggregative Procedures

Most of the classification algorithms employed in regionalization exercises do not incorporate the contiguity constraint. Instead, the goal is to define homogeneous regional types, whether formal or functional: where those identified types do not comprise a single block of contiguous territory, they may then be subdivided into separate regions that do meet that criterion. Thus, in discussing methods of regionalization, the focus here is on classification procedures. Particular cases that differ from this are discussed later.

In formal terms, most regionalization–classification algorithms proceed through an analogy with the analysis of variance procedure within the general linear model. The goal is to minimize within-region variance and maximize between-group variance for any given number of regions. This will maximize both the internal homogeneity of the regions and the external heterogeneity among them. It can be achieved in two ways, given data on a number of areas within a larger block of territory (e.g., counties within a state or census tracts within a city):

1. By a process of division, in which the regionalization begins with the single block and then divides it into two or more separate regions on the predefined criteria.
2. By a process of aggregation, whereby the regionalization begins with a large number of small areas, which are grouped together into a smaller number of larger areas according to the agreed criteria.

Most of the classification–regionalization algorithms commonly applied, such as those in the SPSS package, use the latter approach. They take data on the relevant characteristics for a number of areas—the building blocks for the regionalization–classification—and aggregate them into larger, relatively homogeneous, areas using set principles. All such procedures are essentially inductive: they search for the "best" groupings within the population.

Aggregation Algorithms

Most regionalization exercises do not begin with a predetermined number of regions to be defined. Instead, they proceed inductively, generating a range of regionalizations from which analysts have to choose that which best suits their purposes. Thus, whereas most algorithms will proceed directly to a required number of regions if the analyst so specifies, they will also produce a range of regions from $(n - 1)$, where n is the number of building blocks, to 1 or a specified subset within that range.

The algorithms using aggregation strategies are based on square matrices of either similarity or dissimilarity measures, in which the rows and columns are the building blocks and the cell values contain the measure of similarity/ difference between each pair. The procedure operates as follows:

1. The matrix is scanned and the two most similar (least dissimilar) building blocks according to the cell values are joined together to form a region.
2. The matrix is recalculated, with the two building blocks forming the new region removed and replaced by a joint unit. The measures of similarity–difference between this new unit and the others in the matrix are calculated and added to the matrix.
3. The within-region and between-region variation in the variables (or some other measures of internal homogeneity and external heterogeneity) are calculated.
4. The process returns to step 1, and the next grouping is produced, which will involve (i) creating a new region by grouping together two or more of the original building blocks, (ii) extending one of the already created regions by adding one of the original building blocks to one of the regions, or (iii) aggregating two of the already created regions into a larger, single region.

If the goal is to produce regions rather than regional types, a further decision is involved at the first step. If the two units to be combined are contiguous, then the grouping takes place; if they are not, then the grouping is rejected and the matrix is rescanned for the next most similar (least dissimilar) pair, which will then be grouped together if they are contiguous.

The regionalization process is usually summarized in two diagrams. The first shows which units are grouped with which: the building blocks are arranged along the horizontal axis and a measure of the "efficiency" of the grouping forms the vertical axis. The sequence of groups can then be read off the diagram, which is usually referred to as a dendrogram (or linkage tree). The second diagram shows the efficiency measure graphed against the steps in the regionalization sequence.

Preclassification Decision Making

Before any of these aggregation procedures can be applied, several decisions have to be made regarding measurement issues. The first is which measure of similarity–difference to use in the matrix that is scanned to find the most similar pair at each stage of the regionalization. The following are among a range of possibilities:

- If the regionalization is being undertaken on a single variable only measured on an interval or ratio scale, such as average July temperature, then the difference between each pair can readily be measured as the difference (or distance) between its members on that variable.

- If the regionalization is being conducted on three or more interval or ratio variables, then the difference can be measured as the sum of the differences (or distances) between the members of a pair across all variables.
- If these variables employ different metrics, then they can first be standardized to zero mean and unit variance to ensure that each has equal weight in the regionalization (other weightings can be introduced, if desired).
- If there is collinearity among some of the variables, this can be removed by replacing them with a smaller number of composite variables, produced by a principal components analysis, with the component scores forming the individual variables from which the distances in the reduced space can be calculated.
- If the variables are part of a closed number set (such as the percentage of the local workforce employed in a range of industries), then a measure of dissimilarity, such as the Gini coefficient, can be employed.
- If the variables are measured on a nominal scale only, the difference between two units can be measured by the number of variables on which they agree.

All of the these are measures of difference/dissimilarity. In addition, measures of similarity can be employed:

- If the building blocks are being compared across a range of variables, their similarity across that full profile can be assessed using relevant correlation coefficients (e.g., product–moment for ratio/interval data).

The second is how to measure the distances in similarity–dissimilarity matrices when it is not the original building blocks that are being grouped together but rather groups that have already been created (i.e., the second and third types previously outlined). Again, there are several options, including

- The nearest distance method, which represents the distance between a pair of groups (or between an individual and a group) as the shortest distance between any pair of individuals selected one from each of the two groups.
- The furthest distance method, which takes the distance between a pair of groups (or between an individual and a group) as the longest distance between any pair of individuals selected one from each of the two groups.
- The total distance method, which takes the distance between a pair of groups (or between an individual and a group) as the sum of the distances between all pairs of individuals selected from the two groups across all variables.
- The average distance method, which takes the distance between a pair of groups as the distance between their mean (or median) points, often termed the group centroid.

Each of these has problems, as discussed later.

The third issue concerns the measure of the efficiency of each stage in the regionalization process. Again, a number of options is available, including

- The size of the within-group variation, defined as the sum of the squared distances between each individual unit's value on each variable and the mean for its group on that variable: with no groups formed, this is 0.0, and as groups are formed the degree of internal heterogeneity increases.
- The ratio of the within- to between-group variation, with the within-group variation defined as previously discussed and the between-group variation calculated as the sum of the squared differences between the mean for each group on each variable and the overall mean.
- The contribution of the grouping to the error sum of squares (defined by Ward in 1963 as the within-cluster sum of squared distances).

Whichever is used, the value of the efficiency measure can be graphed against the step in the regionalization process; breaks in this graph may identify steps at which the grouping could be stopped.

A commonly employed method, available in statistical packages such as SPSS, is that devised by Ward in 1963 and almost invariably known as Ward's method. Ward argues that the goal of any regionalization should be to produce groups in which the distances between group members and the group's centroid are minimized: the group members should be clustered together as tightly as possible in the m-dimensional space (where m is the number of variables deployed in the regionalization). To achieve this, the formal goal is to minimize the variance of the within-group distances in the next group to be created, what Ward terms the error sum of squares (ESS):

$$\mathrm{ESS} = \sum_{j=1}^{m} \left\{ \frac{\left[\sum_{i=1}^{n} D_{ix}^2\right]}{n} \right\},$$

where D is the distance between building block i and the group centroid x, and n is the number of members of the group, with summation over all m variables. At each step in the regionalization process, all possible groupings are evaluated, and that with the smallest ESS is adopted.

All these decisions are subjective to an extent. Which option is selected depends on the analyst's preferences with regard to the uses planned for the regionalization–classification: There is no best option because there is no agreement on which criteria should prevail.

A Simple Illustration

A simple illustration of the procedure is provided by the following example. Seven glaciers (G_1–G_7) have

advanced at different rates during 1 year (Table I), and we want to aggregate them into groups with similar rates of advance (in meters). The diagram in Fig. 1 shows them arranged along a single axis according to the grouping variable. The first step involves calculating an interglacier dissimilarity matrix, in which the off-diagonal entries are the differences between each pair in their rate of advance. We use the shortest distance method for grouping pairs of individual glaciers and the average distance method for groupings involving already created groups. The shortest distance, 1 m, is between glaciers 4 and 5, and these are grouped together to create a new group, G_A. The within-group variation after this step is 0.5: Both G_4 and G_5 are 0.5 m from the group centroid, and the sum of these squared deviations is $(0.25 + 0.25) = 0.5$.

A new interglacier dissimilarity matrix is now produced in which G_4 and G_5 are replaced by G_A (Fig. 1) and distances between each of the remaining glaciers and the new group are computed as the distance to the centroid of G_A, which is the same as the average distance between the individual glacier and the two glaciers forming that group. There are now two pairs of individual glaciers that are the same distance (2 m) apart, so these form new groups—G_B and G_C, respectively—at the next step, as shown in the bottom of Table I. This adds an additional 2.0 units to the within-group variation, resulting in a total over both steps of 4.5 (the final column in the bottom block of data in Table I).

A further interglacier dissimilarity matrix (not shown) is now created, with G_B replacing G_1 and G_2 and G_C replacing G_6 and G_7 (Fig. 1). The next step of the grouping involves combining G_3 with G_B to form group G_D. After two more steps, all seven glaciers are in a single group—G_F.

The full sequence of steps is shown in a dendrogram (Fig. 2), in which the glaciers are arrayed along the horizontal axis and the extra contribution of each grouping to the within-group variation forms the vertical axis. The formation of the groups is shown in the body of the graph: with every step, the amount of intragroup homogeneity declines as the contribution to the within-group variation increases.

With such an inductive classification procedure, the decision at which step to end the agglomeration and use that number of groups/regions is subjective. The information in Fig. 2 may be used in this regard, along with the data on the within-group variation in the final two columns of the bottom block of Table I. These two columns of data are shown in graphical form in Fig. 3, in which both the extra contribution to the within-group variation at each step and the total variation are shown. These indicate that up to step 4, both the extra contributions and the total variation are relatively small: after this step, the graphs become much steeper. This suggests that the most appropriate stage to stop the process may be after step 4, when there are three groups—G_A–G_C: at the next step, the within-group variation increases substantially.

Table I A Simple Classification: Glacier Advance

The seven glaciers and their rate of advance

G_1	G_2	G_3	G_4	G_5	G_6	G_7
10	12	15	6	7	19	21

The interglacier distance matrix

	G_1	G_2	G_3	G_4	G_5	G_6	G_7
G_1	0	2	5	4	3	9	11
G_2	2	0	3	6	5	7	9
G_3	5	3	0	9	8	4	6
G_4	4	6	9	0	1	13	15
G_5	3	5	8	1	0	12	14
G_6	9	7	4	13	12	0	2
G_7	11	9	6	15	14	2	0

The revised interglacier distance matrix after the first step of the grouping

	G_1	G_2	G_3	G_6	G_7	G_A
G_1	0	2	5	9	11	3.5
G_2	2	0	3	7	9	5.5
G_3	5	3	0	4	6	8.5
G_6	9	7	4	0	2	12.5
G_7	11	9	6	2	0	14.5
G_A	3.5	5.5	8.5	12.5	14.5	0.0

The steps in the grouping of the glaciers

1	G_4	G_5	G_A	0.5	0.5
2 =	G_1	G_2	G_B	2.0	
2 =	G_6	G_7	G_C	2.0	4.5
4	G_B	G_3	G_D	10.67	15.17
5	G_A	G_D	G_E	40.83	56.00
6	G_E	G_C	G_F	153.45	194.28

Suboptimal Groupings and Local Optima

One problem with hierarchical agglomeration procedures is that after a region is formed at one step of the grouping process, it cannot be disaggregated at a later step. This

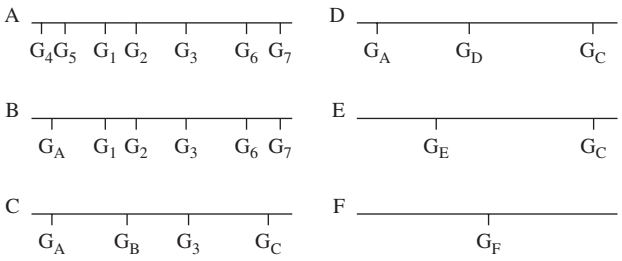


Figure 1 The seven glaciers (G_1–G_7) arranged according to their rate of advance.

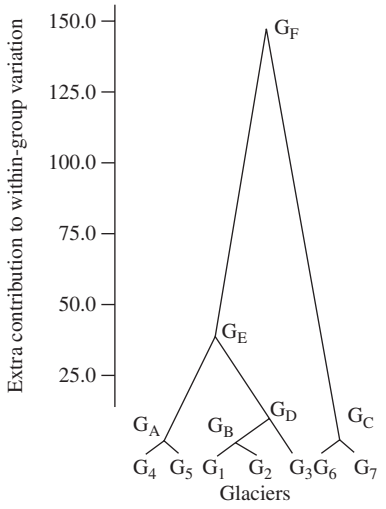


Figure 2 Dendrogram showing the classification of the seven glaciers according to the volume of within-group variation.

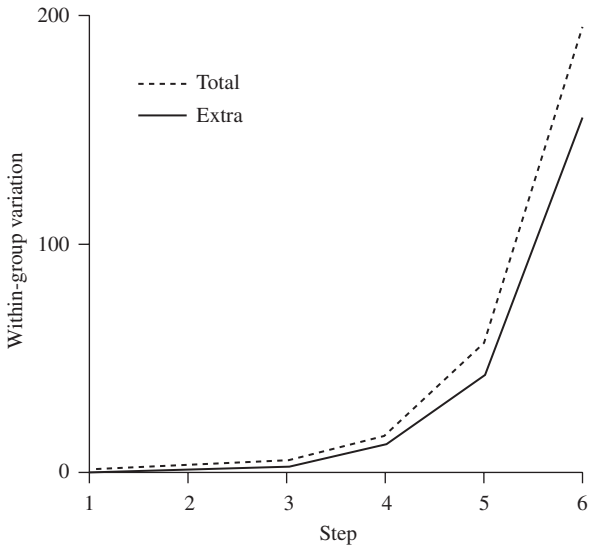


Figure 3 The efficiency of the classification of the seven glaciers showing both the extra contribution to the within-group variation and the total within-group variation at each step.

may mean that whereas the regionalization is optimal at some steps, it is not at others. This is illustrated by the seven-glacier example. At the penultimate step of the classification, there are two groups, G_C and G_E: the former contains glaciers G_6 and G_7 and the latter contains the other five. The within-group variation is 56.00, and the between-group variation is 138.28. Is this the optimal two-group regionalization? Glacier G_3 is a member of group G_E, but it is closest to a member of G_C (G_2). If G_3 were moved from the former to the latter group—making groups with four and three members, respectively—then the within-group variation would be 41.41 and the between-group variation would be 152.87. This grouping is clearly better than that originally produced: the ratio of the two variation components is 0.40 in the first case and 0.27 in the second.

How is such a suboptimal solution produced? G_3 is closest to G_2, as shown on the original dendrogram, but the latter is closest to G_1, with which it is grouped at the second step in the procedure. G_3 is now equidistant between the mean for the new group G_B, comprising glaciers G_1 and G_2, and glacier G_6. However, G_6 has already been grouped with its nearest neighbor G_7, and G_3 is farther from the mean for that group (G_C) than it is from the mean for G_B. So at the next step G_3 is combined with G_B to produce G_D.

The potential problem generated by hierarchical clustering procedures, therefore, is that when grouping previously formed groups, it cannot determine whether the group memberships remain optimal. The optimum at one step in the procedure (or local optimum) is built-in to later stages. In the example, if we had initially opted for just two groups, then glaciers G_3, G_6, and G_7 would have been placed in one group and the other four in the second group. Because we had undertaken an inductive search, however, G_3 was "misclassified" at an early step in the grouping sequence.

This can be corrected in two ways. First, when an inductive, agglomerative search has been undertaken and a desired number of regions determined, a further clustering can be performed specifying that number of regions. Alternatively, Berry suggested subjecting the identified set of regions to a discriminant analysis using the original independent variables from which the dissimilarity matrix was derived to determine if any region members were misclassified.

Ward's Method

As already noted, Ward's method is one of the most commonly used hierarchical grouping procedures because it results in cohesive (or tight) groups with low within-group variation, as measured by the ESS value used to determine the next stage in the grouping. Table II illustrates the steps in grouping the seven glaciers using this procedure.

At the first step, the smallest ESS value is for the grouping of G_4 and G_5: their mean value (centroid) is

$$[(6+7)/2] = 6.5,$$

so the ESS is

$$\left[(6-6.5)^2 + (7-6.5)^2\right]/2 = [(0.25) + (0.25)]/2 = 0.25.$$

Table II Applying Ward's Method to the Grouping of the Glaciers[a]

Step 1	*ESS*	*Step 2*	*ESS*	*Step 3*	*ESS*	*Step 4*	*ESS*	*Step 5*	*ESS*	*Step 6*	*ESS*
G_4, G_5	**0.25**	G_A, G_1	2.89	G_A, G_B	5.69	G_A, G_B	5.69	**G_A, G_B**	**5.69**	G_E, G_F	28.41
G_5, G_1	2.25	**G_1, G_2**	**1.00**	G_B, G_3	4.22	G_B, G_3	4.22	G_B, G_D	17.04		
G_1, G_2	1.00	G_2, G_3	2.25	G_3, G_6	4.00	**G_3, G_C**	**2.38**				
G_2, G_3	2.25	G_3, G_6	4.00	**G_6, G_7**	**1.00**						
G_3, G_6	4.00	G_6, G_7	1.00								
G_6, G_7	1.00										

[a] The pair grouped at each step is shown in bold. ESS, error sum of squares.

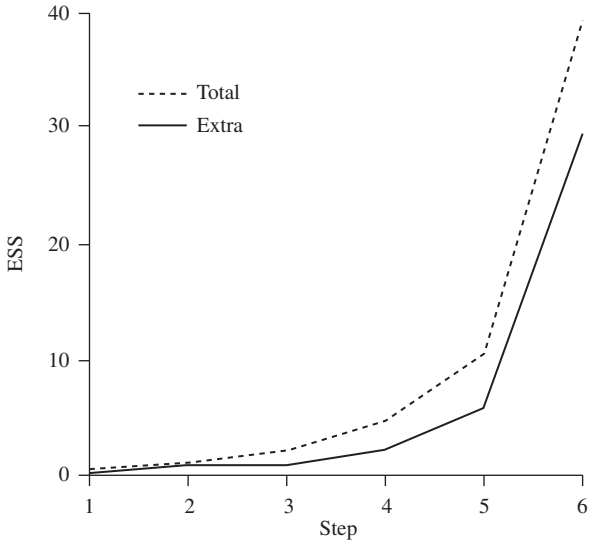


Figure 4 The efficiency of the classification of the seven glaciers using Ward's method showing both the extra contribution to the error sum of squares (ESS) and the total ESS at each step.

These are grouped to form G_A, and at the next step the ESS value is computed for a group involving G_A plus its nearest neighbor, G_1. Their centroid position is

$$[(6 + 7 + 10)/3] = 7.67,$$

so the ESS is

$$\left[(6 - 7.67)^2 + (7 - 7.67)^2 + (10 - 7.67)^2\right]/3 = 2.89.$$

The grouping procedure continues, using the lowest ESS at each step.

Graphing of the ESS and total ESS values at each step (Fig. 4) suggests the same "stopping rule application" as with the previous method (Fig. 3). However, closer inspection of the membership of the groups shows that the problem of suboptimality in step 4 is not repeated: G_3 is grouped with G_C (a combination of G_6 and G_7) rather than with G_B. This illustrates one of the main advantages of Ward's method: because it uses the potential new group centroids in calculating the ESS, rather than the distance between existing group centroids (or between individuals and existing centroids), it is less likely to produce a suboptimal grouping, although this is not invariably the case, and checks should always be made.

Classifications with Closed Number Set Data

One problem with some applications of aggregative classification is that the data form closed number sets—that is, the sum of the values for any building block across all of the variables is a set value (usually 100, if the data are in percentages, or 1.0 for proportions). Use of such data in procedures that employ measures derived from the general linear model, such as correlation coefficients, is invalid, and thus the grouping procedure cannot proceed.

This problem can be avoided, however, by using an alternative procedure that is not constrained by the requirements of the general linear model. It compares the profiles of the various building blocks across the set of variables (which are converted into z scores, with zero mean and unit variance) and groups building blocks with the most similar profiles according to a between-group variation criterion. The groups can then be characterized by their mean profiles.

Factorial Ecology

Widespread use of the variants of these procedures using component or factor analysis as an initial stage for classification and regionalization since the late 1960s led some to give the various applications the collective name of "factorial ecologies." The method was initially deployed by sociologists investigating the residential patterning of cities: it was termed factorial ecology by Sweetser in 1965 and adopted as a general term for multivariate analyses of such patterns. More generally, however, it has been applied as a general descriptive term for classifications of spatial data at a range of scales.

Flow Data and Functional Regionalizations

Classifications with flow data, to create functional regions, involve identifying areas with similar spatial interaction profiles. In the simplest form, these regions can be created by taking a matrix of flows—of the volume of goods moved between two places during a given time period, for example—and for each place identifying the other with which it has the greatest amount of contact. With a square matrix, in which the rows and columns are the places (the building blocks for the regionalization), this simply involves identifying the largest value in each row (excluding the main diagonal). This is illustrated in Table III, which shows the flows among 12 places in a small network, with the rows being the origins and the columns the destinations; their locations are shown in Fig. 5. The largest flows in each row are shown in bold. There are two clear groups of places: the first (B–D) have their main flows to A, whereas the second (F–L) have their main flows to E. This suggests two functional regions based on A and E, respectively. In addition, the main flows for A and E are with each other, suggesting that they are the main nodes in a hierarchy of centers. (Their functional regions are grouped around them in a contiguous formation, although this need not be the case.) More sophisticated examples of such procedures have been devised to produce functional regionalizations of countries used for defining metropolitan regions and reporting census data.

A variant on this approach is the creation of regions as the hinterlands around centers that are the exclusive providers of a certain function, such as education. There may be n different zones with children to be allocated to a school and m schools. The goal might then be to allocate the children to the various schools in order to minimize the total distance that they travel to school, which in effect would mean allocating each n zone to its nearest m school. There may be constraints, however, such as a capacity constraint for each school (in which case the goal would be to minimize the total distance traveled by pupils to school, without any school being allocated an excess capacity), for which linear programming solutions may be sought. Further constraints could include a maximum distance for any child to travel.

Rather than produce a regionalization based simply on the largest flow from each place, an alternative is to compare the profile of flows either to or from a place. This would involve clustering the places in a square matrix (as in Table III) either by the similarity between their rows (which would involve grouping places with similar export patterns) or by the similarity between the columns (grouping places by where their imports come from). This could use one of the agglomerative procedures discussed previously; the outcome would be a regionalization of trade blocks.

Divisive Classification Algorithms

As already noted, most inductive classification procedures employ aggregative algorithms. An alternative procedure

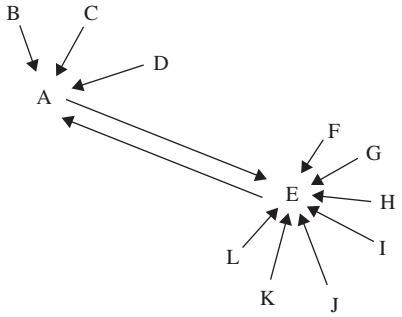


Figure 5 The simple functional regionalization showing the main links among the 12 places.

Table III Example of a Simple Functional Regionalization

From/to	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
A	—	2	3	1	**8**	3	1	2	1	0	0	1
B	**12**	—	3	2	0	1	1	1	1	1	1	1
C	**11**	3	—	3	1	2	1	0	0	0	0	1
D	**10**	2	3	—	1	0	0	0	0	0	0	0
E	**15**	1	2	1	—	2	1	3	2	1	2	3
F	1	1	2	0	**9**	—	2	1	0	0	0	1
G	2	0	0	1	**10**	2	—	2	0	0	1	0
H	0	0	0	0	**11**	1	2	—	4	2	1	0
I	0	0	0	0	**13**	0	1	2	—	4	2	1
J	0	0	0	0	**12**	0	0	1	3	—	4	1
K	2	1	1	0	**10**	2	0	0	0	2	—	3
L	1	0	0	0	**11**	3	1	0	0	1	2	—

is to use divisive methods, which are particularly useful for binary data. One such method is association analysis, developed by plant ecologists.

Take a data set comprising a number of sample sites (m) at each of which the presence or absence of a range of plant species (n) is recorded. One can then calculate the chi-square (χ^2) statistic to evaluate the similarity in the distributions of any pair of species: the greater the dissimilarity (i.e., the probability that where one species is present, the other is absent), the larger the χ^2 value. This produces a matrix of between-species χ^2 values. The species are then divided into two groups using the species that is most dissimilar from all others in its distribution as the discriminator. This is done by summing all the χ^2 values in each column of the matrix and selecting the species with the largest total. The sites are then placed into one of two classes depending on whether the discriminator species is present there or not.

The next step takes the two groups of sites separately and for each computes a matrix of χ^2 values involving the $(n-1)$ species remaining; that is, the species involved in the first step is discarded because it is either present in or absent from all the sites in the relevant group. A further discriminator species is then determined for each of the groups, and a division of each is undertaken. The sequence continues until all the groups contain a single site only: as it does so, the size of the χ^2 value used to determine the next division will decline, indicating that within-group homogeneity is decreasing and between-group heterogeneity is increasing. The divisive process can be halted at any step, and the decision making can be aided by a diagram similar to a dendrogram, with the size of the maximum χ^2 involved in the division on the vertical axis.

This procedure can be applied to the data matrix in two ways. Either it can be deployed to group the sites on the basis of the species that are present/absent there or the original data matrix can be transposed to allow a grouping of the species on the basis of the sites where they are present/absent.

Algorithms Using Artificial (or Computational) Intelligence

Recent developments in computer technology have seen the emergence of computer algorithms for classification employing neural network procedures. Many are widely used in remote sensing, whereby space imagery of environmental and other features is employed for a range of mapping and other activities: classification and regionalization procedures are routinely deployed as means of generalizing large amounts of data. The data relate to small spatial areas called pixels (small squares, as on a TV screen). Each image comprises a value for a large number of such areas in a raster (or checkerboard) format; much of the initial analysis of such data involves their classification into similar types (e.g., for land cover, pixels with mixtures of land use will appear differently from those with a single use).

These inductive procedures begin with a representative sample k of the n building blocks being selected as the "cores" for the regions, and then the genetic algorithm allocates the remaining $(n-k)$ building blocks to those cores through a learning process until homogeneous regions are created. The algorithm is rerun a large number of times with a different sample of k cores, and the best classification is then selected. Such inductive procedures are often termed unsupervised because the cores are randomly selected. An alternative—supervised—procedure involves creating "ideal type" cores as the nodes for the regions; there must therefore be a predetermined number of regions. Such cores have a predetermined population mix, which may be derived either a priori or empirically, from that or another data set. Having created the cores, the building blocks are then regionalized using neural net methods, with the nature of the cores changing during the learning process.

Some of the available procedures employ fuzzy logics (or soft classifications) in which the assumption of mutually exclusive categories deployed in almost all classifications and regionalizations using the procedures discussed previously is relaxed. Individuals can either be placed in several classes, perhaps with some measure of their degree of similarity to each class, or they can be subdivided between two or more classes. In remote sensing, this is especially valuable because the building blocks employed are areas that may have a mixture of land uses (or whatever variables are being used in the classification). As noted previously, many of the units employed in social science classifications and regionalizations are also areas that may be heterogeneous, and such fuzzy logic classification procedures are likely to become more popular in the future. (Fisher argues the value of such approaches because many of the concepts underlying geographical regionalizations, such as the distinction between urban and rural areas, are inherently fuzzy, especially given that they are identified using data that refer to heterogeneous areas.)

Geodemographics

A major application of classification–regionalization procedures during the past two decades has been in an area commonly known as geodemographics. This developed from factorial ecology. The procedures and algorithms involved were adopted by market research firms to identify areas with common socioeconomic and demographic characteristics that could be used for targeting in

various sales campaigns. The procedures were enhanced after the development of geographical information systems allowed the census information to be linked with other data sets (many of them proprietary) based on consumer and other surveys; with these, it is possible to update data sets regularly rather than relying on census data, which may be 10 or more years old. In addition, in many countries the address files associated with the postal and electoral systems are also available for purchase, which can also be used to update address files for targeting.

The types identified using such procedures classify areas according to their residents' lifestyles based on information relating to such characteristics as newspaper readership, types of TV programs watched, and frequency of purchase of various products as well as indicators derived from censuses (such as age and socioeconomic status): they are sometimes called lifestyle databases. these allow firms to target their marketing at certain types of areas, and thus customers, thereby either avoiding areas whose residents are unlikely to purchase their products or identifying areas for potential expansion of sales. In addition, such databases may be used by other bodies, such as political parties seeking to identify target groups of voters for particular policies.

Redistricting as Regionalization

A particular form of regionalization involves the creation of legislative districts for electoral purposes, such as congressional districts in the United States and parliamentary constituencies in the United Kingdom. In such regionalization, the goal is to create a given number of districts, each comprising a contiguous block of territory and within a set margin of error for its total population or electorate. The number of regions required is predetermined. Algorithms have been written to identify groupings of the component areas—the building blocks that are aggregated up to form the districts.

These algorithms operate as follows. Given n building blocks (small areas within the relevant territory being districted) to be aggregated into k districts (where $k < n$), with each district comprising a contiguous block of territory, they begin by selecting k "seeding cores" from among the building blocks using a stratified random selection procedure so that the cores are not clustered together in one portion of the map. Then, they build outward from these by aggregating neighboring areas into each district until they meet the size criterion, and the process continues until all the building blocks are associated with a district and all the districts are within the prescribed size range. Again, this proceeds using random selection procedures. The core district to have the next building block aggregated to it is selected at random, and then one of its contiguous building blocks is selected at random to join the district.

Because of the random elements built in to such procedures, if run many times the computer algorithm will almost certainly identify a number of solutions to the district-building regionalization problem. (The algorithm is also likely to fail many times because one of its identified districts is too small relative to the criterion specified but cannot be expanded because it is surrounded by other districts.) These various solutions can then be evaluated on other criteria, such as their shape, how frequently they cross the boundaries of other administrative areas (such as counties within U.S. states), and how different they are from the existing districts. Decision makers then have to choose which of the options they prefer—a choice that may be made on partisan grounds. For example, in 1973 Morrill found that using the same criteria regarding size, the Republican and Democratic parties derived very different sets of districts for the Washington State legislature. (The Republicans produced a map of districts for the State House in which they would probably win 48 seats compared to the Democrats 38, with 12 marginal; the Democrats' map gave them assured victory for 50 seats, compared to 36 for their opponents, with the remaining 12 too close to call. This, of course, is the long-practiced art of gerrymandering.)

Multicriteria districting is also possible, incorporating additional variables such as conformity with other administrative areas and previous sets of districts. One important criterion applied in areas of the United States in recent years is the racial composition of districts. Under the Voting Rights Act, the voting power of racial minorities should not be diluted by districting schemes, so states have sought to produce a number of districts with black majorities consistent with their proportion of the total population (the so-called minority/majority districts).

Even when the regionalization is a multicriterion exercise, however, a number of feasible solutions will almost certainly be found—and that number could be very large if the ratio of the building blocks (n) to the desired number of regions (k) is also large: the combinatorial possibilities of producing 10 regions using several thousand building blocks are manifold. Thus, the final choice of which regionalization to adopt has to be based on criteria (which may be entirely subjective) that are not incorporated within the algorithm.

Regionalization and the Modifiable Areal Unit Problem

As previously noted, regionalization involves issues of scale, which are crucial to a number of areas of spatial analysis. These combine to form what has become known as the modifiable areal unit problem (MAUP). In certain

types of regionalization, such as the construction of districts for electoral purposes discussed previously, there is a very large number of possible solutions to the problem, only one of which will be optimal (i.e., the best on the selected criterion or criteria), but many others fall within an "optimality range" and are thus acceptable. However, different solutions within this optimality range may produce different outcomes on one or more other criteria.

The MAUP therefore arises when there is a large number of different solutions to the same regionalization problem within predefined constraints. This can pose serious problems both for potential users of a regionalization (such as with a set of congressional districts) and for those who employ the regionalizations as the units for further analyses.

The MAUP and Other Criteria

The first of these problems can be illustrated by the redistricting issues discussed previously. In the Washington State example, the two political parties derived very different regionalizations—with very different potential election results—and the Grand Master employed by the court to produce a "politically nonpartisan" regionalization created yet another, which would have given the Republicans 38 safe seats in the State House and the Democrats 41, with 19 too close to call (recall that the two parties' plans each contained only 12 marginal districts).

In 1982, Johnston and Rossiter developed a computer algorithm that would identify all the possible parliamentary constituencies in a given city, within preset constraints regarding variation in their populations. This was applied to four English cities, and they estimated for each solution to the constituency-building problem how many seats the Labour party would win, given its percentage of the votes in each ward—the building blocks used.

Table IV Redistricting Options in Four English Cities

	City			
	Coventry	*Hull*	*Leicester*	*Sheffield*
No. of building blocks	18	21	16	27
No. of constituencies	4	3	3	6
Labour % of all votes	57	68	59	65
No. of solutions	244	100	214	15,937
No. of constituencies that would be won by Labour				
6	—	—	—	697
5	—	—	—	12,327
4	43	—	—	2913
3	193	100	160	0
2	8	0	54	0
1	0	0	0	0
0	0	0	0	0

The results of this exercise are shown in Table IV. The larger the problem, the greater the number of possible solutions. For example, in the largest city (Sheffield) there were 15,397 different ways in which the 27 wards could be combined into six contiguous constituencies (with a maximum size variation of 12%): of these, Labour (with slightly less than two-thirds of the votes overall) was likely to win five of the six seats in 77% of all the solutions and all six seats in an additional 4%. In none of the cities would Labour win a smaller percentage of the seats than of the votes in any solution, although only in Hull was it bound to win all the seats in every solution (even though it only won two-thirds of the votes). The pattern of votes for Labour across the wards there, and their relative position in the city, meant that it was impossible to create even one constituency in which Labour would not win a majority of the votes.

Cirincione *et al*. used a similar algorithm but a much larger data set. Their example was congressional redistricting in South Carolina in the 1990s, where 3259 census block groups were to be aggregated into six contiguous congressional districts. They ran the algorithm four times using different criteria (size and contiguity only, size and contiguity plus compactness, size and contiguity plus county integrity, and size and contiguity plus compactness and county integrity) and generated 2500 different solutions in each case, giving a total of 10,000 separate sets of six districts. They used these as the equivalent of a sampling distribution to assess whether the actual redistricting plan adopted in the state was likely to have been constructed on racial grounds (i.e., that the racial composition of the districts was a further criterion deployed by the political cartographers). Blacks comprised 30% of the state population in 1990, evenly distributed across the state. Under the requirements of the Voting Rights Act, blacks should have been a majority in at least one, and perhaps two, of the six districts—the so-called minority–majority districts. However, none of the 10,000 plans generated by Cirincione *et al*. contained even a single black-majority district, suggesting (as in the Hull example) that the geography of blacks across the state made it very difficult, if not impossible, to produce a district with a black majority. However, the adopted plan did contain one (a very odd-shaped District 6), and they concluded that this indicated that the redistricting must have been constructed with the production of such a district in mind, thereby violating traditional redistricting criteria in order to comply with the Voting Rights Act.

The MAUP and Statistical Analyses

The second problem relates to the use of the defined regions for statistical analyses. It was first noted by Gehlke

and Biehl in 1934 and further generalized by Yule and Kendall in 1950. Gehlke and Biehl had data on the number of juvenile delinquents and median household incomes in the 252 census tracts in the Cleveland metropolitan area: The correlation between these two variables was −0.502. However, as the tracts were aggregated into larger units (of approximately similar size), the correlation increased: with 175 units it was −0.580, with 125 units it was −0.662, with 100 units it was −0.667, and with 25 units it was −0.763. Using data on wheat and potato yields in English counties, Yule and Kendall found even greater variation according to the number of units. Using all 48 English counties, the correlation was only 0.22; with 24 groups of counties it was 0.30, with 12 groups it was 0.58, with 6 groups it was 0.76, and with 3 groups it was 0.99.

In these studies, only one example of each aggregation (e.g., into six groups of English counties) was employed; however, as shown here, in almost all cases there are many different ways in which n units can be aggregated into k regions, where k is smaller than n. Each of these regionalizations may produce a different correlation coefficient, as demonstrated by Openshaw and Taylor in 1979. They took the 99 counties of the state of Iowa, for which they had data on the percentage of the population older than 60 years of age and the percentage of the votes cast in the 1968 presidential election obtained by the Republican candidate. The correlation between these two at that scale was 0.346. With 30 zones (i.e., aggregations of contiguous counties), the average correlation was 0.33, but with a standard deviation of 0.11; as the number of zones decreased, the standard deviation increased. Even this figure concealed the extent of the variation: with 30 zones, for example, it was possible to get a correlation of 0.98 between the two variables at one extreme and −0.73 at the other. In effect, as Openshaw demonstrated in 1982, it is virtually possible to produce a regionalization of the 99 Iowa counties to produce any desired ordinary least squares relationship between the two variables—both the correlation and the slope coefficient.

One of the major conclusions from these and other findings is that analysts should use the smallest areal unit possible in their analyses in order to avoid the problems just demonstrated. In this regard, the MAUP is a particular case of the ecological inference problem. Another conclusion is that for any specific analysis, one should employ the best possible zoning scheme available. However, as noted previously, much of the data used in spatial analyses—and much other social science research—relate to population aggregates, most frequently for various administrative areas such as those created for censuses (e.g., blocks and tracts). In most analyses of spatial data, therefore, it is necessary to recognize that the patterns and relationships identified are for just one of many alternative realizations of the same aggregation of individual into areal data, and the analysts has only partial control over the aggregation process. As Openshaw (1982) notes, "It is a geographical fact of life that the results of spatial study will *always* depend on the areal units being studied" (p. 37).

The Two Components of the MAUP

There are two components to the MAUP. The aggregation problem occurs because different clusters of building blocks into the same number of regions can result in different geographies, and hence different statistical relationships. The scale problem occurs because the results obtained are also a function of the size of the regions employed—relative, that is, to the original building blocks. In many cases, there is a clear relationship between the scale of the analysis and the results obtained—for example, in the size of the correlation coefficient. However, the interaction of scale and aggregation effects can considerably blur such a general tendency.

Is the MAUP important in many statistical analyses? As Openshaw notes, its impact is fairly trivial if there is agreement on the right geographical object for any particular analysis—the right scale and aggregation. That is extremely unlikely, however, although it may occur in some cases. In the study of the personal vote for House of Representatives members, for example, the obvious unit to employ is the congressional district. Openshaw and Taylor explore three other responses:

1. It is an insoluble problem and thus can only be ignored.
2. It is a problem that can be assumed away, with the results obtained from the particular available data set being accepted as "the real ones."
3. It is a very powerful analytical device for exploring various aspects of geography and spatial variations since alternative regionalizations can be produced. This allows the creation of frequency distributions with which one regionalization can be compared (as with the example from Cirincione *et al*. cited previously) and of optimal regionalizations for particular purposes.

Most geographers and other spatial analysts have at least implicitly adopted the first of these (and perhaps even the second), in part because they lack the available data that would allow the third response to be adopted.

Given the importance of the MAUP, the issue arises as to how it can be incorporated into classification procedures such as regionalizations for census data reporting at small spatial scales. In the UK census from 1951 to 1991, for example, the smallest spatial scale for data reporting was the enumeration districts (EDs)—the small areas used for the administration of the census collection and tabulation procedures. Each ED was a separate unit for data collection. Such EDs were defined for

administrative convenience only and bore no necessary relationship to the underlying geography they were portraying. Many users—private and public sector, plus academic researchers—argued the desirability of these being defined so that they were relatively homogeneous on predetermined characteristics, within prescribed constraints, notably relating to their size (e.g., population or household minima or maxima). Using data for even smaller areas than EDs (postcode sectors), Martin *et al.* used one of Openshaw's classification algorithms (AZP) to produce a set of output areas after the census data had been collected, which met certain criteria—a target population (i.e., average population) with given population minima and homogeneity on specified other variables (such as housing tenure) plus a shape constraint. The result is a geography for publishing data from the 2001 UK census with reporting units that are much more homogeneous than heretofore.

Regionalizing Point Patterns

Most of the agglomerative clustering procedures discussed here involve data sets in which the observation units (or building blocks) are points in an n-dimensional space (such as the orthogonal components derived from a principal components analysis). In many cases, the points in multivariate space refer to the characteristics of areas. The goal is to classify them into groups based on their similarity, as measured by their distance apart in the n-dimensional space.

In these approaches, the goal is to produce the optimal set of regions within certain constraints (or a set of regions within an envelope of constraints). A related approach involves searching for regions that meet other criteria and may indeed overlap. This can be illustrated by some of Openshaw's work on the search for clusters of disease outbreaks. For example, there has been considerable debate in the United Kingdom regarding whether leukemia in children can be associated with proximity to a nuclear power plant or similar installation. Openshaw designed inductive methods, geographical analysis machines (GAMs), to answer this question. The rationale was based on the MAUP: analysis of one particular realization of the geography of a disease may not reveal the clusters, whereas analysis of other realizations (i.e., different regionalizations) may reveal them.

Openshaw's "machines" operated by taking the point pattern of the events whose geography was being investigated, such as leukemia cases. These were integrated in a geographical information system with other data sets (e.g., population data for small areas from censuses) in order to compute rates of occurrence for the disease. The algorithm then selected a random set of points within the relevant map (using grid coordinates) and counted the number of occurrences of the disease within a given distance of that point. The rate of occurrence was then calculated for that region, and its statistical significance was assessed. Repeated a large number of times, perhaps with varying regional sizes (i.e., distances from the randomly selected point), the output was a map of those areas with significant clusters (i.e., unexpectedly high rates). Where there was a large number of significant clusters (i.e., where a number of different random starting points had found such clusters), he argued that this provides a prima facie case that there may well be a localized cause for such outbreaks worthy of investigation.

Openshaw's methods of producing such *ad hoc* regionalizations through inductive, exploratory data analyses have been extended to a range of methods in what geographers call local statistics. Rather than assume that a relationship is constant over an entire map, these methods search for local variations in that relationship, in effect defining regions within regions. Openshaw extended the GAMs to a geographical explanations machine, which identified clusters of rare events and then correlated them with other variables that offered potential explanations for those clusters. Others have extended the methodology to identify regions with similar regression relationships between variables.

Acknowledgments

The author is grateful to Brian Berry, Danny Dorling, Giles Foody, Peter Haggett, Kelvyn Jones, David Martin, and Paul Mather for very valuable comments on the manuscript.

See Also the Following Articles

Aggregation • Demography • Mathematical Demography • Spatial Externalities

Further Reading

Cirincione, C., Darling, T. A., and O'Rourke, T. G. (2000). Assessing South Carolina's congressional districting. *Polit. Geogr.* **19,** 189–212.

Elliott, P., Cuzick, J., English, D., and Stern, R. (eds.) (1996). *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford University Press, Oxford, UK.

Elliott, P., Wakefield, J., Best, N., and Briggs, D. (eds.) (2001). *Spatial Epidemiology: Methods and Applications*. Oxford University Press, Oxford, UK.

Farr, M., and Baxter, R. S. (2001). MOSAIC: From an area classification system to individual classification. *J. Targeting Measurement Anal. Marketing* **10,** 55–65.

Fisher, P. (2001). Sorites paradox and vague geographies. *Fuzzy Sets Systems* **113,** 7–18.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis.* Sage, London.

Gordon, A. D. (1999). *Classification,* 2nd Ed. Chapman & Hall, London.

Johnston, R. J., Pattie, C. J., Dorling, D. F. L., and Rossiter, D. J. (2001). *From Votes to Seats: The Operation of the UK's Electoral System since 1950.* Manchester University Press, Manchester, UK.

King, G. (1997). *A Solution to the Ecological Inference Problem.* Princeton University Press, Princeton, NJ.

Macmillan, W. I. (2001). Redistricting in a GIS environment: An optimization algorithm using switching-points. *J. Geographical Systems* **3,** 167–180.

Macmillan, W. I., and Pierce, T. (1996). Active computer-assisted redistricting. In *Fixing the Boundaries: Defining and Redefining Single Member Electoral Districts* (I. McLean and D. Butler, eds.), pp. 219–234. Dartmouth, Aldershot, UK.

Martin, D., Nolan, A., and Tranmer, M. (2001). The application of zone-design methodology in the 2001 UK census. *Environ. Planning A* **33,** 1949–1962.

Mitchell, R., Martin, D., and Foody, G. M. (1998). Unmixing aggregate data: Estimating the social composition of enumeration districts. *Environ. Planning A* **30,** 1929–1942.

Monmonier, M. S. (2000). *Bushmanders and Bullwinkles: How Politicians Manipulate Electronic Maps and Census Data to Win Elections.* University of Chicago Press, Chicago.

Murray, A. T. (1999). Spatial analysis using clustering methods: Evaluating central point and median approaches. *J. Geographical Systems* **1,** 367–383.

Openshaw, S. (1982). *CATMOG 38: The Modifiable Areal Unit Problem.* GeoBooks, Norwich, UK.

Openshaw, S. (1998). Building automated geographical analysis and explanation machines. In *Geocomputation: A Primer* (P. Longley, S. Brooks, R. MacDonnell, and B. Macmillan, eds.), pp. 95–116. Wiley, Chichester, UK.

Openshaw, S., and Openshaw, C. (1997). *Artificial Intelligence in Geography.* Wiley, Chichester, UK.

Tso, B., and Mather, P. M. (2001). *Classification Methods for Remotely Sensed Data.* Taylor & Francis, London.

Wakefield, J., Quinn, M., and Raab, G. (eds.) (2001). Disease clusters and ecological studies [Special issue]. *J. R. Statistical Soc. Ser. A* **164**(1).

Reliability

Duane F. Alwin
Pennsylvania State University, University Park, Pennsylvania, USA

Glossary

congeneric measures Univocal measures whose true scores are perfectly correlated (i.e., linearly related).
internal consistency A general approach to estimating the reliability of a composite score based on a set of univocal measures.
measurement error The difference between an observed variable and the true score it measures.
multitrait–multimethod approach A measurement design intended to partition reliability into two orthogonal parts due to true score trait variance and true score method variance.
parallel measures Univocal measures with tau-equivalent true scores and equal error variances.
propensity distribution A density function of a hypothetical response distribution for a fixed person.
quasi-simplex model A model for the estimation of reliability in longitudinal data.
reliability The proportion of observed score variance due to the latent true score variable.
tau-equivalent measures Univocal measures with tau-equivalent true scores and unequal error variances.
true score The expected value of the hypothetical propensity distribution for a fixed person.
univocity The property of a measure indicating that it measures one and only one underlying variable; measures having this property are said to be univocal.

Reliability of measurement encompasses the design strategies and statistical estimation methods used for assessing the relative consistency of measurement in specified populations using maximally similar efforts to measure the same quantity or attribute.

Introduction

Issues of measurement quality are among the most critical in scientific research because the analysis and interpretation of empirical results depend intimately on the ability to accurately and consistently measure the phenomena of interest. This may be more difficult in social and behavioral sciences, in which the targets of measurement are often not well specified; even when they are, the variables of interest are often impossible to observe directly. For example, concepts such as social status, personality, intelligence, attitudes, values, psychological or emotional states, deviance, or functional status may be difficult to measure precisely because they reflect difficult to define variables and are not directly observable. Even social indicators that are more often thought to directly assess concepts of interest (e.g., education level or race) are not free of conceptual specification errors that lead to imprecision. The inability to define concepts precisely in a conceptually valid way produces errors of measurement, but measurement problems are also critically related to the nature of the communication and cognitive processes involved in gathering data.

Sometimes, the term reliability is used very generally to refer to the overall stability or dependability of research results, including the absence of population specification errors, sampling error, nonresponse bias, as well as various forms of measurement errors. Here, the term is used in its more narrow psychometric meaning, focusing specifically on the absence of measurement errors. Even then, there are at least two different conceptions of error—random and nonrandom (or systematic) errors of measurement—that have consequences for research findings. Within the psychometric tradition, the concept of reliability refers to the absence of random error. This conceptualization of error may be far too narrow for many research purposes, where reliability is better understood as the more general absence of measurement error. However, it is possible to address the question of reliability separately from the more general issue of measurement error, and later the relationship between

random and nonrandom components of error is discussed.

Errors of measurement occur in virtually all measurement, regardless of content, and the factors contributing to differences in unreliability of measurement are worthy of scrutiny. It is well-known that statistical analyses ignoring unreliability of measures generally provide biased estimates of the magnitude and statistical significance of the tests of mean differences and associations among variables. Although the resulting biases tend to underestimate mean differences and the strength of relationships making tests of hypotheses more conservative, they also increase the probability of type II errors and the consequent rejection of correct, scientifically valuable hypotheses about the effects of variables of interest.

This article discusses the major approaches to estimating measurement reliability. There are two traditions for assessing reliability: (i) the classical test theory or psychometric tradition for continuous latent variables and (ii) the recent approach developed for categoric latent variables. From the point of view of either tradition, reliability estimation requires repeated measures across multiple levels of the variable. This article focuses mainly on how repeated measures are used in social research to estimate the reliability of measurement for continuous latent variables.

Key Notation and Symbols

The key notation and symbols used in this article are summarized in Table 1. This article follows the convention of using uppercase symbols to denote random variables and vectors of random variables and uppercase Greek symbols to represent population matrices relating random variables. Lowercase symbols are used to denote person-level scores and within-person parameters of propensity distributions.

Basic Concepts

Reliability for Continuous Variables

On the simplest level, the concept of reliability is founded on the idea of consistency of measurement. Consider a hypothetical thought experiment in which a measure of some quantity of interest (Y) is observed: It could be a child's height, the pressure in a bicycle tire, or a question inquiring about family income in a household survey. Then imagine repeating the experiment, taking a second measure of Y, under the assumption that nothing has changed—that is, neither the measurement device nor the quantity being measured have changed. If across these two replications one obtains consistent results, we say the measure of Y is reliable, and if the results are inconsistent, we say the measure is unreliable. Of course, reliability is not a categoric variable, and ultimately we seek to quantify the degree of consistency or reliability in social measurement.

Classical true score theory (CTST) provides a theoretical model for formalizing the statement of this basic idea and ultimately for the estimation and

Table I Key Symbols Used in Discussion of Reliability

Symbol	*Definition*
Y_{gp}	An *observed score* for variable Y_g on person p
τ_{gp}	The *true score* of person p in measure Y_g defined as $E(y_{gp})$
ε_{pg}	The *error score* for person p in measure Y_g defined as $\varepsilon_{pg} = y_{gp} - \tau_{gp}$
S	A finite population of persons
G	The number of measures in a set of univocal measures
$\mathrm{E}[Y_g]$	The expectation of the observed score random variable Y_g in population S
$\mathrm{E}[T_g]$	The expectation of the true score random variable T_g in population S
$\mathrm{E}[E_g]$	The expectation of the error score random variable E_g in population S
$\mathrm{VAR}[Y_g]$	The variance of the observed score random variable Y_g in population S
$\mathrm{VAR}[T_g]$	The variance of the true score random variable T_g in population S
$\mathrm{VAR}[E_g]$	The variance of the error score random variable E_g in population S
$\mathrm{COV}[T_g, Y_g]$	The covariance of the random variables T_g and Y_g in population S
$\mathrm{COR}[T_g, Y_g]$	The correlation of the random variables T_g and Y_g in population S
K	The number of sets of univocal measures
Σ_{YY}	Covariance matrix for a set of G measures in population S
Λ	The $(G \times K)$ matrix of regression coefficients relating observed measures to true scores in population S
Φ	The $(K \times K)$ matrix of covariances among latent true scores in population S
Θ^2	The $(G \times G)$ matrix of covariances among errors of measurement in population S
M	The number of distinct methods of measurement
P	The number of occasions of measurement

quantification of reliability of measurement. The classical definitions of observed score, true score, and measurement error are reviewed, as well as several results that follow from these definitions, including the definition of reliability. First covered are definitions of these scores for a fixed person, p, a member of the population (S) for which we seek to estimate the reliability of measurement of the random variable Y. Reference to these elements as persons is entirely arbitrary because they may be organizations, work groups, families, counties, or any other any theoretically relevant unit of observation. The reference to "persons" is used because the classical theory of reliability was developed for scores defined for persons and because the application of the theory has been primarily in studies of people. It is important to note that throughout this article the assumption is made that there exists a finite population of persons (S) for whom the CTST model applies and that we wish to draw inferences about the extent of measurement error in that population.

The model assumes that Y is a univocal measure of the continuous latent random variable T, and that there is a set of multiple measures of the random variable $\{Y_1, Y_2, \ldots, Y_g, \ldots, Y_G\}$ that have the univocal property; that is, each measures only one thing, in this case T. An *observed score* y_{gp} for a fixed person p on measure g is defined as a (within-person) random variable for which a range of values for person p can be observed. In the thought experiment performed previously, imagine a hypothetical infinite repetition of measurements creating a propensity distribution for person p relating a probability density function to possible values of Y_g. The true score τ_{gp} for person p on measure g is defined as the expected value of the observed score y_{gp}, where y_{gp} is sampled from the hypothetical propensity distribution of measure Y_g for person p. Figure 1 presents several examples of what such propensity distributions might look like for a Y_g measured on a continuous scale. From this we define measurement error for a given observation as the difference between the true score and the particular score observed for p on Y_g—that is, $\varepsilon_{gp} = y_{gp} - \tau_{gp}$. Note that a different error score would result had we sampled a different y_{gp} from the propensity distribution for person p, and an infinite set of replications will produce a distribution for ε_{gp}.

Several useful results follow from these simple definitions. First, the expected error score is zero—that is, $E(\varepsilon_{gp}) = 0$. Second, the correlation between the true score and the error score for a fixed person is zero—that is, $E(\varepsilon_{gp}, \tau_{gp}) = 0$. These two results follow from the fact that the true score for person p is a fixed constant. Third, the shapes of the probability distributions of ε_{gp} and y_{gp} are identical and the variance of the propensity distribution for y_{gp} is equal to the variance of the

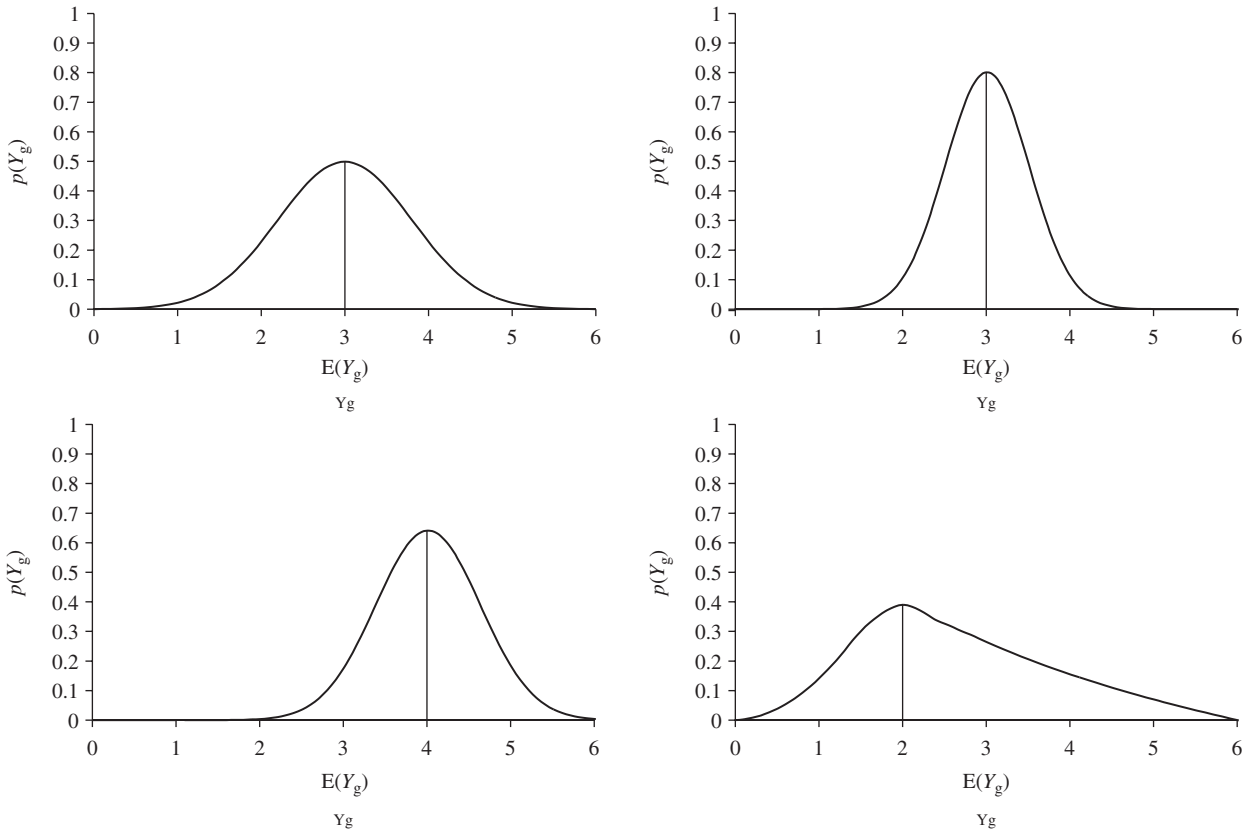


Figure 1 Examples of propensity distribution for a fixed person.

error scores—that is, $\mathrm{VAR}(\varepsilon_{gp}) = \mathrm{VAR}(y_{gp})$. These properties combine to define measurement error under this model as random error.

Given a population of persons for whom the previous model holds, we can write the model $Y_g = T_g + E_g$ for the gth measure of Y, the properties of which are well-known. In this model, the expectation of E_g is zero, from which it follows that $\mathrm{E}[Y_g] = \mathrm{E}[T_g]$. Also under this model, the covariance of the true and error scores is zero, $\mathrm{COV}[T_g, E_g] = 0$, from which it follows that the variance of the observed score equals the sum of the variance of the true score and the variance of the error score, $\mathrm{VAR}[Y_g] = \mathrm{VAR}[T_g] + \mathrm{VAR}[E_g]$. Reliability is defined as a population parameter, namely the proportion of the observed variance that is accounted for by true score variance, which is expressed as the squared correlation between Y_g and T_g: $\mathrm{COR}[Y_g, T_g]^2 = \mathrm{VAR}[T_g]/\mathrm{VAR}[Y_g] = (\mathrm{VAR}[Y_g] - \mathrm{VAR}[E_g])/\mathrm{VAR}[Y_g]$. As a generic concept, then, reliability refers to the relative proportion of random error versus true variance in the measurement of Y_g in a fixed population—that is, variance due to random "noise" versus variance due to "signal," to use a metaphor from telegraphy. As the proportion of error variance in $\mathrm{VAR}[Y_g]$ declines, reliability will approach unity, and as it increases relative to $\mathrm{VAR}[Y_g]$ reliability will approach zero.

Let Y_1 and Y_2 be two measures from the set of measures defined previously, such that $Y_1 = T_1 + E_1$ and $Y_2 = T_2 + E_2$. Assume further that Y_1 and Y_2 are tau equivalent—that is, they have the same true scores, $T = T_1 = T_2$. It follows from this set of definitions that the covariance between Y_1 and Y_2 is equal to the variance of T, $\mathrm{COV}(Y_1, Y_2) = \mathrm{VAR}(T)$. With this result we can define the reliability for the two measures of the random variable of interest, Y_1 and Y_2, in the population of interest as $\mathrm{COV}(Y_1, Y_2)/\mathrm{VAR}[Y_1]$ and $\mathrm{COV}[Y_1, Y_2]/\mathrm{VAR}[Y_2]$, respectively. Such measures are referred to as tau-equivalent measures. If, in addition to tau equivalence, the error variances of the two measures are equal (i.e., $\mathrm{VAR}[E_1] = \mathrm{VAR}[E_2]$), this would imply equal variances, $\mathrm{VAR}[Y_1]$ and $\mathrm{VAR}[Y_2]$, and equal reliabilities for Y_1 and Y_2. In this special case, the reliability of Y_1 and Y_2 can be expressed by their correlation, $\mathrm{COR}[Y_1, Y_2]$, since $\mathrm{COR}[Y_1, Y_2] = \mathrm{COV}[Y_1, Y_2]/(\mathrm{VAR}[Y_1]^{1/2}\ \mathrm{VAR}[Y_2]^{1/2})$. Such measures (with tau equivalence and equal error variances) are said to be parallel measures. Finally, measures are often not tau equivalent, as in the case of different scales or metrics used to measure Y_1 and Y_2, but their true scores are linearly related (i.e., $\mathrm{COR}[T_1, T_2] = 1.0$). When this is the case, the measures are said to be congeneric. Note the nested nature of the relationship between these three models: the tau-equivalent measures model is a special case of the congeneric model, and the parallel measures model is a special case of the tau-equivalence model.

Nonrandom Measurement Error

Usually, we think of measurement error as being more complex than the random error model developed previously. In addition to random errors of measurement, there is also the possibility that Y_{gp} contains systematic (or correlated) errors. The relationship between random and systematic errors can be clarified if we consider the following extension of the classical true score model: $y_{gp} = \tau^*_{gp} + \eta_{gp} + \varepsilon_{gp}$, where η_{gp} is a source of systematic error in the observed score; τ^*_{gp} is the true value, uncontaminated by systematic error; and ε_{gp} is the random error component discussed previously. This model directly relates to the one given previously in that $\tau_{gp} = \tau^*_{gp} + \eta_{gp}$. The idea, then, is that the variable portion of measurement error contains two types of components—a random component, ε_{gp}, and a nonrandom or systematic component, η_{gp}. Within the framework of this model, the goal would be to partition the variance in Y_g into those portions due to τ^*, η, and ε. It is frequently the case that systematic sources of error increase reliability. This is, of course, a major threat to the usefulness of CTST in assessing the quality of measurement. It is important to address the question of systematic measurement errors, but this often requires a more complicated measurement design. This can be implemented using a multitrait–multimethod measurement design along with confirmatory factor analysis, a topic to which we return later.

Common Factor Model Representation of CTST

It is a straightforward exercise to express the basic elements of CTST as a special case of the metric (unstandardized) form of the common factor model and to generalize this model to the specification of K sets of congeneric measures. Consider the following common factor model:

$$\mathbf{Y} = \Lambda\mathbf{T} + \mathbf{E},$$

where $\mathbf{Y}$ is a $(G \times 1)$ vector of observed random variables, $\mathbf{T}$ is a $(K \times 1)$ vector of true score random variables measured by the observed variables, $\mathbf{E}$ is a $(G \times 1)$ vector of error scores, and Λ is a $(G \times K)$ matrix of regression coefficients relating true and observed random variables. The covariance matrix among measures under this model can be represented as follows:

$$\Sigma_{YY} = \Lambda\Phi\Lambda' + \Theta^2,$$

where Σ_{YY}, Φ, and Θ^2 are covariance matrices for the $\mathbf{Y}$, $\mathbf{T}$, and $\mathbf{E}$ vectors defined previously, and Λ is the coefficient matrix as defined previously.

For purposes of this illustration, we consider all variables to be centered about their means. Here, we take the simplest case where $K = 1$; that is, all G variables in Y are measures of T. However, note that the model can

be written for the general case of multiple sets of congeneric measures. In the current case, the model can be represented as follows:

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \bullet \\ \bullet \\ \mathbf{Y}_G \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_{1T} \\ \boldsymbol{\lambda}_{2T} \\ \bullet \\ \bullet \\ \boldsymbol{\lambda}_{GT} \end{bmatrix} \times [\mathbf{T}] + \begin{bmatrix} \boldsymbol{E}_1 \\ \boldsymbol{E}_2 \\ \bullet \\ \bullet \\ \boldsymbol{E}_G \end{bmatrix}.$$

For G congeneric measures of T, there are $2G$ unknown parameters in this model (G λ_{gT} coefficients and G error variances, θ_g^2) with degrees of freedom (df) equal to $0.5G(G+1) - 2G$. In general, this model requires a scale to be fixed for T since it is an unobserved latent random variable. Two options exist for doing this: (i) The diagonal element in Φ can be fixed at some arbitrary value (e.g., 1.0), or (ii) one of the λ_{gT} coefficients can be fixed at some arbitrary value (e.g., 1.0). For G measures of T, the tau-equivalent model has $G+1$ parameters with $\mathrm{df} = 0.5G(G+1) - G + 1$. For a set of G parallel measures there are two parameters to be estimated: VAR[T] and VAR[E], with $\mathrm{df} = 0.5G(G+1) - 2$. Note that both the tau-equivalent measures and parallel measures form of this model invoke the assumption of tau equivalence. This is imposed on the model by fixing all λ_{gT} coefficients to unity. In order to identify the tau-equivalent or parallel measures model, observations on two measures, Y_1 and Y_2, are sufficient to identify the model. For the congeneric model, G must be ≥ 3. It should be clear that the congeneric measures model is the most general and least restrictive of these models, and the tau-equivalent and parallel measures models simply involve restrictions on this model. An internal consistency estimate of the reliability for a set of G measures for which a common factor model holds is developed later, and it is shown that Cronbach's alpha is a special case of the common factor model involving tau-equivalent measures.

What has been stated for the model in the previous discussion can be generalized to any number of G measures of any number of K factors. The only constraint is that the assumptions of the model—univocity and random measurement error—are realistic for the measures and the population from which the data derive. Although there is no way of testing whether the model is correct, when the model is overidentified the fit of the model can be evaluated using standard likelihood ratio approaches to hypothesis testing within the confirmatory factor analysis framework. For example, there is a straightforward test for whether a single factor can account for the covariances among the G measures. Absent such confirming evidence, it is unlikely that a simple true score model is appropriate.

Scaling of Variables

The discussion to this point has assumed interval-level measurement of continuous latent variables and the use of standard Pearson-based covariance approaches to the definition of statistical associations. Observed variables measured on ordinal scales are not continuous (i.e., they do not have origins or units of measurement), and therefore should not be treated as if they are continuous. This does not mean that where the observed variables are categorical the underlying latent variable being measured cannot be assumed to be continuous. Indeed, the tetrachoric and polychoric approaches to ordered dichotomous and ordinal polytomous data assume there is an underlying continuous variable Y^*, corresponding to the observed variable Y, that is normally distributed. The use of these approaches is somewhat cumbersome and labor-intensive because it requires that the estimation of the model be done in two steps. First, one estimates the polychoric or tetrachoric correlations for the observed data, which often takes a substantial amount of time, especially when the number of categories is large. In the second step, one estimates the parameters of the CTST model using maximum likelihood or weighted least squares. There are two basic strategies for estimating polychoric/tetrachoric correlations. One is to estimate the polychoric/tetrachoric correlations and thresholds jointly from all the univariate and bivariate proportions in a multiwave contingency table. This approach is computationally very complex and not generally recommended. The other, which almost always produces identical results, is to estimate the thresholds from the univariate marginals and the polychoric correlations from the bivariate marginals. In 1994, Jöreskog presented a procedure for estimating the asymptotic covariance matrix of polychoric correlations, which requires the thresholds to be equal.

Another approach to examining measurement errors, which also assumes continuous latent variables, appropriate for categoric data, is based on item response theory (IRT). Test psychologists and others have used IRT models to describe item characteristic curves in a battery of items. One specific form of the IRT model, namely Rasch models, has been suggested as one approach to modeling measurement errors. Duncan's work illustrates how the Rasch model can be applied to dichotomous variables measured on the same occasion. These approaches have not explicitly included parameters for describing reliability as defined here.

Designs for Reliability Estimation

Two general design strategies exist for estimating the reliability of measurement using repeated measures: replicate (or similar) measures during the same occasion

of measurement (cross-sectional measurement) or replicate measures in reinterview designs (longitudinal measurement). The application of either design strategy is problematic, and in some cases the estimation procedures require assumptions that are inappropriate given the data gathered in such designs. Estimating reliability from information collected within the same interview is especially difficult, owing to the virtual impossibility of replicating questions exactly. Researchers often employ similar, although not identical, questions and then examine correlation or covariance properties of the data collected. In other words, rather than multiple or repeated measures, investigators often use multiple indicators as a substitute. However, it is risky to use covariance information from multiple indicators to estimate item reliability because items that are different contain specific components of variance, orthogonal to the quantity measured in common, and because difficulties in separating reliable components of specific variance from random error variance present significant obstacles to this estimation approach. Because of potentially biasing effects of memory and other cognitive processes, it is virtually impossible to obtain unbiased estimates of reliability from cross-sectional surveys.

A second approach to estimating reliability in survey data uses the multiple-wave reinterview (i.e., test–retest) or panel design. Such longitudinal designs also have problems for the purpose of estimating reliability. For example, the test–retest approach using a single reinterview must assume that there is no change in the underlying quantity being measured. With two waves of a panel study, the assumption of no change, or even perfect correlational stability, is unrealistic, and without this assumption little purchase can be made on the question of reliability in designs involving two waves. The analysis of panel data must be able to cope with the fact that people change over time so that models for estimating reliability must take the potential for individual-level change into account. Given these requirements, techniques have been developed for estimating measurement reliability in panel designs where $p \geq 3$ in which change in the latent true score is incorporated into the model. With this approach, there is no need to rely on multiple measures within a particular wave or cross section in order to estimate the measurement reliability. This approach is possible using modern structural equation models for longitudinal data.

Reliability Models for Composite Scores

The most common approach to assessing reliability in cross-sectional data is through the use of multiple indicators of a given concept and the estimation of the reliability of a linear composite score made up of those measures. Let Y symbolize such a linear composite defined as the sum $Y_1 + Y_2 + \cdots + Y_g + \cdots + Y_G$; that is, $\sum_g (Yg)$, where g is an index that runs from 1 to G. Such estimates of reliability are referred to as internal consistency estimates of reliability (ICR). In this case, we can formulate a reliability model for the composite as $Y = T + E$, where T is a composite of true scores for the G measures, and E is a composite of error scores. This assumes that for each measure the random error model holds, that is, $Y_g = T_g + E_g$, and thus $T = \sum_g (Tg)$ and $E = \sum_g (Eg)$. The goal of the internal consistency approach is to obtain an estimate of $\mathrm{VAR}(T)/\mathrm{VAR}(Y) = [\mathrm{VAR}(Y) - \mathrm{VAR}(E)]/\mathrm{VAR}(Y)$. This can be defined as a straightforward extension of the common factor model of CTST given previously. The following identities result from the previous development:

$$\mathrm{VAR}(Y) = \sum_j \sum_i \Sigma_{YY}$$
$$\mathrm{VAR}(T) = \sum_j \sum_i [\Lambda \Phi \Lambda']$$
$$\mathrm{VAR}(E) = \sum_j \sum_i \Theta^2.$$

(Note that i and j represent indexes that run over the rows and columns of these matrices, where $i = 1$ to G and $j = 1$ to G.) In other words, the common factor representation of the CTST model given previously for the population basically partitions the composite observed score variance into true score and error variance. These quantities can be manipulated to form an internal consistency measure of composite reliability as follows:

$$\mathrm{ICR} = \frac{\sum_j \sum_i \Sigma_{YY} - \sum_j \sum_i \Theta^2}{\sum_j \sum_i \Sigma_{YY}}.$$

The most common estimate of internal consistency reliability is Cronbach's α, computed as follows:

$$\boldsymbol{\alpha} = \frac{\boldsymbol{G}}{\boldsymbol{G} - \mathbf{1}} \left[\mathbf{1} - \frac{\sum_g \mathbf{VAR}(\boldsymbol{Y}_g)}{\mathbf{VAR}(\boldsymbol{Y})} \right].$$

This formula is derived from the assumption of G unit-weighted (or equally weighted) tau-equivalent measures. The logic of the formula can be seen as follows. First, rewrite $\sum_j \sum_i \Theta^2$ in the previous expression for ICR as equal to $\sum_j \sum_i \Sigma_Y - \sum_j \sum_i \Sigma_T$, where Σ_Y is a diagonal matrix formed from the diagonal elements of Σ_{YY}, and Σ_T is a diagonal matrix formed from the diagonal of $\Lambda \Phi \Lambda'$. Note further that under tau-equivalence $\Lambda = 1$ (a vector of 1s), so this reduces to φI, where φ is the variance of T_g, and I is

a $(G \times G)$ identity matrix. Note that in the population model for tau-equivalent measures, all the elements in Σ_{YY} are identical and equal to φ, the variance of the true score of T_g. From these definitions, we can rewrite ICR as follows:

$$\text{ICR} = \frac{\sum_j \sum_i \Sigma_{YY} - \sum_j \sum_i \Sigma_Y + \sum_j \sum_i \varphi I}{\sum_j \sum_i \Sigma_{YY}}.$$

Note further that $\sum_j \sum_i \Sigma_{YY} - \sum_j \sum_i \Sigma_Y = G(G-1)\varphi$, and $\sum_j \sum_i \varphi I = G\varphi$, and thus Cronbach's α can be derived from the following identities:

$$\begin{aligned}
\text{ICR} &= [G(G-1)\varphi + G\varphi] / \sum_j \sum_i \Sigma_{YY} \\
&= [G/G-1][G(G-1)]\varphi / \sum_j \sum_i \Sigma_{YY} \\
&= [G/G-1]\left[\sum_j \sum_i \Sigma_{YY} - \sum_j \sum_i \Sigma_Y\right] \Big/ \sum_j \sum_i \Sigma_{YY} \\
&= [G/G-1]\left[1 - \left[\sum_j \sum_i \Sigma_Y / \sum_j \sum_i \Sigma_{YY}\right]\right].
\end{aligned}$$

The final identity is equivalent to the formula for Cronbach's α given previously. The point of this derivation is that the ICR approach actually has a more general formulation (the congeneric measures model) for which Cronbach's α is but a special case (i.e., ICR $= \alpha$ when the G measures are tau equivalent).

These methods can be generalized to the case of weighted composites, where Yw is the composite formed from the application of a set of weights to the G variables in Y. However, we will not consider this case here, except to note that when the vector of weights, w, is chosen to be proportional to $\Theta^{-2}\Lambda$, such a set of weights will be optimal for maximizing ICR.

There have been other variations to formulating ICR. Heise and Bohrnstedt, for example, defined an ICR coefficient, named Ω, based on the use of U^2 in place of Θ^2 in the previous formulation for ICR, where U^2 is a diagonal matrix of unique variances from an orthogonal common factor analysis of a set of G variables without the CTST assumptions of univocity (e.g., $K > 1$). They proposed partitioning Ω into its contributions from the common factors of the model, arbitrarily labeling the first factor common variance as "valid" variance and successive factor common variance as "invalid" variance.

Although it is a very popular approach, ICR coefficients have several major shortcomings. First, ICR is an unbiased estimate of composite reliability only when the true score model assumptions hold. To the extent the model assumptions are violated, it is generally believed that ICR approaches provide a lower bound estimate of reliability. However, at the same time, there is every possibility that ICR is inflated due to correlated errors (e.g., common method variance among the items), and that some reliable variance is really invalid in the sense that it represents something about responses other than true score variation, such as nonrandom sources of measurement error. ICR therefore captures systematic sources of measurement error in addition to true score variation and in this sense cannot be unambiguously interpreted as a measure of data quality.

Reliability Models for Single Measures

There are two questionable assumptions of the CTST approach. The first is that the measures are univocal measures of a single underlying variable. The second is that the errors of the measures are independent of one another. These assumptions rule out, for example, the operation of memory in the organization of responses to items in a series of questions. Obviously, respondents are fully cognizant of the answers they have given to previous questions, so there is the possibility that memory operates to distort the degree of consistency in responses. These assumptions for reliability estimation in cross-sectional studies also rule out the operation of other types of correlated errors, such as the operation of systematic method factors. Thus, in cross-sectional data it may be impossible to assume that measurement errors are independent, since similar questions are often given in sequence or at least included in the same battery. Given the shortcomings of the ICR approaches, attention has turned to more closely examining the sources of variation at the item level. Two variants on the basic CTST model have been developed for this purpose: the multitrait–multimethod measurement design and the quasi-simplex approach for longitudinal data.

Multitrait–Multimethod Models

There is an increasing amount of support for the view that shared method variance inflates ICR estimates. One approach to dealing with this is to reformulate the CTST along the lines of a multiple-factor approach and to include sources of systematic variation from both true variables and method factors. With multiple measures of the same concept, as well as different concepts measured by the same method, it is possible to formulate a multitrait–multimethod model. In general, the measurement of K traits measured by each of Q methods (generating $G = KQ$ observed variables) allows the specification of such a model.

Following from the discussion of nonrandom measurement errors, we can formulate an extension of the common factor representation of the CTST given previously as follows:

$$\mathbf{Y} = \Lambda_{T^*}\mathbf{T}^* + \Lambda_M\mathbf{M} + \mathbf{E},$$

where $\mathbf{Y}$ is a $(G \times 1)$ vector of observed random variables, $\mathbf{T}^*$ is a $(K \times 1)$ vector of "trait" true score random variables, $\mathbf{M}$ is a $(Q \times 1)$ vector of "method" true score random variables, and $\mathbf{E}$ is a $(G \times 1)$ vector of error scores. The matrices Λ_{T^*} and Λ_M are $(G \times K)$ and $(G \times Q)$ coefficient matrices containing the regression relationships between the G observed variables and the K and Q latent trait and method latent variables. Note that with respect to the CTST model given previously, $\Lambda T = \Lambda_{T^*}\mathbf{T}^* + \Lambda_M\mathbf{M}$. The covariance structure for the model can be stated as

$$\Sigma_{YY} = [\Lambda_{T^*} \mid \Lambda_M]\, \Phi_T[\Lambda_{T^*}|\Lambda_M]' + \Theta^2,$$

where Φ_T has the following structure:

$$\Phi_T = \left[\begin{array}{c|c} \Phi_{T^*} & 0 \\ \hline 0 & \Phi_M \end{array}\right].$$

Note that the specification of the model places the constraint that the trait and method factors are uncorrelated. The estimation of this model permits the decomposition of reliable variance in each of the observed measures into valid and invalid parts.

Quasi-Simplex Models

A second approach to estimating reliability of single items uses the reinterview approach within a longitudinal design, or what are often called "panel" designs. The limitation of the test–retest approach using a single reinterview is that it must assume there is no change in the underlying quantity being measured. To address the issue of taking individual-level change into account, both Coleman and Heise developed a technique based on three-wave quasi-simplex models within the framework of a model that permits change in the underlying variable being measured. This approach can be generalized to multiwave panels. This class of autoregressive or quasi-Markov simplex model specifies two structural equations for a set of P over-time measures of a given variable Y_t (where $t = 1, 2, \ldots, P$) as follows:

$$Y_t = T_t + E_t$$
$$T_t = \beta_{t,t-1}T_{t-1} + Z_t.$$

The first equation represents a set of measurement assumptions indicating that over-time measures are assumed to be tau equivalent, except for true score change, and that measurement error is random. The second equation specifies the causal processes involved in change of the latent variable over time. Here, it is assumed that Z_t is a random disturbance representing true score change over time. This model assumes a lag-1 or Markovian process in which the distribution of the true variables at time t is dependent only on the distribution at time $t-1$ and not directly dependent on distributions of the variable at earlier times. If these assumptions do not hold, then this type of simplex model may not be appropriate. In order to estimate such models, it is necessary to make some assumptions regarding the measurement error structures and the nature of the true change processes underlying the measures. All estimation strategies available for such three-wave data require a lag-1 assumption regarding the nature of the true change. This assumption in general seems a reasonable one, but erroneous results can be obtained if it is violated. The various approaches differ in their assumptions about measurement error. One approach assumes equal reliabilities over occasions of measurement. This is often a realistic and useful assumption, especially when the process is not in dynamic equilibrium (i.e., when the observed variances vary with time). Another approach to estimating the parameters of the previous model is to assume constant measurement error variances rather than constant reliabilities. Where $P = 3$, either model is just-identified, and where $P > 3$ the model is overidentified with degrees of freedom equal to $0.5[P\ (P+1)] - 2P$. The four-wave model has two degrees of freedom, which can be used to perform likelihood ratio tests of the fit of the model.

One of the main advantages of the reinterview design, then, is that in appropriate circumstances it is possible to eliminate the confounding of the systematic error component discussed earlier, if systematic components of error are not stable over time. In order to address the question of stable components of error, the panel survey must deal with the problem of memory because in the panel design, by definition, measurement is repeated. Therefore, although this overcomes one limitation of cross-sectional surveys, it presents problems if respondents can remember what they say and are motivated to provide consistent responses. If reinterviews are spread over months or years, this can help rule out sources of bias that occur in cross-sectional studies. Given the difficulty of estimating memory functions, estimation of reliability from reinterview designs makes sense only if one can rule out memory as a factor in the covariance of measures over time, and thus the occasions of measurement must be separated by sufficient periods of time to rule out the operation of memory.

Reliability Models for Categorical Latent Variables

The approaches previously discussed rely on models in which the latent variable is assumed to be continuous

and in which the data can be assumed to vary according to an interval scale or an approximation (e.g., one that is at least ordinal in character). These assumptions, however, are clearly problematic for categorical latent variables. Latent-class models can be used to assess the extent of measurement error in measures of categorical variables. Several investigators have explored discrete-time Markov chain models, where the Markovian property is posited to hold at the level of the latent classes measured repeatedly. These models provide an analog to the conception of reliability involved in the structural equation modeling approaches for panel data discussed previously.

Conclusions

The foregoing discussion has concentrated on the theoretical background for strategies aimed at quantifying the reliability of measurement in social research. All definitions, relationships, and results were given for a hypothetical finite population (S), and nothing has been stated up to this point about sampling. In order to estimate reliability parameters for a given population of interest, one will need to sample the specific population using probability methods. In so doing, all the usual corrections for sample design effects and for sampling error must be taken into account in drawing inferences about reliability of measurement. This set of considerations is stressed in order to reinforce the fact that not only is the level of reliability influenced by the properties of the measuring device and the conditions of measurement but also, as a population parameter expressed by the ratio of true score to observed score variance, it is obviously influenced by the characteristics of the population to which the measures are applied.

See Also the Following Articles

Measurement Error, Issues and Solutions • Validity Assessment • Validity, Data Sources

Further Reading

Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality Quantity* **23,** 277–331.

Alwin, D. F. (2002). The reliability of survey data. Final report to the National Science Foundation (SES-9710403). Institute for Social Research, University of Michigan, Ann Arbor.

Alwin, D. F., and Jackson, D. J. (1979). Measurement models for response errors in surveys: Issues and applications. In *Sociological Methodology 1980* (K. F. Schuessler, ed.), pp. 68–119. Jossey-Bass, San Francisco.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Browne, M. W. (1984). The decomposition of multitrait–multimethod matrices. *Br. J. Math. Stat. Psychol.* **37,** 1–21.

Clogg, C. C., and Manning, W. D. (1996). Assessing reliability of categorical measurements using latent class models. In *Categorical Variables in Developmental Research—Methods of Analysis* (A. von Eye and C. C. Clogg, eds.), pp. 169–182. Academic Press, New York.

Coleman, J. S. (1968). The mathematical study of change. In *Methodology in Social Research* (H. M. Blalock, Jr., and A. B. Blalock, eds.), pp. 428–478. McGraw-Hill, New York.

Collins, L. M. (2001). Reliability for static and dynamic categorical latent variables: Development measurement instruments based on a model of growth processes. In *New Methods for the Analysis of Change* (L. M. Collins and A. G. Sayer, eds.), pp. 271–288. American Psychological Association, Washington, DC.

Duncan, O. D. (1984). The latent trait approach in survey research: The Rasch measurement model; Rasch measurement: Further examples and discussion. In *Surveying Subjective Phenomena* (C. F. Turner and E. Martin, eds.), Vols. 1 and 2, pp. 210–229, 367–440. Russell Sage Foundation, New York.

Greene, V. L., and Carmines, E. G. (1979). Assessing the reliability of linear composites. In *Sociological Methodology 1980* (K. F. Schuessler, ed.), pp. 160–175. Jossey-Bass, San Francisco.

Heise, D. R. (1969). Separating reliability and stability in test–retest correlation. *Am. Sociol. Rev.* **34,** 93–191.

Heise, D. R., and Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. In *Sociological Methodology 1970* (E. F. Borgatta and G. W. Bohrnstedt, eds.), pp. 104–129. Jossey-Bass, San Francisco.

Jöreskog, K. G. (1970). Estimating and testing of simplex models. *Br. J. Math. Stat. Psychol.* **23,** 121–145.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* **36,** 109–133.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika* **59,** 381–389.

Langeheine, R., and van de Pol, F. J. R. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociol. Methods Res.* **18,** 416–441.

Lord, F. M., and Novick, M. L. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **49,** 115–132.

Saris, W. E., and Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In *Measurement Errors in Surveys* (P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, eds.), pp. 575–597. Wiley, New York.

Reliability Assessment

Edward G. Carmines
Indiana University, Bloomington, Indiana, USA

James A. Woods
West Virginia University, Morgantown, West Virginia, USA

Glossary

measurement The process of linking abstract concepts to empirical indicators of those concepts.
random measurement error All of the chance factors that confound the measurement of any phenomenon.
reliability The extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials.
validity The extent to which an indicator of some abstract concept measures what it purports to measure.

Reliability focuses on the extent to which an empirical measurement yields consistent results in repeated trials. The more consistent a measure is, the more reliable it is. If a bathroom scale repeatedly registers a person's weight at 5 pounds less than the true weight, that scale is consistent on repeated trials—it is reliable. However, that scale is not a valid indicator of weight, because it is never correct. That is the difference between these two basic properties of empirical measurement. Validity is the extent to which the indicator measures what it purports to measure. Reliability is the extent to which the indicator is consistent across attempts.

Measurement

Measurement is crucial to science. From the perspective of the social sciences, measurement may be thought of as the process of linking abstract concepts to empirical indicators of those concepts. It focuses on the relationship between the empirically grounded indicator, the observable response, and the underlying unobservable concept. Thus, measurement allows the social scientist to move from abstract, indirectly observable concepts and theories to empirical, directly observable indicators of those theoretical concepts. Many studies in the social sciences represent abstract concepts by a single empirical indicator. For example, social status is sometimes measured by occupational prestige. Or the state of the economy is measured by per capita income. From a measurement perspective, this reliance on single indicators is undesirable for two reasons. First, when using a single indicator, it is almost always impossible to estimate reliability. To do this, there must be some *a priori* information available, and this is usually not the case. Second, even if the reliability of a single indicator can be estimated, it is more affected by random error, as compared to a composite measure made up of two or more indicators.

Random and Nonrandom Measurement Error

Two basic types of errors affect empirical measurements: random error and nonrandom error. Random error consists of all of the chance factors that confound the measurement of any phenomenon. If an indicator is a reliable indicator of a theoretical concept, it will produce consistent results on repeated observations because random error does not cause systematic fluctuation from one observation to the next. That is, to take a very simple example, if a scale records an individual's weight as 5 pounds less than the true weight on the first attempt,

5 pounds more than the true weight on the second attempt, 8 pounds more than the true weight on the third attempt, and 10 pounds less than the true weight on the fourth attempt, that scale is not very reliable as a measure for weight. The more random the error, the more unreliable is the measure.

Classical Test Theory

Classical test theory is used to assess random measurement error. By determining the amount of random error, reliability can be estimated.

Reliability

Random error is present in any measure. Reliability focuses on the assessment of random error and estimating its consequences. Although it is always desirable to eliminate as much random error from the measurement process as possible, it is even more important to be able to detect the existence and impact of random error. Because random error is always present to at least a minimum extent, the basic formulation in classical test theory is that the observed score is equal to the true score that would be obtained if there were no measurement error plus a random error component, or $X = t + e$, where X is the observed score, t is the true score, and e is the random disturbance. The true score is an unobservable quantity that cannot be directly measured. Theoretically, it is the average that would be obtained if a particular phenomenon was measured an infinite number of times. The random error component, or random disturbance, indicates the differences between observations.

Classical test theory makes the following assumptions about measurement error:

Assumption 1: The expected random error is zero,

$$E(e) = 0.$$

Assumption 2: The correlation between the true score and random error is zero,

$$\rho_{(t,e)} = 0.$$

Assumption 3: The correlation between the random error of one variable and the true score of another variable is zero,

$$\rho_{(e_1,t_2)} = 0.$$

Assumption 4: The correlation between errors on distinct measurements is zero,

$$\rho_{(e_1,e_2)} = 0.$$

From these assumptions, we see that the expected value of the observed score is equal to the expected value of the true score plus the expected value of the error:

$$E(X) = E(t) + E(e).$$

However, because, by assumption 1, the expected value of e is zero, $E(e) = 0$, then,

$$E(X) = E(t).$$

This formula applies to repeated measurements of a single variable for a single person. However, reliability refers to the consistency of repeated measurements across persons and not within a single person. The equation for the observed score may be rewritten so that it applies to the variances of the single observed score, true score, and random error:

$$\mathrm{Var}(X) = \mathrm{Var}(t + e) = \mathrm{Var}(t) + 2\mathrm{Cov}(t, e) + \mathrm{Var}(e).$$

Assumption 2 stated that the correlation (and covariance) between the true score and random error is zero, so $2\ \mathrm{Cov}(t, e) = 0$. Consequently,

$$\mathrm{Var}(X) = \mathrm{Var}(t) + \mathrm{Var}(e).$$

So the observed score variance equals the sum of the true score variance and the random error variance. Reliability can be expressed as the ratio of the true score variance to the observed score variance:

$$\rho_x = \frac{\mathrm{Var}(t)}{\mathrm{Var}(X)}.$$

That is, ρ_x is the reliability of X as a measure of t. Alternatively, reliability can be expressed in terms of the error variance as a proportion of the observed variance:

$$\rho_x = 1 - \left[\frac{\mathrm{Var}(e)}{\mathrm{Var}(X)}\right].$$

This equation makes it clear that reliability varies between 0 and 1. If all observed variance consists of error, then reliability will be 0, because $1 - (1/1) = 0$. At the other extreme, if there was no random error in the measurement of some phenomenon, then reliability will be 1, because $1 - (0/1) = 1$.

Parallel Measurements

One estimate of a measure's reliability can be obtained by correlating parallel measurements. Two measurements are defined as parallel if they have identical true scores and equal error variances. Thus, X and X' are parallel if $X = t + e$ and $X' = t + e'$, where $\mathrm{Var}(e) = \mathrm{Var}(e')$ and $t = t$. Thus, parallel measures are functions of the same true score, and the differences between them are the result of purely random error.

Assessing Reliability

There are four basic methods for estimating the reliability of empirical measurements. These are the retest method, the alternative-form method, the split-halves method, and the internal consistency method.

Retest Method

The most straightforward and intuitive method of assessing reliability is to correlate the same measures administered at different points in time. Figure 1 shows a representation of the retest method. The equations for the tests at times 1 and 2 can be written as follows:

$$X_1 = X_t + e_1$$

and

$$X_2 = X_t + e_2.$$

Because in parallel measures $t = t$ and $\mathrm{Var}(e_1) = \mathrm{Var}(e_2)$, and by assumption $\rho_{(e_1,t_2)} = 0$ and $\rho_{(e_1,e_2)} = 0$, it follows that $\rho_x = \rho_{x_1x_2}$. That is, reliability is the correlation between the scores on the same test obtained at two points in time. If the retest reliability coefficient is exactly 1.0, the results on the two administrations of the test are the same. However, because there is almost always random measurement error, the correlations across time will not be perfect.

Although test–retest correlations are a simple and intuitively appealing way to assess reliability, they have some serious problems and limitations. First, often researchers can obtain a measure only at a single point in time. It may be too expensive or impractical to measure the phenomenon at different times. Moreover, if the test–retest correlations are low, it may not indicate unreliability but rather the possibility that, in the interim, the theoretical concept of interest has undergone a change. For example, if the measure of support for a candidate at the beginning of a campaign differs from that at the end of a campaign, the suspicion would probably be that there was real change in the candidate's popularity during the course of the campaign—not that the measurement of popularity was necessarily unreliable.

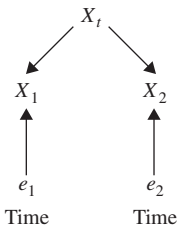


Figure 1 A representation of the retest method for estimating reliability.

Another problem that affects test–retest correlations and lowers reliability estimates is reactivity. Sometimes the very act of measuring a phenomenon can induce change in the phenomenon. A person may become sensitized to the concept of being measured and thus change his/her answer based only on the earlier measurement. Lowered reliability estimates are not the only effects. If the period between time 1 and time 2 is short enough, the subject may remember his/her answer at time 1 and thus appear more consistent than is actually the case. These memory effects can inflate reliability estimates.

Alternative-Form Method

The alternative-form method of assessing reliability is similar to the retest method in that it requires administering a test to the same people at different points in time. The difference between this method and the retest method is that an alternative form of the same test is given on the second testing. Because the two forms of the same test are intended to measure the same theoretical concept, they should not differ systematically from each other. Using random procedures to select items for the different forms of the test can ensure that the tests do not differ systematically. In this case, the reliability is the correlation between the alternative forms of the test. The two forms should be administered about 2 weeks apart, to allow for day-to-day fluctuations in the individual.

This method for assessing reliability is superior to the retest method because it reduces the extent to which an individual's memory can inflate the reliability estimate. However, this method, like the retest method, does not allow the researcher to distinguish true change from the unreliability of the measure if the measurement takes place only on two occasions. Therefore, the results of alternative-form reliability studies are easier to interpret if the phenomenon being measured is relatively enduring and not subject to rapid and radical alteration. A limitation of this method is the difficulty of designing alternative forms that are truly parallel.

Split-Halves Method

Reliability estimated by using the split-halves method, unlike the retest or alternative-form methods, can be conducted on only one occasion. In the split-halves method, the total set of parallel items is divided into halves and the scores on the halves are correlated to yield an estimate of reliability. The halves can be considered approximations to alternative forms. The correlations between the two halves would be the reliability for each half of the test and not for the total test. To estimate the reliability of the entire test, a statistical correction, called the Spearman–Brown prophecy formula, must be estimated.

Because the total test is twice as long as each half, the Spearman–Brown prophecy formula is expressed as

$$\rho_{xx''} = \frac{2\rho_{xx'}}{1 + \rho_{xx'}}, \quad (1)$$

where $\rho_{xx''}$ is the reliability coefficient for the whole test and $\rho_{xx'}$ is the split-half correlation. For example, if the correlation between the halves is 0.75, the reliability for the whole test would be $[(2)(0.75)]/(1 + 0.75) = 0.857$. The reliability coefficient varies between 0 and 1, taking on these extreme values if the correlation between the halves is 0.00 or 1.00.

The more general version of the Spearman–Brown prophecy formula is

$$\rho_{x_n x_n''} = \frac{N\rho_{xx'}}{1 + (N-1)\rho_{xx'}}. \quad (2)$$

This formula estimates the reliability of a scale that is N times longer than the original scale. A researcher can also use the Spearman–Brown prophecy formula to determine the number of items that would be needed to attain a given reliability. To estimate the number of items required to obtain a particular reliability, the following formula is used:

$$N = \frac{\rho_{xx''}(1 - \rho_{xx'})}{\rho_{xx'}(1 - \rho_{xx''})}, \quad (3)$$

where $\rho_{xx''}$ is the desired reliability, $\rho_{xx'}$ is the reliability of the existing test, and N is the number of times the test would be lengthened to obtain reliability of $\rho_{xx''}$. For example, if a 10-item test has a reliability of 0.60, then the estimated lengthening required to obtain a reliability of 0.80 would be $N = 0.8(1 - 0.6)/0.6(1 - 0.8) = 2.7$. In other words, 27 parallel items would be required to attain a reliability of 0.80.

There is an element of indeterminateness in using the split-halves technique to estimate reliability, due to the different ways that the items can be grouped into halves. The most typical way to divide the items is to place the even-numbered items in one group and the odd-numbered items in the other group. But other ways of partitioning the total item set are also used, including separately scoring the first and second halves of the items and randomly dividing the items into two groups. For a 10-item scale, there are 126 different splits, and each will probably result in a slightly different correlation between the two halves, which will lead to a different reliability estimate. It is therefore possible to obtain different reliability estimates even if the same items are administered to the same individuals.

Internal Consistency Methods

Because of the indeterminateness of the split-halves technique, other methods that do not require either splitting or repeating items have been developed to estimate reliability. These techniques go under the general rubric of "measures of internal consistency." The most popular of these measures is Cronbach's alpha. Cronbach's alpha is equal to the average of all possible split-half correlations for a composite scale $2N$ items long, and is calculated from the variance–covariance matrix as follows:

$$\alpha = \frac{N}{N-1}\left[1 - \frac{\sum \mathrm{Var}(Y_i)}{\mathrm{Var}_x}\right], \quad (4)$$

where N is the number of items, $\sum \mathrm{Var}(Y_i)$ is the sum of the item variances, and Var_x is the variance of the total composite. If the correlation matrix rather than the variance–covariance matrix is being used, the formula becomes

$$\alpha = \frac{a}{a-1}\left[1 - \frac{a}{a+2b}\right], \quad (5)$$

where a is the number of items in the composite and b is the sum of the correlations among the items. Alpha is a lower bound to the reliability of an unweighted scale of N items, i.e., the reliability is always equal to or greater than alpha. It is equal to the reliability if the items are parallel. The reliability of a scale can never be lower than alpha, even if the items depart substantially from being parallel measurements. Thus, alpha is a conservative estimate of reliability.

Cronbach's alpha is a generalization of Kuder and Richardson's procedure, which was designed to estimate the reliability of scales composed of dichotomously scored items. These items are scored either 0 or 1, depending on whether the individual possesses the particular characteristic of interest. Because Cronbach's alpha can handle dichotomously scored items or items that can take on three or more values, because it encompasses the Spearman–Brown prophecy formula, because it makes use of all of the information contained in the items, and because it is easy to compute, it is a very popular estimate of reliability.

Correction for Attenuation

No matter which specific method is used for obtaining an estimate of reliability, one of the estimate's most important uses is to "correct" correlations for unreliability due to random measurement error. If the reliability of each variable can be estimated, these estimates can be used to determine what the correlation between the two variables would be if they were made perfectly reliable. This process is called correction for attenuation. The formula for the correction for attenuation is

$$\rho_{x_t y_t} = \frac{\rho_{x_t y_t'}}{\sqrt{\rho_{xx'}\rho_{yy'}}}, \quad (6)$$

where $\rho_{x_t y_t}$ is the correlation corrected for attenuation, $\rho_{xx'}$ is the reliability of X, $\rho_{yy'}$ is the reliability of Y, and $\rho_{x_t y_t'}$ is the observed correlation between x and y. By correcting correlations between measures for attenuation due to unreliability, the researcher gains a more accurate estimate of the correlation between the underlying theoretical concepts.

See Also the Following Articles

Inter-Rater Reliability • Randomization • Reliability • Split-Half Reliability • Test–Retest Reliability • Validity Assessment • Validity, Data Sources

Further Reading

Carmines, E. G., and Zeller, R. A. (1979). *Reliability and Validity Assessment.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-017. Sage Publ., Newbury Park, CA.

DeVellis, R. F. (1991). *Scale Development: Theory and Applications.* Sage Publ., Newbury Park, CA.

Lord, R. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA.

Pedhazur, E. J., and Pedhazur, L. P. (1991). *Measurement, Design, and Analysis: An Integrated Approach.* Lawrence Erlbaum Assoc., Hillsdale, NJ.

Traub, R. E. (1994). *Reliability for the Social Sciences: Theory and Applications.* Sage Publ., Thousand Oaks, CA.

Zeller, R. A., and Carmines, E. G. (1980). *Measurement in the Social Sciences: The Link Between Theory and Data.* Cambridge University Press, Cambridge.

Religious Affiliation and Commitment, Measurement of

David E. Campbell
University of Notre Dame, Notre Dame, Indiana, USA

Glossary

black Protestants/Black Church A religious tradition composed of denominations historically governed by and in service of African Americans. Although these denominations share many beliefs and organizational characteristics with white evangelical churches, they are nonetheless distinct owing to their doctrinal emphases on social justice and civil rights and to their social role as one of the few kinds of organizations free from white control throughout African American history.

denomination The specific religious organization to which someone belongs, often affiliated with a national or international organization (e.g.. Southern Baptists).

evangelical Protestants A religious tradition characterized by a conservative theology. Evangelical Protestants generally believe that an individual must be personally converted to a belief in Jesus Christ. They also emphasize the importance of the Bible and stress the need for missions in order to convert others.

mainline Protestants A religious tradition characterized by a more modernist (some would say liberal) theology and interpretation of scripture, with less emphasis on personal piety.

religious tradition A general grouping of denominations, united by a common history and common beliefs (e.g., evangelical Protestants).

The religious diversity within the United States makes religion simultaneously important and difficult to understand. Survey researchers must distinguish between an individual's religious affiliation (church membership) and level of religious commitment. This entry offers practical guidance for researchers relying on existing data sets, such as the General Social Survey and American National Election Studies, as well as scholars designing their own data collection.

Introduction

The last decade or so has seen a revival of the study of religion within the social sciences. This has been especially true in the United States, where religiously themed issues regularly ignite controversy in the public square and a large majority of Americans profess adherence to religion (the predictions of secularization theorists to the contrary notwithstanding). Along with this increasing scholarly interest in religion has come increasing sophistication regarding the best methods to measure individuals' religious involvement. Even scholars whose primary interest is not religion can benefit from these advances in the measurement of religious variables because accurate measures are necessary to avoid biasing the interpretation of other statistical relationships in models that control for religious engagement.

Scholars of religion have come to recognize that the measurement of religious engagement requires information about both religious affiliation and commitment, or what the literature on the subject often calls belonging and behaving (a third dimension, believing, is not relevant to this particular discussion). Belonging refers to religious identity, one's denomination and/or religious tradition. Behaving entails religious activities, whether public (such as attending worship services) or private (such as praying). These two dimensions of religious involvement are not orthogonal, but religious researchers have nonetheless found them to be empirically and

theoretically distinctive enough to warrant separate consideration.

Belonging: Religious Affiliation

The sheer number of religious denominations within the United States can be overwhelming to any researcher trying to make sense of the religious groups to which Americans belong. The codebook for the 2000 American National Election Study (ANES), for example, lists over 130 possible denominations, a list that grows with each extension of the ANES time series. In contrast, in the 1950s the sole religiously relevant item in the ANES was simply whether the respondent identified as Catholic, Protestant, or Jewish. This expanding list of denominations is partly a function of the growing religious diversity within the United States, but also reflects social scientists' increasingly nuanced measurement of religious affiliation over the last 50 years.

The first step in measuring denominational affiliation in survey research, therefore, is ensuring that the design of the survey enables respondents to identify their denomination. Experience suggests that neither confronting respondents with a massive list of denominations nor asking them to recall the specific affiliation of their church without a prompt is likely to be very successful. A better method is to follow the model of the ANES, which uses a branching questionnaire to determine religious affiliation (see Table I). Respondents are first asked if they ever attend religious services. If they do, the follow-up inquires whether their place of worship is Protestant, Roman Catholic, Jewish, or something else. Those who do not attend services are asked, "Regardless of whether you now attend religious services, do you ever think of yourself as part of a particular church or denomination?" If the answer is yes, they too are asked to identify themselves as Protestant, Catholic, Jewish, or another faith. Protestants are then asked about their specific affiliation (Baptist, Methodist, etc.), with precise follow-up questions to distinguish among subgroups within each denomination (e.g. Evangelical Lutheran Church in America vs. Missouri Synod Lutherans). Similarly, Jews are asked whether they consider themselves Orthodox, Conservative, or Reform. In parallel fashion, respondents choosing the "Other" category also have the opportunity to specify their denomination. Although the branching format of this item is best suited to a telephone or Internet survey, it can be adapted for self-administered surveys. Whenever possible, any survey should also collect the verbatim name of the church a respondent attends. This can provide invaluable information when classifying ambiguous responses.

Table 1 Questions to Determine Religious Affiliation[a]

(If respondent attends religious services): Do you mostly attend a place that is Protestant, Roman Catholic or what?
If respondent does not attend religious services: Regardless of whether you now attend any religious services do you ever think of yourself as part of a particular church or denomination? (If yes): Do you consider yourself Protestant, Roman Catholic, Jewish, or what?
Protestants:
(If Baptist): With which Baptist group is your church associated? Is it the Southern Baptist Convention, the American Baptist Churches in the USA, the American Baptist Association, the National Baptist Convention, USA, an independent Baptist church or some other Baptist group? (If independent Baptist): Are you affiliated with any larger Baptist group or is this strictly a local church?
(If Lutheran): Is this church part of the Evangelical Lutheran Church in America, the Missouri Synod, or some other Lutheran group?
(If Methodist): Is your church part of the United Methodist Church, African Methodist Episcopal, or some other Methodist group?
(If Presbyterian): Is this the Presbyterian Church in the USA or some other Presbyterian group?
(If Reformed): Is this the Christian Reformed Church, the Reformed Church in America, or some other Reformed group?
(If Brethren): Is this the Church of the Brethren, the Plymouth Brethren, or what?
(If Christian or "just Christian"): When you say "Christian" does that mean the denomination called the "Christian Church"
(Disciples of Christ) or some other Christian denomination, or do you mean to say "I am just a Christian?"
(If Church or Churches of Christ): Is this the Church of Christ or United Church of Christ?
(If Church of God): Is this the Church of God of Anderson, Indiana, the Church of God of Cleveland, Tennessee, the Church of God in Christ, or some other Church of God?
(If Holiness or Pentecostal): What kind of church is that? What is it called exactly? Is that part of a larger church or denomination? What is that church called?
(If other): What is it called exactly? Is that church part of a denomination? Is that group Christian?
(If Jewish): (If respondent attends religious services): Do you usually attend a synagogue or temple that is Orthodox, Conservative, Reform, or what? (If respondent does not attend religious services but considers self Jewish): Do you consider yourself Orthodox, Conservative, Reform, or what?

[a] From the American National Election Studies.

Although providing an array of denominational options allows the analyst to make subtle distinctions among the churches to which respondents belong, when analyzing the data the scope of denominational choices may seem to be as much a curse as a blessing. The generally accepted strategy for dealing with the complexities of religious affiliation is to collapse the denominations into a smaller set of religious traditions. As defined in 1996 by Kellstedt *et al*., a religious tradition is "a group of religious communities that share a set of beliefs that generates a distinctive worldview"—a family of denominations with a common history and trajectory.

Moving from an array of denominations to a set of religious traditions requires an appreciation for the subtleties of America's religious landscape. The Evangelical Lutheran Church in America (ELCA) is a good example of why sensitivity to the wrinkles of denominationalism is important. Notwithstanding the denomination's name, members of the ELCA are not classified as members of the evangelical tradition. This is because "evangelical" is generally applied to theologically conservative Protestants, and the ELCA is more accurately classified as a mainline (i.e., liberal) denomination.

To assist analysts in finding their way through the thickets of America's many denominations, Steensland *et al*. developed in 2000 a classification system that groups denominations into a manageable number of religious traditions, based on their historical commonalities. Building on the work of Kellstedt and his colleagues, Steensland *et al*. emphasized that traditions should be treated as nominal, not ordinal, categories. This contrasts with the approach of Smith in 1990, who classified denominations by where they fell along a continuum defined by the terms fundamentalist, moderate, and liberal. Smith's approach has become quasi-institutionalized; a variable in the public General Social Survey (GSS) data file places each respondent who identifies with a denomination along this continuum. Steensland and his colleagues, however, argued that this system is not conceptually clear and offers limited analytical leverage for understanding the impact religion has on Americans' social and political attitudes. Instead they recommended that individuals be grouped into a relatively small number of denominational categories, assuming no ordinal relationship among them.

Within Protestantism, Steensland and his colleagues recommended that the analyst divide respondents into three mutually exclusive religious traditions. The first is the black Protestant tradition. Because the Black Church has historically been the central institution in the African American community, its doctrinal emphases and religious practices differ from white Protestant denominations. Although there is widespread consensus regarding the need to classify black Protestants as a distinct tradition, researchers have employed two slightly different methods of classifying members of the Black Church. One is to limit the category to African Americans who attend a church that is affiliated with a historically black denomination; the other is to include all African Americans with any Protestant affiliation, regardless of the specific denomination. The decision of which method to use depends on the theoretically grounded assumptions of the analyst. In the first method, the analyst is assuming that the distinctiveness of black Protestants is dependent on their attending a specific type of church. In the second, the implicit assumption is that the most important criterion is the respondents' race and not denominational affiliation. Empirically, differences between the two methods are minor because almost all blacks who affiliate with a church belong to one of seven predominantly black denominations: African Methodist Episcopal (A.M.E.) church; African Methodist Episcopal Zion (A.M.E.Z.) church; Christian Methodist Episcopal (C.M.E.) church; National Baptist Convention, U.S.A, Incorporated (NBC); National Baptist Convention of America, Unincorporated (NBCA); Progressive National Baptist Convention (PNBC), and Church of God in Christ (COGIC).

In addition to a division along racial lines, among white Protestants a distinction should be made between members of mainline and evangelical denominations. Mainline Protestant churches are characterized as having a theology that seeks accommodation with modernity, with more emphasis on social justice than personal piety. They typically do not hold to a literal interpretation of scripture. Evangelical denominations, on the other hand, generally have a strict interpretation of the Bible, stress personal piety as paramount, and emphasize the need for their members to remain separate from the wider cultural milieu. These differences are due to the fact that evangelical churches are generally associated with sect movements, which emphasize separation from more established denominations, whereas mainline churches are more attuned to ecumenism.

Table II contains a list of denominations classified as belonging to the mainline or evangelical traditions. Although this and similar lists include a large number of denominations covering most of the Protestant universe, the sheer number of churches makes it impossible for them to be totally comprehensive. Upon encountering a denomination that is not classified, researchers will find the *Encyclopedia of American Religions* invaluable. As the most comprehensive source of information about individual denominations in the United States, this encyclopedia generally provides enough information about a particular church to classify it as evangelical or mainline.

Another significant religious tradition in the United States is Roman Catholicism, a category with little ambiguity. Similarly, self-identified Jews are easily classified. Given the widespread use of the terms, researchers

Table II Classifying Denominations into Religious Traditions[a]

Black Protestant

African Methodist
African Methodist Episcopal Church
African Methodist Episcopal Zion Church
American Baptist Association
American Baptist Churches in the USA
Apostolic Faith
Baptists, Don't Know Which (if respondent is black)
Christian Tabernacle
Church of God in Christ
Church of God in Christ Holiness
Church of God, Saint, and Christ
Disciples of God
Federated Church
Holiness, Church of Holiness
House of Prayer
Methodist, Don't Know Which (if respondent is black)
Missionary Baptist
National Baptist Convention of America
National Baptist Convention, USA, Inc.
Other Baptist Churches (if respondent is black)
Other Methodist Churches
Pentecostal Apostolic
Primitive Baptist
Sanctified, Sanctification
Southern Baptist Convention (if respondent is black)
United Holiness
Witness Holiness
Zion Union
Zion Union Apostolic
Zion Union Apostolic-Reformed

Evangelical Protestant

Advent Christian
Amish
American Baptist Association
Apostolic Christian
Apostolic Church
Assembly of God
Baptist, Don't Know Which
Bible Missionary
Brethren Church, Brethren
Brethren, Plymouth
Brother of Christ
Calvary Bible
Chapel of Faith
Charismatic
Chinese Gospel Church
Christ Cathedral of Truth
Christ Church Unity
Christian and Missionary Alliance
Christian Calvary Chapel
Christian Catholic
Christian, Central Christian
Christian Reformed
Christ in Christian Union
Christ in God
Churches of God (Except with Christ and Holiness)
Church of Christ
Church of Christ, Evangelical
Church of Daniel's Band
Church of God of Prophecy, The
Church of Prophecy
Church of the First Born
Church of the Living God
Community Church
Covenant
Dutch Reformed
Evangelical Congregational
Evangelical Covenant
Evangelical, Evangelist
Evangelical Free Church
Evangelical Methodist
Evangelical United Brethren
Faith Christian
Faith Gospel Tabernacle
First Christian
Four Square Gospel
Free Methodist
Free Will Baptist
Full Gospel
Grace Brethren
Holiness Church of God
Holiness (Nazarene)
Holy Roller
Independent
Independent Bible, Bible, Bible Fellowship
Independent Fundamental Church of America
Laotian Christian
Living Word
Macedonia
Mennonite
Mennonite Brethren
Missionary Baptist
Missionary Church
Mission Covenant
Nazarene
New Testament Christian
Lutheran Church-Missouri Synod
Open Bible
Other Baptist Churches
Other Fundamentalist
Other Lutheran Churches
Other Methodist Churches
Other Presbyterian Churches
Pentecostal
Pentecostal Assembly of God
Pentecostal Church of God
Pentecostal Holiness, Holiness Pentecostal
People's Church

continues

Table II *continued*

Pilgrim Holiness
Primitive Baptist
Salvation Army
Seventh Day Adventist
Swedish Mission
Triumph Church of God
Way Ministry, The
Wesleyan
Wesleyan Methodist-Pilgrim
Southern Baptist Convention
Wisconsin Evangelical Lutheran Synod

Mainline Protestant

American Baptist Churches in the USA
American Lutheran Church
American Reformed
Baptist (Northern)
Christian Disciples
Congregationalist, First Congregationalist
Disciples of Christ
Episcopal Church
Evangelical Lutheran
Evangelical Reformed
First Christian Disciples of Christ
First Church
First Reformed
Friends
Grace Reformed
Hungarian Reformed
Latvian Lutheran
Lutheran Church in America
Lutheran, Don't Know Which
Methodist, Don't Know Which
Moravian
Presbyterian Church in the USA
Presbyterian, Don't Know Which
Presbyterian, Merged
Quaker
Reformed
Reformed Church of Christ
Reformed United Church of Christ
Schwenkfelder
United Brethren, United Brethren in Christ
United Church of Canada
United Church of Christ
United Church of Christianity
United Methodist Church
United Presbyterian Church in the USA

Other Affiliation

Conservative Nontraditional
Christadelphians
Christian Scientist
Church of Jesus Christ of the Restoration
Church Universal and Triumphant
Jehovah's Witnesses
Jesus LDS
LDS
LDS-Mormon
LDS-Reorganized
Mormon
True Light of Church of Christ
Worldwide Church of God

Liberal Nontraditional
Christ Church Unity
Eden Evangelist
Mind Science
New Age Spirituality
New Birth Christian
Religious Science
Spiritualist
Unitarian, Universalist
United Church, Unity Church
Unity

From Steensland *et al.* (2000) with permission.

will profit from asking Jewish respondents to identify themselves as belonging to a specific tradition with Judaism, such as Orthodox, Reform, and so on, as in the ANES.

In spite of their relatively small numbers, most studies treat Jews as being distinct from other religious groups. This is not the case for a group of other denominations that do not easily fit within the aforementioned religious traditions, some of which are similar in size, and (arguably at least) equally distinctive. In spite of their heterogeneity, these nontraditional churches are typically grouped together in a catch-all "Other" category. Typical examples include the Latter-day Saints (LDS) and Christian Scientists, as well as non-Judeo-Christian religions (such as Islam) and Eastern Orthodox churches. Some analysts further subdivide some members of this catch-all group and distinguish between conservative nontraditional and liberal nontraditional denominations. LDS is an example of the former, and Unitarian-Universalist is an example of the latter.

For most surveys of the general population, particularly those with a national scope, the small number of respondents in these categories make any further distinctions statistically meaningless. However, some sampling frames could produce sizable numbers of respondents within the groups normally relegated to the "Other" or "Nontraditional" categories. For example, because of the

LDS' geographic concentration, a survey within western states will probably have a large number of LDS respondents. Owing to their distinctive theology and social cohesion, whenever their numbers warrant it they should be classified as a separate group. The same applies to other distinctive groups such as Muslims and members of Eastern Orthodox denominations, populations that are also geographically concentrated in some areas of the United States.

Thus far, the discussion has centered only on respondents who identify with a specific religious denomination. Between 2 and 5% of respondents to surveys such as the GSS, however, are classified as Protestants without a denomination. The ambiguity of nondenominationalism leaves the analyst with the challenge of determining whether nonaffiliation is a reflection of a person's secularism or of membership in an explicitly nondenominational church. This is hardly a trivial question because nondenominationalism is a growing trend in American Protestantism. Most notably, mega-churches, which are capturing an ever-increasing share of the religious market, are usually nondenominational. Steensland and his colleagues addressed this issue by dividing respondents identified as nondenominational Protestants between those who report attending church once a month or more and those who report attending less frequently. They classify nonaffiliated respondents who attend church at least monthly as evangelical Protestants, on the grounds that the nondenominational movement is restricted to evangelical churches. In 1996, Kellstedt and his colleagues, who proposed a classification system that is otherwise similar to the Steensland *et al*. scheme, labeled nonaffiliated Protestants with a minimal level of religious commitment as secularists, a category that also includes respondents who report that they are atheists or agnostics. Although this is not a self-identified tradition in the same sense as, say, evangelicalism, Kellstedt *et al*. nonetheless argued persuasively that secularists constitute a distinct group in contemporary American society, an argument bolstered by Layman's 2001 evidence of secularists' impact on the political landscape of the United States.

Behaving: Religious Commitment

The introduction of secularists as a religious tradition underscores that measuring denominational affiliation without also gauging religious commitment paints an incomplete picture of religion's role in American society. There are numerous measures of religious behavior as a form of commitment, but by far the most common is attendance at religious services. Being common, however, does not mean that the measure of church attendance is without controversy. On the contrary, the accuracy of the church attendance measures in surveys such as the GSS is a matter of considerable dispute.

For decades, numerous sources of data have estimated a strikingly similar rate of church attendance in the United States—around 40% of Americans report attending worship services each week. This reported rate of church attendance is remarkably robust to alternative survey questions. Gallup asks respondents if they attended church "last week"; other surveys, notably the GSS, typically ask respondents how frequently they attend church. Both methods produce essentially the same estimated levels of church attendance. In spite of this consistency, however, some scholars have questioned whether surveys accurately gauge the rate of attendance at religious services. The standard questions probably facilitate telescoping, which occurs when a respondent reports having been in church over a wider time span than specified by the interviewer (e.g., someone who attended church 2 weeks prior reporting that she had attended the previous week). They also do little to minimize social desirability bias, the tendency for people to report that they attend church frequently because of the normative value attached to religious commitment in the United States.

These suspicions were articulated most forcefully by Hadaway, Marler, and Chaves in 1993; their research casts doubt on the conventional wisdom that 40% of Americans attend church every week. They collected data on the attendance at religious services in Protestant churches in a single county and a selection of Catholic dioceses nationally and concluded that survey-based estimates are roughly twice the observed rate of church attendance. The research of Hadaway, Marler, and Chaves has proven to be controversial, provoking a plethora of studies revisiting the question of how frequently Americans attend religious services and how church attendance should be measured. In a follow-up to their seminal 1993 article, Marler and Hadaway in 1999 answered many of their critics' objections by testing the accuracy of self-reported church attendance within a single church's congregation. Marler and Hadaway counted the attendance at a particular church by telephone and then surveyed that congregation by telephone over the next week, asking respondents whether they had attended church within the previous 7 days. As with the earlier study, they found substantial overreporting of church attendance, in the range of 59 to 83% (depending on various assumptions).

At this writing, no consensus has developed in the wake of the claims made by scholars who question the validity of survey-based estimates of how many Americans regularly darken the doors of their churches. One camp is adamant that surveys produce hopelessly inaccurate estimates of church attendance, whereas researchers on the other side of the debate are quick to point to the problems endemic

to head counts in the pews. This is not the place to attempt to resolve this debate; instead, the discussion here centers on what can been learned about measuring church attendance from the literature on the subject.

In response to Hadaway and his colleagues, Woodberry pointed out that an extremely important but typically ignored factor affecting the validity of any survey-based measure of religious behavior is simply the quality of the survey's sample. Woodberry noted that surveys with a high refusal rate oversample church-goers because "highly religious respondents are generally . . . more cooperative and thus presumably are less likely to refuse an interview." Whereas a high response rate is always desirable to minimize sampling bias, for a valid measure of church attendance it is essential. Woodberry also stressed that atheoretical weighting schemes designed to correct for sampling bias can accentuate the inaccuracy of survey-based church attendance measures. For example, weighting by gender undersamples women who work full-time, who attend church less frequently than do homemakers. Woodberry also pointed out that telephone surveys oversample churchgoers because the 5–8% of the U.S. population without a phone is less likely to be regular church attenders than those who do have a phone.

In addition to ensuring that the quality of the sample is as high as possible, researchers should pay careful attention to the wording and context of the survey item gauging church attendance. Based on an experiment conducted in the 1996 GSS, Smith recommended in 1998 that telescoping be minimized by repeatedly reminding respondents that their responses to the church attendance question are only to cover the previous 7 days. Desirability bias can be avoided by including church attendance in a list of other activities, such as visiting a doctor (see Table III). Using this method rather than the standard GSS question significantly lowered the percentage of respondents who reported attending church in the previous week, from the usual 40% to just over 30%, which Smith interpreted as weeding out a substantial portion of inaccuracies.

Smith also reported evidence regarding what people mean when they indicate they have attended "religious services" because this term could mean anything from a worship service to Bible study group to watching a religious program on television. Again, data from the 1996 GSS speak to this question. Respondents who answered affirmatively when asked about attending religious services were asked to indicate whether in the previous 7 days they had attended a weekly worship service, participated in another type of religious meeting, or tuned into a religious program on television or radio (see Table III).

Of those who indicated that they attended religious services in the first question, 4.6% have an inconsistent response in the follow-up question, and another 6.5% were uncertain. Because space on a survey questionnaire is always at a premium, theoretical concerns should determine whether or not this follow-up question is included when measuring church attendance. For some purposes, it may be enough to know the frequency with which the respondent has a communal religious experience without greater specificity; for other purposes, specifying attendance at a worship service is necessary.

Of course, these suggestions for measuring church attendance apply only to analysts designing their own survey instruments. When using secondary sources of data, researchers are limited by the measure employed. In this case, researchers should at least be aware that the marginal frequencies of church attendance are likely to be biased, although the precise extent of that bias remains in dispute.

Attendance at religious services is not the only measure of religious behavior frequently employed by survey researchers. Other behavioral indicators of religious commitment include questions regarding private devotionalism, such as praying or scripture reading.

Table III Questions to Measure Attendance at Religious Services[a]

Now I'm going to ask you about things you did during the last seven days. I'm only interested in what you did during the last seven days. From (last day of week) to today did you . . .

a. Go to see a doctor or receive medical treatment at a clinic or hospital?
b. Have a meal (breakfast, lunch, or dinner) at a restaurant (including fast food places and take-out)?
c. Go to a movie theater to see a film?
d. Attend religious services?

On what day or days did you attend religious services during the last seven days? (Probe: Did you attend religious services on any other days during the last seven days? (Ask until respondent says, "No").

During the last seven days did you do the following:

a. Attend a regular, weekly worship service at a church/synagogue (e.g., mass or Sunday morning services). Don't include watching a service on TV or listening to one on the radio.
b. Watch a religious program on television or listen to a religious program on the radio
c. Attend some other type of religious event or meeting (e.g., prayer breakfast, Bible study group, choir practices, church sponsored lectures, adult fellowship meetings)?

[a] From 1996 General Social Survey; see also Smith (1998).

Scholars of religion generally also treat questions that determine the extent to which a respondent's life is guided by religion as a measure of commitment.

The difficulty in employing questions such as these is that the normative significance attached to such activities differs across religious traditions. In 1996, Leege stressed that the standard measures of religious commitment in surveys such as the ANES have a Protestant bias. For example, many Protestant denominations, especially within the evangelical tradition, place great emphasis on reading the Bible frequently. This is not the case among Roman Catholics. Therefore, whether a Catholic reads the Bible frequently is not as good an indication of commitment to the precepts of Catholicism because it is an indication of commitment to the norms of the evangelical tradition. Perhaps the point can be made most clearly by thinking about an obviously inappropriate question, such as asking Christians whether they adhere to kosher dietary restrictions.

The solution, it would seem, requires measures of religious commitment that are specific to each religious tradition—a difficult task given America's denominational diversity. A clever way around this problem was proposed in 2001 by Mockabee, Monson, and Grant. They detailed a method to weight individuals' participation in a religious activity by its normative significance within that individual's tradition. This is done by weighting the behavior by the proportion of each respondent's coreligionists who report participating in that activity. More specifically:

$$C_i = (\Sigma\delta_{kt})^{-1}(\Sigma\delta_{kt}B_k),$$

where C_i is the commitment score for each individual I, B_k measures how frequently the individual participates in each of k activities, and δ_{kt} is the proportion of people in the same religious tradition who report participating in that activity. Upon employing these weights, Mockabee, Monson, and Grant found that, as expected, the average level of religious commitment rises among Catholics. However, they also concluded that the weights have a limited substantive impact in models predicting political behavior, such as the 1996 presidential vote. They did not test the difference their weighting scheme has when modeling other types of dependent variables, but it seems likely that the weights would result in moderate corrections for bias.

The bottom line, therefore, is that although tradition-specific measures of religious commitment are probably best, weighting the standard behavioral measures by the norms of each tradition appears to be a workable solution. And depending on the dependent variable being modeled, the weights themselves may not even have all that much of an impact, providing reassurance that the bias in these measures is within reasonable bounds. As the United States becomes more and more religiously diverse, we should expect the bias in standard behavioral measures of religious commitment to increase and a weighting system such as that proposed by Mockabee, Monson, and Grant to have increasing analytical utility. However, weights are only a post hoc correction for invalid measures. We hope that scholars of religion will continue to develop better measures of religious commitment that apply to diverse religious traditions.

Conclusion

For many researchers, particularly those intending to use religious affiliation and commitment as control variables only, the examples discussed here will provide a starting place for constructing valid measures of religious affiliation and commitment. Some studies, however, will explore new angles on religion and thus will require adapting these methods of measurement or even the development of new methods. Such efforts can also be guided by principles distilled from the research cited here.

Whether the researcher is developing new methods of measurement or not, the study of religion's impact on social behavior should be guided by one paramount principle: the specifics of religious affiliation and commitment aside, the most important thing to learn from the literature on religion is the need for understanding the nuances of the American religious environment. All scholars studying religion should be prepared to learn to navigate through America's religious labyrinth. Given the myriad ways in which religion affects social attitudes and behavior, this is an investment that has the potential to pay a big dividend.

See Also the Following Articles

Survey Design • Surveys

Further Reading

A Symposium on Church Attendance in the United States. (1998). *Am. Sociol. Rev.* **32,** 112–145.

American Association for Public Opinion Research. (2000). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* AAPOR, Ann Arbor, MI.

Hadaway, C. K., Marler, P. L., and Chaves, M. (1993). What the polls don't show: A closer look at U.S. church attendance. *Am. Sociol. Rev.* **58**(6), 741–752.

Harris, F. C. (1999). *Something Within: Religion in African-American Political Activism.* Oxford University Press, New York.

Hunter, J. D. (1983). *American Evangelicalism: Conservative Religion and the Quandary of Modernity.* Rutgers University Press, New Brunswick, NJ.

Kellstedt, L. A., and Green, J. C. (1993). Knowing God's many people: Denominational preference and political

behavior. In *Rediscovering the Religious Factor in American Politics* (D. C. Leege and L. Kellstedt, eds.), pp. 53–71. M. E. Sharpe, Armonk, NY.

Kellstedt, L. A., Green, J. C., Guth, J. L., and Smidt, C. E. (1996). Grasping the essentials: The social embodiment of religion and political behavior. In *Religion and the Culture Wars* (J. C. Green, J. L. Guth, C. E. Smidt, and L. A. Kellstedt, eds.), pp. 174–192. Rowman and Littlefield, Lanham, MD.

Layman, G. (2001). *The Great Divide: Religious and Cultural Conflict in American Party Politics.* Columbia University Press, New York.

Leege, D. C. (1996). Religiosity measures in the national election studies: A guide to their use, 2. *Votes Opinions* **2,** 6–9.

Leege, D. C., and Kellstedt, L. A. (eds.) (1993). *Rediscovering the Religious Factor in American Politics*. M. E. Sharpe, Armonk, NY.

Lincoln, C. E., and Mamiya, L. H. (1990). *The Black Church in the African American Experience.* Duke University Press, Durham, NC.

Marler, P. L., and Hadaway, C. K. (1999). Testing the attendance gap in a conservative church. *Sociol. Religion* **60**(2), 175–186.

Melton, J. G. (ed.) (1993). *The Encyclopedia of American Religions*. 4th Ed. Gale Research, Detroit, MI.

Mockabee, S. T., Monson, J. Q., and Grant, J. T. (2001). Measuring religious commitment among Catholics and Protestants: A new approach. *J. Sci. Stud. Religion* **40**(4), 675–690.

Princeton Religious Research Center. (1994). Do that many people really attend worship services? *PPRC Emerging Trends* **16**(5), 1–3.

Roof, W. C., and McKinney, W. (1987). *American Mainline Religion: Its Changing Shape and Future.* Rutgers University Press, New Brunswick, NJ.

Sherkat, D. E. (1999). Tracking the "other": Dynamics and composition of "other" religions in the General Social Survey, 1973–1996. *J. Sci. Stud. Religion* **38**(4), 551–560.

Smith, C. (1998). *American Evangelicalism: Embattled and Thriving.* University of Chicago Press, Chicago, IL.

Smith, T. W. (1990). Classifying Protestant denominations. *Rev. Religious Res.* **31,** 225–245.

Smith, Tom W. (1998). A review of church attendance measures. *Am. Sociol. Rev.* **32**(1), 131–136.

Steensland, B., Park, J. Z., Regnerus, M. D., Robinson, L. D., Wilcox, W. B., and Woodberry, R. D. (2000). The measure of American religion: Toward improving the state of the art. *Soc. Forces* **79**(1), 291–318.

Wald, K. D., and Smidt, C. E. (1993). Measurement strategies in the study of religion and politics. In *Rediscovering the Religious Factor in American Politics* (D. C. Leege and L. A. Kellstedt, eds.), pp. 26–49. M. E. Sharpe, Armonk, NY.

Woodberry, R. D. (1997). *The missing fifty percent: Accounting for the gap between survey estimates and head-counts of church attendance.* Department of Sociology, University of Notre Dame, Notre Dame, IN.

Woodberry, Robert D. (1998). When surveys lie and people tell the truth: How surveys oversample church attenders. *Am. Sociol. Rev.* **63**(1), 119–122.

Wuthnow, R. (1988). *The Restructuring of American Religion.* Princeton University Press, Princeton, NJ.

Remote Sensing

John A. Kupfer
University of Arizona, Tucson, Arizona, USA

Charles W. Emerson
Western Michigan University, Kalamazoo, Michigan, USA

Glossary

electromagnetic radiation (EMR) A combination of oscillating electric and magnetic fields propagating through space and carrying energy from one place to another; electromagentic radiation is generally classified by its wavelength (e.g., visible light, 0.4–0.7 μm).
multispectral Imagery consisting of multiple bands, each of which was recorded simultaneously at different wavelengths.
pixel A picture element in a digitized image that represents the area corresponding to a given measurement of electromagnetic radiation.
platform The vehicle or unit from which a remote sensing system collects data (e.g., a satellite or airplane).
reflectance Ratio of the amount of energy incident on an object to the amount reflected.
sensor A device that converts electromagnetic radiation into an electrical signal that can be recorded, displayed, and analyzed.

Remote sensing has been defined by the American Society for Photogrammetry and Remote Sensing as "the measurement or acquisition of information of some property of an object or phenomenon by a recording device that is not in physical ... contact with the object or phenomenon under study." In most cases, the information being measured is electromagnetic radiation (EMR), with the sun serving as the energy source in passive measurements of reflected radiation, the earth as the source in measurements of emitted thermal radiation, and an energy-emitting device as the source in the case of active remote sensing systems such as radar. Measurements of the interactions of EMR with the earth's surface are recorded by a sensor that is typically mounted on an airplane or satellite platform. Data received by the sensor are recorded and written to a digital data file composed of a two-dimensional grid of spatial objects called pixels, each of which represents a specified area on the earth's surface. Additional dimensions of information are provided by measuring radiation in a range of spectral bands (e.g., energy reflected from objects in different subsets of wavelength) and by examining characteristics of the same location at various times.

The Remote Sensing Process

Physical Principles

The amount of energy that passes onto, off, or through a surface per unit time is called *radiant flux*, and the characteristics of the radiant flux and what happens to it as it interacts with the earth's surface are a fundamental focus of remote sensing research. Because different objects have different spectral properties (e.g., the amount of energy reflected in various wavelengths), it is possible to infer characteristics of surface features by monitoring the nature of radiant flux. A sensor's ability to detect surface properties and, consequently, a scientist's ability to extract meaningful information from remotely sensed imagery are a function of (i) the number and dimension of specific wavelength intervals to which a given remote sensing instrument is sensitive (spectral resolution), (ii) the amount of area on the earth's surface that is represented by a pixel (spatial resolution), (iii) the frequency with which a sensor records imagery for a given

area (temporal resolution), and (iv) the ability of an imaging system to discriminate differences in radiant flux—that is, the range of digital values that is recorded (radiometric resolution).

Sensor Technology

When a sensor element is exposed to electromagnetic radiation (EMR), an electrical charge or change in resistance is produced and converted to a digital number ranging from 64 steps for early satellites to 4096 steps for some modern sensors. Some passive digital remote sensors are panchromatic in that they cover a relatively wide range of EMR wavelengths in a single band of grayscale values. The result is comparable to a black-and-white photograph. Multispectral sensors, on the other hand, divide the electromagnetic spectrum into a range of individual wavelength bands. The Landsat Multispectral Scanner (MSS), for example, records EMR in four wavelength bands: 0.5–0.6 μm (green), 0.6–0.7 μm (red), 0.7–0.8 μm (near infrared), and 0.8–1.1 μm (near infrared). Hyperspectral scanners go one step further, utilizing dozens or even hundreds of narrow, contiguous bands, thus approaching a continuous spectral response curve.

A trade-off that must be made when going from a single panchromatic band to a larger set of narrower multispectral or hyperspectral bands is the reduction in the amount of reflected radiation recorded per band. When the range of wavelengths reaching the sensing elements used in satellite and aerial scanners is restricted, they have to view a relatively larger patch of the earth's surface in order to yield a measurable electrical response. The newer earth resources satellites therefore generally include a finer spatial resolution panchromatic "sharpening" band in addition to multispectral bands having coarser spatial resolutions. The increased radiometric depth of some sensors, along with the need for higher spatial or spectral resolution, has led to huge increases in file sizes that continue to push the performance envelope of modern computer workstations.

Clear skies are a basic condition for passive remote sensing of the earth's surface. However, in many areas of the world, such as the tropics, this requirement is rarely met, limiting the utility of optical and thermal sensors in these areas. Active radar sensors overcome this problem by providing their own source of microwave EMR, which is not significantly absorbed or reflected by light rain or snow, clouds, or smoke particles. Side-looking radar or side-looking airborne radar systems have fixed antennas that send out pulses of radiation as the platform moves along. The interval between the time a pulse is sent and received is converted to a distance from the platform's path, and the intensity of the received pulse is converted to a grayscale value. Radar images appear very different from optical images since microwave radiation interacts differently with Earth surface materials. Having the source of radiation at the sensor also changes the geometry of radar images compared to optical images. Synthetic aperture radar systems electronically simulate a series of antennas as the sensor moves along track. Mathematical combination of larger numbers of these simulated antennas at increasing distances leads to a constant spatial resolution in the across-track direction.

LIDAR (light detection and ranging) systems use pulses of laser light directed toward the ground to measure the platform–ground distance. This technology, when combined with accurate measurements of the sensor location and attitude (using the global positioning system and inertial navigation systems in the case of aircraft-borne sensors), facilitates the creation of detailed digital elevation models, digital representations of the earth's surface composed of regularly spaced point locations with an elevation attribute. Small-footprint LIDAR systems use a highly focused beam that is approximately 0.1 m in diameter, depending on the aircraft altitude. Multiple pulses are collected for the same area on the ground to distinguish the height of the vegetation canopy from the ground surface. Large footprint systems that survey ground patches with a diameter of 5–15 m are currently under development. These systems use the waveform of the reflected light to derive a vertical profile of the vegetation canopy.

Image Processing

After digital data are recorded by a sensor and stored, they typically undergo digital image processing, including a preprocessing phase and a postprocessing phase. Preprocessing of the data involves (i) radiometric corrections, which correct the data to avoid error or distortions due to internal sensor errors, sun angle, topography, and various atmospheric effects, and (ii) geometric corrections, which include correcting for geometric distortions due to sensor–Earth geometry variations and converting the data to real-world coordinates (Table I). Postprocessing involves (i) image enhancement, which may be used to improve the appearance of the imagery to assist in interpretation and analysis of the data, and (ii) image transformations, which can be used to combine or transform the original bands into new images that better display or highlight certain features in the image.

Postprocessing may also include image classification, which entails grouping pixels into nominal classes using combinations of the digital numbers that represent reflectance in each band. The result is a classified image (e.g., of land use or vegetation type). Unsupervised classification methods are inductive approaches that statistically cluster pixels into a predefined number of classes based on their spectral properties. The meaning of the classes is subsequently established by relating the pixels in

Table I Summary of Selected Image Enhancement and Transformation Techniques

Technique	*Example*	*Purpose*
Geometric correction	Deskewing	Remove systematic distortions inherent in sensor
	Registration	Convert image to ground coordinates
Radiometric enhancement	Sun elevation correction	Normalize multidate images for illumination differences
	Contrast stretch	Expand range of grayscale or color values to enhance viewing
Spectral enhancement	Band ratioing	Reduce effects of shadows
	Principal components analysis	Reduce dimensionality (No. of bands)
Spatial enhancement	Low-pass filter	Smooth image to enhance broad trends in brightness
	High-pass filter	Emphasize details
Change detection	Image differencing	Highlight differences between two images obtained on different dates

each class to corresponding features on the earth's surface (e.g., "forest" pixels). Supervised classification methods, in contrast, utilize validated training sites that represent the desired output classes. The spectral signatures of the training sites are then used by the classification algorithm to classify other pixels that have similar spectral responses.

Remote Sensing History and Prominent Satellite Platforms

Photography and Satellite Imagery

In 1824, Joseph Nicéphore Niépce obtained the first fixed image of a landscape. The first known aerial photograph was taken from a balloon by Gaspard Felix Tournachon (a.k.a. "Nadar"), a magician who photographed Bievre, France, in 1858. The practical benefits of aerial photography were realized early on, with balloons, kites, and even homing pigeons serving as the camera platform until the invention of the airplane in 1903. More than 1 million aerial reconnaissance photographs were taken during World War I, and after the war civilian development of aerial photography interpretation continued, particularly in the area of resource assessment and management.

In the 1930s, panchromatic color film was developed, which approximated the human eye's response to visible light. The spectral range was extended with the development of color infrared film in World War II, which was sensitive to EMR in the near-infrared range. Radar, another World War II invention, was turned toward the ground after the war, resulting in the first side-looking airborne radar images. This technology has benefited from the rapid advances in electronics so that modern synthetic aperture radar systems can provide high-resolution imagery from space or aircraft platforms. The Cold War led to the development of a number of space-based military reconnaissance programs, such as CORONA, ARGON, and LANYARD. More than 800,000 of these high-resolution satellite images taken between 1959 and 1972 were recently declassified and offered for sale by the U.S. Geological Survey.

Civilian spin-offs of space technology included meteorological satellites such as TIROS (Television Infrared Observation Satellite), which became operational in 1960. The use of satellites for meteorological and related applications continues today with the GOES (Geostationary Operational Environmental Satellite) satellites, which are the data source for the satellite loops seen on daily weather reports. These satellites, which have one visible and four thermal bands with resolutions from 1 to 8 km, occupy an orbit that has a 24-hour period matching the earth's rotation: They are thus geostationary and continuously remain above the same location on the equator. The polar-orbiting National Oceanic and Atmospheric Administration series of satellites carrying the Advanced Very High Resolution Radiometer sensor have a relatively coarse 1.1-km resolution but provide twice-daily information that is used for a range of applications. ORBIMAGE also operates OrbView-1, which provides 10-km resolution imagery of severe weather and lightning, and OrbView-2, which provides 1.1-km imagery of the entire globe for ocean monitoring and agricultural applications.

The success of space-based military and meteorological imaging in the 1960s led to a joint NASA/U.S. Department of the Interior feasibility study on Earth Resources Technology Satellites (ERTS). Six satellites were planned, with ERTS-1 being launched in July 1972. The program name was changed to Landsat with the launch of the second satellite in 1975. Landsats 1–3 had a Return Beam Vidicon sensor in addition to a four-band Multispectral Scanner (MSS), with the latter becoming the primary sensor. Landsat-4 (1982) and -5 (1984) carried a new, higher resolution sensor called the Thematic Mapper (TM), in addition to a slightly modified MSS, which was included to provide continuity with earlier satellites. In comparison to the four-band, 80-m nominal spatial resolution (pixel size) of MSS imagery, TM images had six bands with 30-m nominal spatial

resolution in the visible, near-, and mid-infrared wavelengths and a 120-m thermal infrared band that detected emitted terrestrial radiation. Landsat-6 failed to achieve orbit in 1993, and Landsat-7 was launched in 1999, carrying the Enhanced Thematic Mapper Plus (ETM+) sensor. This sensor provides continuity with Landsat-4 and -5 because it has the same six reflective bands as the TM sensor plus a higher resolution (60-m) thermal infrared band and a 15-m panchromatic band. In May 2003, a scan line corrector in the ETM+ sensor failed, and the system has been operating in a reduced capacity mode since July 2003.

A number of other countries and organizations, including India, the Soviet Union/Russia, Japan, China, Brazil, Canada, and the European Space Agency, have had imaging satellites in the past or are planning future launches. The French government, with the participation of Belgium and Sweden, developed the Système Pour l'Observation de la Terre (SPOT) program in 1978. SPOT-1, launched in 1986, employed a linear array of high-resolution visible (HRV) sensors (not unlike a desktop flatbed scanner) with pointable optics that allowed stereoscopic viewing of areas imaged from adjacent tracks. The HRV sensor operated in a 10-m resolution panchromatic mode and a 20-m resolution multispectral mode with three bands in the green, red, and near-infrared ranges of the electromagnetic spectrum. The later HRVIR (high-resolution visible and infrared) sensor carried on SPOT-4 added a 20-m mid-infrared band to aid in vegetation monitoring and soil moisture mapping and a "vegetation" instrument, which provided 1-km resolution imaging of the whole globe daily. The vegetation instrument is also carried aboard SPOT-5 (launched in May 2002), along with twin high-resolution geometric scanners with a 2.5- or 5-m panchromatic resolution and a 10-m multispectral resolution (20-m mid-infrared). The high-resolution stereographic sensor collects panchromatic imagery in a fore-and-aft fashion to allow preparation of digital elevation models at 10 m resolution.

In 1999, Space Imaging Corporation launched IKONOS, a high-resolution (1-m panchromatic, 4-m multispectral) satellite, as a purely commercial venture. This was followed by the launch of EROS-A in 2000, DigitalGlobe's Quickbird satellite in 2001, and ORBIMAGE's OrbView-3 in June 2003. All of these satellites fill a niche for high-resolution space imagery that is not met by the fleet of governmental satellites (with the possible exception of SPOT-5). In some sense, they can replace mid- to high-altitude aerial photography, with the added benefit of regular repeat cycles.

Sources of Imagery

Current and historical remotely sensed images are available from a number of private and governmental sources. The Earth Observing System (EOS), the primary component of NASA's Earth Science Enterprise, seeks to provide data, modeling capabilities and eventually an understanding of human impacts on the environment. As of late 2003, two satellites, Terra and Aqua, had been launched as part of this program, with the latter carrying five sensors: ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer), MODIS (Moderate-Resolution Imaging Spectroradiometer), CERES (Clouds and the Earth's Radiant Energy System), MISR (Multiangle Imaging Spectro-Radiometer), and MOPPITT (Measurements of Pollution in the Troposphere). The EROS (Earth Resources Observation Systems) Data Center, a unit of the U.S. Geological Survey, is a major source of aerial photographs, satellite imagery, and other mapping products. Local and state governments often make satellite images, photographs, and other products of aerial survey firms available at low or no cost via the Internet. Private firms, such as Space Imaging, Digital Globe, and ORBIMAGE, also distribute their own products.

Remote Sensing Applications

Remote Sensing of Terrestrial Features

Many natural and anthropogenic features on the earth's surface can be identified, mapped, and studied on the basis of their spectral properties, the proportion or amount of energy reflected, absorbed, transmitted, or emitted by an object at various wavelengths. An object's spectral properties depend on its composition and condition, so the effective utilization of remote sensing data requires an understanding of the properties of the features under investigation and the factors that influence these characteristics. Although different features may be indistinguishable in one spectral band, they may be very different in others, underscoring the value of examining properties in a range of wavelengths. A graph of the spectral reflectance of an object as a function of wavelength is called a spectral reflectance curve (Fig. 1), and the configuration of the curve provides insights into the nature and characteristics of an object and influences the choice of wavelengths in which remote sensing data are acquired for a particular application.

Spectrally, vegetation can be distinguished from inorganic materials by its high absorption of red and blue light, moderate reflectance of green light, and high reflectance of near-infrared energy. Pigments in a typical green plant, including chlorophyll *a* (maximum absorption, 0.44 and 0.67 μm), chlorophyll *b* (maximum absorption, 0.49 and 0.65 μm), and β-carotene (maximum absorption, 0.45 μm), are responsible for high absorption in the red and blue portions of the visible light spectrum. Near infrared reflectance (0.7−1.2 μm) tends to be high for

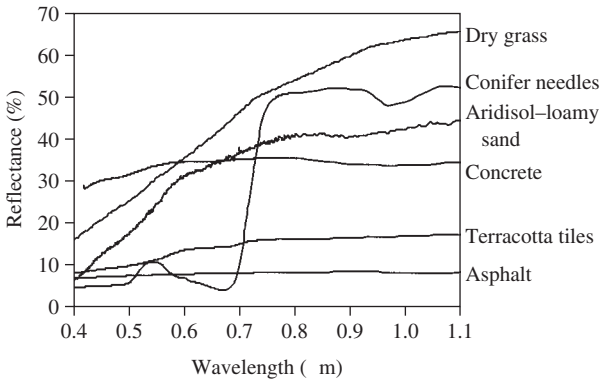


Figure 1 Spectral reflectance curves in the visible and near infrared wavelengths for dry grass, green conifer needles, a light yellowish brown loamy sand—aridisol soil, concrete, terracotta roofing tiles, and asphalt. Data reproduced from the ASTER Spectral Library through the courtesy of the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California.

healthy green vegetation due to internal scattering of EMR at the cell wall–air interfaces within the leaf. At longer wavelengths, water within the leaves (especially in the spongy mesophyll) is a strong absorber of middle-infrared wavelengths; thus, as the water content of leaves increases, reflection in these bands (especially between 1.5–1.8 and 2.1–2.3 μm) decreases.

Different vegetation types (e.g., grassland, deciduous forest, and desert scrub) can often be distinguished in images due to differences in leaf size and shape, plant morphology, water content, and vegetation density, enabling the creation of vegetation maps that are useful in fields such as natural resource management, forestry, and ecology. When coupled with faunal surveys, vegetation maps can be used in applications such as endangered species habitat mapping and timber harvest planning. Information on temporal dynamics of vegetation can be gained by utilizing multiple image dates, taken either at different times over a growing season or over multiple years. Examples of research using multitemporal analyses include the documentation of changes in land use and land cover related to human activities (e.g., agricultural conversion and urban growth) and the delineation of spatial patterns and effects of disturbances (e.g., clear-cutting, wildfires, and insect outbreaks). Because many plant species undergo relatively unique seasonal (phenological) changes, multitemporal remote sensing can also play a role in monitoring crop development and health and in projecting agricultural yields.

Further insight into vegetation characteristics can be gained by applying vegetation indices derived from the original data (e.g., the Normalized Difference Vegetation Index). Such indices are dimensionless, radiometric measures that function as indicators of relative abundance or activity of green vegetation, including leaf area index, percentage green cover, chlorophyll content, green biomass, or photosynthetically active radiation. For instance, there is a well-known inverse relationship between spectral response in the visible spectrum and plant biomass, allowing scientists to make regional and global estimates of biomass and productivity that are central in the study of global change effects.

In addition to providing a means for examining vegetation characteristics, remote sensing can be used to identify, categorize, and map anthropogenic features. Maps designed to show both natural and human-created features, known as land use and land cover maps, may be of tremendous value to groups as diverse as urban planners, economists, transportation managers, real estate developers, demographers, natural resource managers, and conservationists. The premise of land use and land cover mapping is the same as that for vegetation studies: different urban materials, such as concrete, blacktop asphalt, and asphalt shingles, have unique spectral properties that distinguish them from other such materials and vegetation. Two other similarities to remote sensing of vegetation are that different portions of the electromagnetic spectrum are better suited for extracting different types of information (e.g., estimating building perimeter and area vs identifying different land use types), and there is a tradeoff between the detail of the information needed and the spatial resolution of the data needed to capture such features. In general, because of the fine scale of many objects of interest in urban and suburban landscapes, it is frequently important to have data with extremely high spatial resolution (typically <5 m and often <1–2 m), limiting the utility of some of the primary sensor platforms more commonly used in Earth resource applications.

Remote sensing can also play an important role in the study of soils, minerals, geomorphology, and topography, especially when vegetation is sparse or absent. Several factors influence soil reflectance in remotely sensed images, including mineral composition (e.g., iron oxide content), organic matter content, soil texture, moisture content, and surface roughness. When the effects of these factors on the spectral properties of soils are understood, remote sensing can be used in the identification, inventory, and mapping of soil types and properties that can be used to inform decisions about crop nutrition and herbicide usage, short-term stresses (e.g., drought), and susceptibility of soil to erosion. Observed differences in soil texture or moisture can also be used in archeological applications, for instance, to detect the impact of humans on the soil that may be related to past land use practices.

Similarly, general geologic information, such as chemical composition of rocks and minerals on the earth's

surface, lithology, geologic structure, drainage patterns, and landform characteristics, can be extracted from remotely sensed data. This information can be valuable in the production of geologic maps that, when coupled with information collected in the field, can provide surficial clues to the locations of subsurface deposits of ore minerals, oil and gas, and groundwater. Geological information developed from imagery is also valuable in hazards planning and civil engineering applications.

Remote Sensing of Aquatic and Atmospheric Phenomena

When measuring the spectral properties of water bodies, total radiance recorded by a sensor is a function of EMR received from four sources: (i) radiation that never reaches the water surface (atmospheric noise or path radiance); (ii) radiation that reaches the water surface but is reflected in the top few millimeters; (iii) radiation that penetrates the air–water interface, interacts with the water and organic/inorganic constituents, and then exits the water column without contacting the bottom (subsurface volumetric radiance); and (iv) radiation that penetrates the water, reaches the bottom of the water body, and is propagated back through and exits the water column (bottom reflectance). The goal of most aquatic remote sensing is to extract or isolate the radiance of interest from all of the other components.

Remote sensing has been used to examine and map a wide range of hydrologic variables. Because water bodies absorb much more of the incident radiant flux in the near- and mid-infrared wavelengths than do land surface features, remote sensing can be used to delineate the land–water interface and monitor the surface extent of water bodies. Multitemporal measurements can be used to document flood timing and extent, fluctuations in lake size at a variety of timescales (seasonal, annual, and decadal), drought effects, and changes in wetland area. Recent research has similarly focused on ways by which remote sensing can be used to measure the extent, cover, and volume of snowpack, glaciers, and ice shelves, providing information relevant to water management issues and global environmental change studies. Data from thermal infrared wavelengths is pivotal in monitoring ocean-wide trends in sea surface temperatures that are indicators of El Niño/Southern Oscillation and La Niña events that greatly affect regional climate patterns and thus bear on water management issues.

Spectral properties of pure water differ from those of water with suspended sediment and minerals as a function of both the quantity and characteristics (e.g., particle size and absorption) of the material. By collecting *in situ* measurements of suspended material concentrations and relating these measures to remotely sensed data, it is possible to derive estimates of the type, amount, and spatial distribution of suspended materials in inland and near-shore waters. This allows scientists to examine soil erosion, reservoir and harbor sedimentation, and water quality. Estimates of suspended sediments are also valuable because such material can impede the transmission of solar radiation into the water column and influence photosynthesis of submerged aquatic vegetation and phytoplankton.

In addition to providing information on inorganic constituents of water bodies, remote sensing can be used to examine organic components of aquatic ecosystems. Of particular interest has been the ability of remote sensing systems to detect and estimate concentrations of phytoplankton, small single-celled plants that contain chlorophyll and other photosynthetically active pigments. Because different types of phytoplankton contain varying concentrations of chlorophyll, it is possible to estimate the amount and general type of phytoplankton, which can provide information on the health and chemistry of the water body. For reasons similar to those discussed previously for terrestrial vegetation, numerous studies have documented a relationship between selected spectral bands and the concentration of aquatic chlorophyll. Near-surface estimates of chlorophyll concentration can thus be derived to estimate biomass and productivity, which in turn can be used to better understand the dynamics of ocean and coastal currents, assess the ocean's role in the global carbon cycle, and clarify marine influences on global climate change. In recent years, remote sensing has also been used to monitor declines in coral reef health related to water pollution or environmental variability.

Finally, remote sensing is increasingly used to examine a range of atmospheric phenomena. Various aspects of precipitation (location, intensity, and amount) can be measured or estimated directly, through the use of active microwave sensors such as the Nexrad radar systems operated by the National Weather Service in the United States, or indirectly, through measures of cloud reflectance, cloud-top temperatures, and the presence of frozen precipitation aloft. Although ground-based radar provides real-time information on precipitation with high spatial resolution, the spatial coverage of such systems is localized. Conversely, meteorological satellites have global coverage, but the information provided is at a coarser timescale and provides only an estimate of local conditions such as actual rainfall amounts. Nonetheless, the latter have proved valuable in estimating global rainfall patterns and monitoring drought development. Lastly, satellite imagery is used to examine cloud cover and atmospheric moisture in the mid- and upper level of the troposphere and, recently, to gather information on a broader range of atmospheric conditions, including atmospheric chemistry (e.g., aerosol concentrations)

and temperatures in the troposphere and lower stratosphere.

Conclusions

Although remote sensing techniques have primarily been viewed as a means for gathering data that are then interpreted by the user, they are increasingly serving other roles in scientific and applied research. Remotely sensed data on natural and anthropogenic features such as vegetation cover, land use, topography, and hydrography now serve as input to a range of simulation models, including hydrologic, climatic, ecological, and economic models. Classified images of land use and land cover are combined with surveys of demographic and socioeconomic variables to develop models that allow scientists to better understand processes such as deforestation and land use conversion. Remotely sensed imagery and the products derived from analysis of imagery are also important sources of data for geographic information systems (GIS). In fact, most comprehensive image analysis software packages now include GIS functions for change detection overlays, local spatial analysis techniques, conversions between raster (i.e., pixel-based grids) and vector (i.e., points, lines, and polygons defined and displayed on the basis of two-dimensional Cartesian coordinate pairs) data structures, and other not strictly image-related processes. GIS software packages by necessity work with raster data and images in a number of formats, and they increasingly include analysis functions that were previously only found in specialized image analysis packages.

Compared to field-based sampling, remote sensing cannot provide measures of human and environmental phenomena, such as water quality, vegetation composition, soil properties, or plant health, with the same amount of detail. However, when coupled with field surveying, remote sensing offers the ability to view and map large areas of the earth's surface at multiple times and to obtain information for areas that would otherwise be difficult or impossible to sample due to physical or financial constraints. Remote sensing has thus become a valuable tool in research and applications in a wide range of disciplines, such as engineering, geology, geography, urban planning, forestry, and agriculture. Furthermore, the Internet has increased the availability and dissemination of remote sensing products, and decreasing costs coupled with continuous improvements in spatial, spectral, radiometric, and temporal resolutions are making remote sensed data accessible to a broader range of end users and expanding the role of remote sensing in society.

See Also the Following Articles

Geographic Information Systems • Spatial Databases • Spatial Pattern Analysis

Further Reading

Avery, T. E., and Berlin, G. L. (2004). *Fundamentals of Remote Sensing and Airphoto Interpretation*, 6th Ed. Prentice Hall, Upper Saddle River, NJ.

Jensen, J. R. (1996). *Introductory Digital Image Processing*, 2nd Ed. Prentice Hall, Upper Saddle River, NJ.

Jensen, J. R. (2000). *Remote Sensing of the Environment: An Earth Resource Perspective.* Prentice Hall, Upper Saddle River, NJ.

Lillesand, T. M., Kiefer, R. W., and Chipman, J. W. (2004). *Remote Sensing and Image Interpretation*, 5th Ed. Wiley, New York.

Russ, J. C. (2002). *The Image Processing Handbook*, 4th Ed. CRC Press, Boca Raton, FL.

Sabins, F. F. (1997). *Remote Sensing Principles and Interpretation*, 3rd Ed. Freeman, New York.

Sample, V. A. (ed.) (1994). *Remote Sensing and GIS in Ecosystem Management*. Island Press, Washington, DC.

Schowengerdt, R. A. (1997). *Remote Sensing: Models and Methods for Image Processing.* Academic Press, San Diego.

Research Designs

Gerardo L. Munck
University of Southern California, Los Angeles, California, USA

Jay Verkuilen
University of Illinois, Champaign-Urbana, Champaign, Illinois, USA

Glossary

case study A study in which one unit is analyzed, typically in an intensive manner that is attentive to time and process. Although, strictly speaking, in a case study the $N = 1$, frequently the effective number of observations is considerably higher.

cross-sectional design A study in which observations on a variable or multiple variables are collected across units at the same point in time.

experimental design A study in which the treatment is consciously manipulated by the researcher and in which units are randomly assigned to treatment and control groups. To distinguish experimental from quasi-experimental designs, the former are sometimes called randomized experiments.

external validity Concept originally introduced by Donald T. Campbell to refer to the generalizability of the finding of a causal relationship between variables beyond the domain of the actual units, spatial and temporal setting, and specific treatments that are examined.

internal validity Concept originally introduced by Donald T. Campbell to refer to a causal relationship between variables in the actual units, spatial and temporal setting, and specific treatments that are examined.

large *N* study A study in which observations are made across a large number of units. Such studies, however, vary significantly in terms of their *N*, with typical cross-national studies in the field of comparative politics and international relations oscillating between 30 and 100 and those using opinion surveys reaching into the thousands.

longitudinal design A study in which multiple observations on variables are collected across time for the same unit. Also known as time series design.

observational studies Nonexperimental studies, also called correlational studies, in which the treatment is not consciously manipulated by the researcher. Rather, researchers simply record the values of variables as they naturally occur. These studies include natural experiments in which particularly sharp and obvious changes in the value of a variable are held to offer an analogy to the introduction of a treatment.

pooled time series, cross-sectional design A design that combines a cross-sectional design and a longitudinal design; includes panel studies and repeated measures design.

quasi-experimental design Design in which the treatment is consciously manipulated by the researcher, as in an experiment, but in which, unlike in experiments, units are not randomly assigned to treatment and control groups.

research design A key aspect of the research process that revolves around the direct impact on the prospects of causal inference of four core questions: How is the value of the independent variable(s) assigned? How are units selected? How many units are selected? and How are comparisons organized (i.e., whether temporally and/or spatially)? In addition, research designs can be evaluated in terms of their indirect impact on causal inference in light of their requirements and contributions vis-à-vis theory and data.

small *N* study A study in which observations are made across a small number of units. Typically, each unit is treated as a case study so that multiple observations on each unit are made.

The methodology of research design hinges on the choices made with regard to four core questions: How is the value of the independent variable(s) assigned? How are units

selected? How many units are selected? and How are comparisons organized (i.e., whether temporally and/or spatially)? These choices can be assessed in terms of their direct impact but also their indirect impact—due their requirements and contributions vis-à-vis theory and data—on the prospects of making causal inferences. Three research traditions—experimental and quasi-experimental, quantitative, and qualitative—represent distinct responses to these methodological choices and each has important strengths but also significant weaknesses. Thus, the need for choices about research design to be explicitly addressed and justified, and the need to actively construct bridges across research traditions, is emphasized.

Introduction: Goals, Problems, and Options

The pioneering work on research design by Donald T. Campbell and associates has made such a major contribution that it is difficult to think about research design without, in one way or another, drawing on their insightful discussions. They provide a valuable template for thinking about research design, helpfully framing the discussion in terms of the basic goal of validity, a range of problems or threats to validity, and a set of options or design choices that can be pursued as a way to guard against these threats. At the same time, their work displays some limitations and biases and ultimately fails to offer a clear, encompassing, and balanced understanding of the challenges involved in research design. This assessment and comparison of research designs thus adopts Campbell *et al.*'s basic template and some of their key ideas about the goals, problems, and options of research design but also parts company with them in some significant ways.

First, Campbell *et al.*'s discussion of the goals of research design in terms of the concept of validity is both somewhat confusing and biased. Initially, Campbell introduced the concepts of internal and external validity, which disaggregated the ultimate goal of research design—to increase the prospects of making causal inferences—and aptly distinguished different problems of causal inference that are affected and potentially solved by different design choices. Over time, however, these two concepts have been awkwardly relabeled and, more important, defined in different ways in different texts. In addition, two other types of validity—statistical conclusion validity and construct validity—that pertain in part to research design but spill over into questions of data analysis and measurement, respectively, have been introduced into the discussion, further complicating matters. Another problem with this typology of validity is that it is somewhat biased. Thus, Lee Cronbach has argued that Campbell *et al.* gave undue primacy to internal over external validity and did not recognize the importance of generalizability. This is a critique that they have acknowledged and sought to address in their latest statement. However, they still do not fully incorporate the difficulties of generalizing on the basis of experimental and quasi-experimental designs in their overall assessment of the strengths and weaknesses of design options.

To avoid these problems, we only consider internal and external validity to be the core goals directly relevant to a discussion of research design, and both retain the classic labels of internal and external validity and follow the original definitions offered by Campbell and Julian Stanley in 1966. As they argue, the establishment of the internal validity of a causal proposition involves showing that a factor is the cause of an effect or, more modestly, probing alternative hypotheses and opting for those that stand up best to attempts at disconfirmation. In contrast, the verification of the external validity of a causal proposition entails demonstrating that a causal proposition can be generalized beyond the domain of the actual units, spatial and temporal setting, and specific treatments that are actually studied. In turn, both internal and external validity can be viewed as distinct and equally essential goals of research design.

Second, Campbell *et al.*'s discussion of the problems of causal inference is organized around a fairly *ad hoc* and cumbersome list of threats to validity. To be sure, the listed threats to validity are all important, and their analysis of the way different designs get around or fail to get around these threats to validity is exemplary. Indeed, they offer many specific recommendations that creatively respond to thorny and complex problems routinely encountered in the conduct of substantive research. However, they do little to present their list of threats in a logically explicit manner, to distinguish threats that are relevant to experimental designs from those that pertain to nonexperimental designs, and to offer a sense of the extent to which the threats to validity they discuss constitute a complete list. In place of their list, we propose a scheme whereby research design choices are evaluated in light of a set of problems of causal inference that are directly and indirectly affected by design choices.

Figures 1 and 2 provide a graphic representation of the assumptions and potential problems of causal inference that design options impact in a direct manner. With regard to the analysis of single units, the core concerns are the need to ascertain that the posited direction of causality is correctly described and that the proposed model of the link among causal variables is fully and correctly specified. With regard to the analysis of multiple units, the key issues are the need to validate the assumptions of unit independence and unit homogeneity. In turn, Fig. 3 draws attention to the need to place the discussion of research design in the broader context of the research

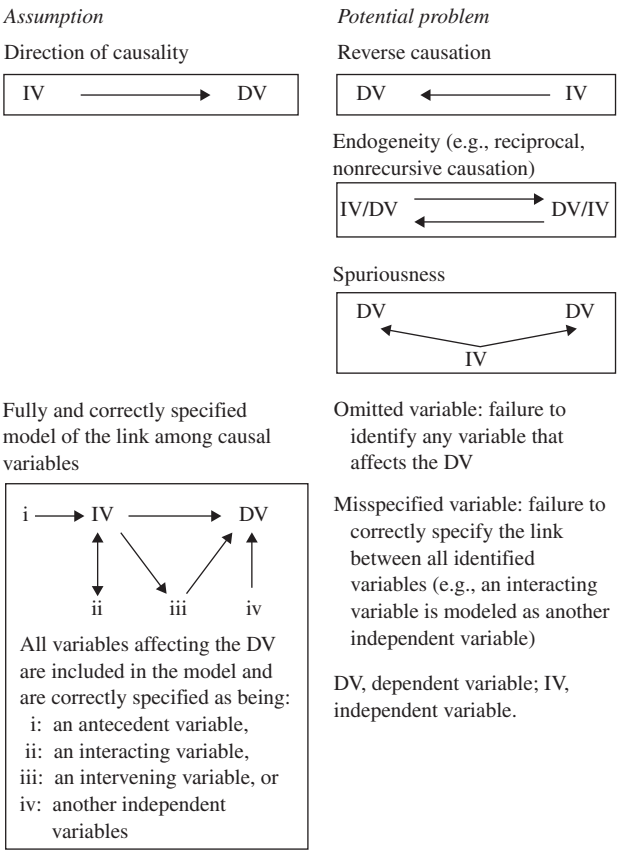


Figure 1 Problems of causal inference: Issues in the analysis of single units.

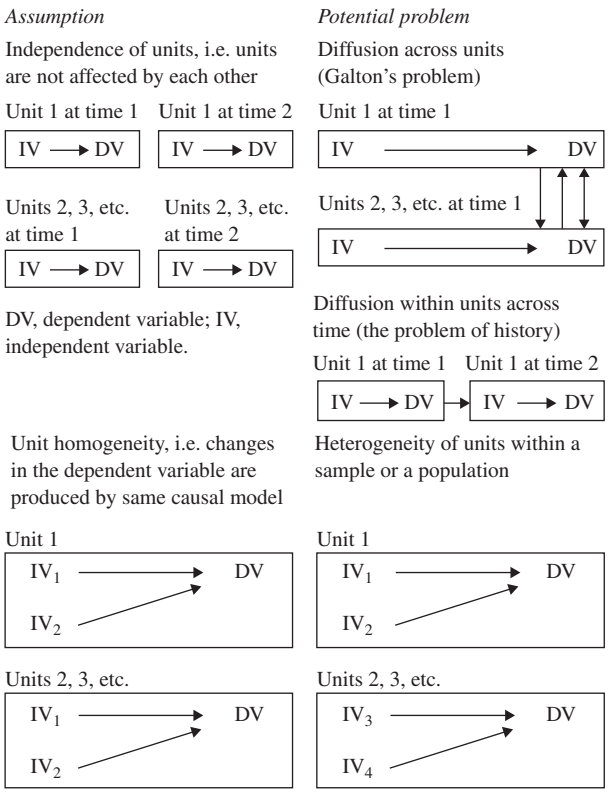


Figure 2 Problems of causal inference: Issues in the analysis of multiple units and singular units across time.

process and to consider how design options have an indirect but important impact on causal inference as a result of their requirements and contributions vis-à-vis theory and data.

Third, Campbell *et al*.'s discussion of design options is fairly limited and their assessment biased toward experimental designs. Indeed, they place such strong emphasis on certain design options—the ability of researcher to consciously manipulate independent variables and randomly assign units to treatment and control groups—that they offer a narrow and unbalanced optic on questions of research design. They do highlight the difficulties of conducting experiments and emphasize the ways in which experiments are likely to never guarantee that all threats to validity are eliminated. Moreover, occasionally they offer an exemplary display of pluralism. However, they tend to overlook some significant shortcomings associated with experimental data and downplay the significant potential virtues of nonexperimental data, and they ignore both the role played by design choices other than those that are defining elements of experiments and the way in which design choices have an indirect impact on causal inference. Thus, to assess the

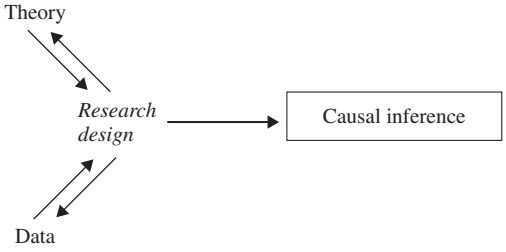


Figure 3 Research design in context.
Although research is frequently portrayed as proceeding in a linear fashion from theory to data collection and causal inference, as a matter of practice research is an interactive and iterative process. Thus, it is important to recognize that design options are (i) affected by the state of available theory and can also affect theory building and (ii) limited by the availability of data and also have an impact on the quality of data that are generated.

strengths and weaknesses of different research designs in a systematic and balanced manner, it is important to recognize that design options revolve around at least four core questions: How is the value of the independent variable(s) assigned? How are units selected? How

Table I Research Designs: Classification Criteria and Options

Classification criteria	*Main options*	*Disaggregate designs*
How is the value of the independent variable(s) assigned?	Manipulation, random assignation of cases to treatment and control groups	Experimental (randomized experiment)
	Manipulation, nonrandom assignation of cases to treatment and control groups	Quasi-experimental
	Nature	Observational
How are units selected?	Random sample	Representative
	Purposive (deliberate or intentional) sample	Typical (mode, mean, or median)
		Heterogeneous (extreme and typical)
	Entire population	Census
How many units are selected?	Many	Large-*N*
	Few	Small-*N*
	One	Case study
How are comparisons organized?	Across units	Cross-sectional
	Across time	Longitudinal

many units are selected? and How are comparisons organized (i.e., whether temporally and/or spatially) (Table I)? Moreover, it is necessary to address the manner in which design choices have both a direct and an indirect impact on causal inference.

Recognizing the seminal nature of the work by Campbell *et al.* but also seeking to overcome its limitations, this article provides an overview of the current state of knowledge on research design. The number of alternative research designs is, in principle, very large. Indeed, focusing only on the main options related to the four criteria presented in Table I, it is clear that the range of possible designs is much greater than usually believed, although not all will be sensible. However, the following discussion is organized in terms of research traditions, given that this is how much research and discussion about design options is carried out, focusing on experimental and quasi-experimental designs first, turning next to large *N* or quantitative designs, and completing the discussion with case study and small *N* or qualitative designs. In each case, the focus is on the prototypical designs associated with each tradition and an assessment and comparison of the strengths and weaknesses of these designs. The need for choices about research design to be explicitly addressed and justified, and the need to actively construct bridges across research traditions, is emphasized.

Experimental and Quasi-Experimental Designs

Experimental designs, also called randomized experiments, are characterized by two distinguishing features: (i) the conscious manipulation by the researcher of a treatment or, more generically, an independent variable of interest, and (ii) the random assignment of units to treatment and control groups (Fig. 4). Experiments also draw upon other design elements, and thus it is possible to distinguish a variety of experimental designs. However, these two features are the defining features of the prototypical experimental design.

The strengths of this design are undeniable. Given that the treatment is administered before a posttest measure on the dependent variable is registered, the direction of causality is clearly established. In addition, the random assignment of units to treatment and control groups ensures that these groups are equivalent; that is, they do not vary systematically on any variable except the manipulated variable. In this manner, multiple unexamined variables are held to not vary in any patterned manner and, hence, control over these alternative hypotheses is ensured by turning other variation into noise. Thus, the beauty of experimental designs is that the data generated lend themselves to a particularly simple form of analysis, in which the differences in posttest measures of the treatment and control groups can be interpreted as a measure of the causal effect of the treatment. Indeed, experiments are the most powerful means of gaining control and establishing the internal validity of a causal claim (Table II).

A quasi-experimental design differs from an experimental design in that in the former the treatment is consciously manipulated by the researcher but the units are not randomly assigned to treatment and control groups. Thus, this design offers less of a basis for assuming the initial equivalence of treatment and control groups and requires researchers to consider how third extraneous variables might confound efforts at ascertaining causal effects and to be explicit about alternative hypotheses. The difficulties this difference between experiments and quasi-experiments introduces are central to the work of Campbell *et al*. Indeed, their work can be understood essentially as an effort to alert researchers to

The units are randomly divided into two groups—the treatment group and the control group—and a treatment is applied to the treatment group and not to the control group.

Measures are taken of the values of the dependent or outcome variable for the treatment and control groups. The generated data are organized as follows:

	Treatment group Units 1, 2, 3, . . . , n	Control group Units 1, 2, 3, . . . , n
Independent variable	Yes (treatment)	No
Dependent variable	*Value*	*Value*

Alternatively, after the units have been randomly divided into two groups, a pretest measure of the value of the dependent variable for both the treatment group and the control group is made. The generated data are organized as follows:

	Treatment group Units 1, 2, 3, . . . , n		Control group Units 1, 2, 3, . . . , n	
	T_1	T_2	T_1	T_2
Independent variable	No	Yes (treatment)	No	No
Dependent variable	Value	*Value*	Value	*Value*

In both instances, the analysis of the data focuses on the difference in values on the dependent variable between treatment and control groups (the italicized values), a difference that can be interpreted as a measure of the causal effect of the treatment.

When a pretest measure on the dependent variable is obtained in addition to a posttest measure, a comparison of the values of units of the treatment and control groups at T_1 (the underlined values) serves to double-check that all units are "equivalent," given that this goal is already ensured by the random assignment of units.

Figure 4 Experimental designs: Defining features.

Table II Experimental Designs: An Assessment

Design elements	*Strengths*	*Weaknesses*
Conscious manipulation of independent variables	Establishes causal direction (internal validity)	Lack of viability, due to practical and/or ethical reasons, to study many important questions
		Unsuitable for the study of action and lack of attention to causal mechanisms (internal validity)
		Unsuitable for an assessment of complex causes
		Requires *a priori* knowledge of plausible independent variables and well-specified causal model; not useful in theory generation
		Tends to generate obtrusive, reactive measurements
Random assignment of units to treatment and control groups	Establishes equivalence of units, helps guard against third, extraneous variables (internal validity)	Lack of viability, due to practical and/or ethical reasons, to study many important questions
Nonrandom selection of units, setting, and treatment		Difficult to generalize from sample to population (external validity)

the various ways in which factors other than the treatment may be responsible for the observed posttest variance between treatment and control groups. Thus, in large part due to the significant effort by Campbell *et al.* to anticipate potential threats to the validity of quasi-experiments—and to offer a range of designs that minimize and help guard against such threats—researchers have a sophisticated road map for consciously taking into consideration how third extraneous variables may exercise a confounding effect. In summary, even though quasi-experiments are more complicated than experiments, both designs offer a powerful basis for making claims

about causality, and it is thus hardly surprising that these designs have been used often in psychology and economics and are increasingly being adopted by political scientists.

The considerable strengths of experimental and quasi-experimental designs notwithstanding, it is important to recognize the serious weaknesses associated with these designs. One standard limitation concerns the viability of conducting experiments, whether for practical and/or ethical reasons, to study the sort of questions that are of interest to many social scientists. The fact that we cannot or would not want to manipulate some variables and/or randomly assign subjects to control and treatment groups has vast implications. Indeed, as Hubert Blalock (1991) argues, "If we were to confine our analyses to experimental and quasi-experimental designs, virtually all of sociology and political science would have to go to the wayside." (p. 332) However, the problem is deeper than the standard concern with viability would indicate and actually derives from the very substance of the subject matter of the social sciences.

The core difference between the social and natural sciences is that the former are first and foremost about agents and actions. However, a basic feature of experimental and quasi-experimental designs—the conscious manipulation of the treatment—is founded on a notable asymmetry, which places the experimenter in an active role and relegates the experimentee to a passive, reactive role. In effect, experiments and quasi-experiments treat subjects as objects and embody an implicit behaviorist perspective that renders them ineffectual instruments for the study of the causal significance of action and, relatedly, forces them to be silent on the critical question of causal mechanisms. Indeed, as William Shadish *et al.* recognize, although experimental and quasi-experimental designs can be used to predict what the effect of a factor is, they are less useful for explaining why and how such effects are generated. This is a significant limitation. Indeed, it is not far-fetched to argue that the internal validity of a causal argument is not fully established until the causal mechanisms that generate the causal effect are properly identified and tested.

Experiments are also of limited use with regard to the study of complex causal relationships. This failure is in part due to the inability to deal squarely with agency through experiments and quasi-experiments. Consequently, these designs are not suitable means for getting at a core theoretical concern in social theory: the interaction between structures and agents, or macro and micro causal factors. In addition, because experiments and quasi-experiments are in essence instruments geared to the study of short-term effects, they are not useful for studying causes that work themselves out over an extended period of time or for assessing the interaction between long- and short-term causal factors. Moreover, because experimenters, as a way to manipulate the treatment, must rigidly assign variables to the status of independent and dependent variables, they do not constitute a means of assessing feedback effects or reciprocal and nonrecursive causation. In short, a range of theories simply cannot be tested through experimental and quasi-experimental designs.

Another significant limitation concerns the generalizability of results derived from experiments. The reason for this is that although a defining feature of experiments is that units are randomly assigned to treatment and control groups; the units, settings, and treatments are usually not randomly selected. Indeed, inasmuch as control is gained through the manipulation of the treatment and the random assignment of units to treatment and control groups, it is practically inevitable that the ability to randomly select these units will decline. The purposive or intentional selection of units that researchers have to use is not without merits, and it can certainly be carried out with an eye to the relationship between the studied sample and the universe. However, even when carefully practiced, purposive selection tends to lead to biased results. Indeed, in many instances the gap between the conditions of experimental research (whether in the laboratory or in field settings) and the phenomenon being studied can be substantial; hence, the ability to generalize beyond the domain of the actual units, spatial and temporal setting, and specific treatments that are examined is compromised. Thus, it is important to recognize that both internal and external validity are critical aspects of knowledge, and that there are good reasons for the standard view that the gains made by experiments in terms of internal validity tend to come at the cost of a loss in external validity.

Finally, two other limitations are indirect consequences of the manipulation of treatments that characterizes experimental and quasi-experimental designs. First, because of this design element, experiments and quasi-experiments approach the two-sided issue of causal theorizing from one side: They focus on the effect of causes rather than on the causes of effects. Thus, they require *a priori* knowledge about plausible independent variables and presume that all that needs to be determined is the effect of preselected causes. Hence, experiments and quasi-experiments are of less use during the early, exploratory stage in the research process, when a typical challenge is to uncover potential independent variables by working backward from a dependent variable. Furthermore, experiments and quasi-experiments assume a well-specified causal model and thus require that the state of theory building already be fairly advanced. Second, the manipulation of treatments makes experiments and quasi-experiments a particularly obtrusive and reactive form of generating data. The gain made by manipulating treatments comes at the cost of the

quality of the data generated for analysis. In summary, as shown in Table II, although experiments especially, but also quasi-experiments, have some important strengths, they are also associated with a number of significant weaknesses.

Observational Designs

The distinction between experimental and observational studies is a fundamental and deep one. The key difference is that in observational studies, control of possible third variables is not attained "automatically" through random assignment. Rather, in observational studies third variables have to be formulated and measured explicitly, and control is sought through the analysis of the data. However, the distinction between large *N* studies, on the one hand, and case and small *N* studies, on the other hand, is equally profound and probably more pervasive. This second distinction is not unique to observational studies. Indeed, the quantitative vs qualitative distinction runs through both the experimental and the observational research communities. However, the discussion of this distinction is developed only in the context of observational studies and focuses primarily on the prototypical quantitative and qualitative studies: the large *N*, cross-sectional study and the small *N* study based on the longitudinal case study, respectively.

Large *N* Studies

A large *N* study has some considerable strengths that make it, in some regards, superior to an experimental study. First, because it uses data generated through the natural course of events, it is a viable design for studying important questions that involve nonmanipulable variables that cannot be addressed with an experimental method. Second, because a large *N* study is not constrained by the requirement to randomly assign units to treatment and control groups, it is more likely to entail a randomly selected sample. This is a major benefit that gives large *N* researchers the ability to generalize beyond the domain of the actual units, spatial and temporal setting, and specific treatments that are actually studied and to establish the generalizability of their findings (external validity), a core weakness of experimental methods. However, much as is the case with regard to random assignment in the context of experimental designs, the beauty of random selection is tarnished by the difficulty of applying this design element to many units, settings, and variables (Table III).

Another important strength of large *N* studies, due to their tendency to study quite large samples, is their ability to use statistical analysis to establish patterns of association with a high degree of precision and confidence. Such results, however, differ significantly from those that can be obtained using experimental data. Indeed, although experimental data offer strong grounds for making claims about causality, because large *N* studies are observational it is crucial to remember the simple but profound point that "association is not causation" and, moreover, that even "a lack of correlation does not disprove causation" (Bollen, 1989: p. 52). This does not mean that claims about causality cannot be made on the basis of large *N* studies. Indeed, such claims can be made legitimately if researchers verify the causal assumptions of their statistical models, including nonspuriousness, the lack of omitted variables, independence of cases, and unit homogeneity. However, it is extremely difficult to substantiate that patterns of association establish causality (internal validity).

Table III Observational Designs: Large-*N* Studies

Design elements	*Strengths*	*Weaknesses*
Assignment of value of the independent variable(s) by nature	Viability of studying important questions that involve nonmanipulable variables	
Random selection of units or selection of an entire population	Establishes generalizability (external validity)	Lack of viability for many units, settings, and variables
Selection of many units that are compared cross-sectionally and/or longitudinally	Establishes patterns of association with a high degree of precision and confidence	Association does not establish causation (internal validity) and it is difficult to verify the causal assumptions in statistical models
	Constitutes a tool for theory generation	Associations are more interpretable when guided by a strong theory, i.e., a theory with few variables and detailed predictions
		Measurement validity is harder to establish the larger the *N* and the larger the number of variables
		Lack of attention to causal mechanisms (internal validity)

In short, claims about causality derived from large *N* studies should be treated with great caution.

Regarding the indirect consequences of design choices, it bears emphasizing that large *N* studies are quite useful for exploratory work. In an experiment, treatments need to be planned in advance and need to be relatively few in number, a particularly difficult and restrictive requirement for many areas of the social science. In contrast, at least with regard to the kinds of questions that are common to macro-oriented inquiries in sociology, political science, and economics, it is possible to obtain more data after examining some relationships. This practice runs the risk of capitalizing on chance if not rigorously tested on new data, but it is potentially an important part of any study.

This benefit notwithstanding, a core problem regarding the demands that large *N* research puts on theory and data should be highlighted. On the one hand, associations are more interpretable when guided by a strong theory—that is, one with few variables and detailed predictions. In other words, good large *N* research not only requires a causal model specified prior to testing but also puts a heavy burden on the ability of theory builders to reduce the number of potential explanatory factors. On the other hand, to ensure that the causal model is fully specified, it is important that potential independent variables are not omitted. This fact complicates the interpretation of results. Moreover, it makes a heavy, sometimes practically impossible, demand concerning data. Indeed, because potentially confounding factors are not controlled "automatically" through random assignment in large *N* studies but rather must be modeled and measured explicitly, the need to collect data on numerous variables across a large number of cases and/or time entails some serious costs. This demand opens the door to well-founded charges concerning the validity of the data. In addition, the data requirements of large *N* designs make it extremely difficult to offer a quantitative study of causal mechanisms, a crucial shortcoming that further weakens the claims about causality (internal validity) that can be made on the basis of a large *N* study. Gains in terms of generalizability and findings about patterns of association are heavily dependent on good theory and good data—valuable resources that are not always available.

Some advances that go beyond the cross-sectional design typically used in large *N* studies help get around some of these limitations. Especially noteworthy are time series and panel studies, event history analysis, and hierarchical modeling. These methods offer fruitful ways of addressing the problems associated with the assumptions of unit independence and unit homogeneity. However, these potential gains tend to make even more imposing demands in the area of data collection than cross-sectional designs and thus are achieved at a cost. In summary, as Table III highlights, much as in the case of experimental and quasi-experimental designs, it is only fair to note the strengths of quantitative or large *N* designs but also to recognize their weaknesses.

Case and Small *N* Studies

The status of the case and small *N* studies tradition in the social sciences has been marked by a notable disjuncture between the high number of its practitioners and its low standing in the broader methodological community. This odd situation, however, has been changing over time. After famously stating that "one-shot case studies [are] of almost no scientific value," Campbell retracted this harsh critique of a staple of qualitative research. In turn, statistician David Freedman has argued that case studies, when well designed, can establish causality in a more powerful manner than is standard in most quantitative research. Finally, a flourishing debate on case and small *N* research in recent years has increasingly made explicit the rationale for choosing a qualitative research design and the methodological foundations of rigorous qualitative research. As a result, the reconstructed logic of qualitative methods is catching up with the logic of qualitative research, and a clear sense of the strengths and weaknesses of this tradition is emerging.

One well-established and important strength of case and small *N* studies is that they represent a viable design to address important questions that involve nonmanipulable variables. This is a virtue shared by small and large *N* studies, but in this regard small *N* researchers have an advantage, particularly in light of their ability to gain access to various types of data and form a complex picture of their cases, including a keen sense of developments over time. For this reason, much of the analysis in the social sciences on a range of critical questions, especially during the initial stages in the research process, is done by qualitative researchers (Table IV).

A feature of case and small *N* studies setting them apart from both large *N* and experimental studies and accounting for a key comparative advantage is the manner in which these studies can be used to analyze the role of agency and hence to establish causal mechanisms. This is a critical point. Indeed, as philosophers of science and methodologists have increasingly insisted, it is not enough to focus on causal effects. Rather, it is necessary to go beyond statements about what the effect of a factor is and to consider how and why the effect operates. This calls for the specification of causal mechanisms, which requires, in most areas of the social sciences, considering agency. In this regard, it is probably not an overstatement to suggest that a qualitative design is the method *par excellence* to study agency and hence to empirically ground the analysis of causal mechanisms.

A related strength of case and small *N* studies is that they offer a basis for assessing causality (internal validity).

Table IV Observational Designs: Case and Small-*N* Studies

Design elements	*Strengths*	*Weaknesses*
Assignment of value of the independent variable(s) by nature	Viability of studying important questions that involve nonmanipulable variables	
Selection of one or a few units that are compared longitudinally and/or cross-sectionally	Addresses agency and establishes causal mechanisms	In the absence of strong theory, it is difficult to establish control and eliminate potential variables (internal validity)
	The study of causal mechanisms, and within- and cross-case analysis, offers a basis for assessing causality (internal validity)	
	Constitutes a powerful tool for theory generation	
	Measurement validity is easier to establish the smaller the *N*	
Purposive selection of a few units		Difficult to generalize from sample to population (external validity)

This is done, most fundamentally, through a within-case form of analysis that uses empirical evidence about causal mechanisms as a way to check expectations concerning the direction of causal processes, to eliminate potential third variables, and to verify the assumption of unit independence and unit homogeneity. Moreover, this goal is also frequently advanced by combining the prototypical longitudinal within-case design with a cross-sectional design, such as a traditional cross-case study or a cross-sectional within-case study either focused on different implications of a theory or cast at a different level of aggregation. Through these means, which capitalize on knowledge about process and the ways in which case studies lend themselves to more observations than are suggested by the strict definition of a case study as an $N = 1$ study, small *N* researchers can make valuable contributions that are important to highlight.

However, the weaknesses of qualitative methods as a tool for establishing causality should also be duly recognized. To use case studies to assess causality, it is necessary to be explicit about the posited causal model—including all its variables, the relationship among the variables, and the form of the effect of each variable or combination of variables—as well as about the causal mechanisms that are considered to be in play. This is a demanding task, but failing to specify these things in advance makes it easy for researchers to simply focus on confirming evidence and to disregard alternative interpretations. Indeed, a drawback of most analyses of causal mechanisms is that they fail to specify formally what mechanisms are posited and what plausible alternative mechanisms should be considered and also to set up the study of mechanisms as a standard test among competing hypotheses.

Moreover, even if such steps are taken, it is necessary to recognize the limits of small *N* analysis, in light of the number of observations they entail, as a tool for causal assessment. Some important exceptions to this general principle exist. First, occasionally qualitative researchers may be able to design their research in a way that resembles a natural experiment, in which the hypothesized causal factor changes markedly while other factors remain the same. Second, it is possible to establish patterns of association with precision even with a relatively small sample using techniques of analysis such as exact tests, permutation tests, resampling, or Bayesian methods that do not rely on asymptotics.

Third, presuming a powerful theory is being tested, a few observations could very well serve as the basis for clear results. However, frequently qualitative researchers seek to assess complex causal relationships and, when this is the case, small *N* designs tend to constitute weak means of controlling for alternative hypotheses and for ruling out chance. Indeed, claims to test complex causes using small *N* designs rely on the quite unreasonable assumptions that (i) the world operates in a deterministic fashion, (ii) the proposed causal model is complete, and (iii) there is no measurement error. In short, the use of case studies and small *N* studies to assess causality tends to rely on some very stringent assumptions concerning the state of theory and data.

Another significant limitation concerns the difficulty of using case studies to make generalizations. When studying a small number of cases, random selection is not advisable. Indeed, random selection offers a basis for generalization only when a large number of cases are selected and hence the law of large numbers takes force. Thus, qualitative researchers have to resort to the purposive selection of their cases. When this is done, it is important to avoid drawing a sample of convenience that bears an unclear relationship to the broader population. Moreover, it is important that researchers be aware of how their choice of cases might introduce bias.

Such efforts to select cases in a conscious and careful manner, however, should not be mistaken as steps providing a sufficient basis for making claims about generalizations (external validity).

Finally, with regard to the indirect consequences of design choices, two strengths of case and small *N* studies deserve mention. One is that this type of design constitutes a powerful tool for theory generation. Indeed, one of the clear benefits of qualitative research is its fruitfulness at a tool for generating ideas about causal variables and mechanisms and for theorizing that is closely informed by knowledge of the substantive problem of interest. Another related benefit is that the detailed knowledge of context that is associated with case studies plays a critical role in helping researchers establish measurement validity. In summary, as is the case with the experimental and quantitative traditions, the qualitative tradition is characterized by a mix of strengths and weaknesses (Table IV) that must be considered in any balanced assessment of the potentials of different research designs.

Conclusion: Choices and Bridges

The broad message concerning research design we have sought to convey is that making causal inferences about the complex realities of interest to social scientists is probably more difficult than is generally believed and that questions of research design play a key role in determining whether researchers have a solid basis for making claims about causal relationships. Technical fixes cannot, in general, get around design problems, but more attention goes to the former than the latter. Indeed, given the centrality of research design to the research process as a whole, it is probably fair to say that research design is a relatively unappreciated aspect of methodology and that it deserves more attention from methodologists and practicing researchers alike. Beyond this generic admonition, two further points that build on but go beyond the previous discussion offer material for further consideration.

One point is the need for choices about research design to be explicitly addressed and justified. There is a tendency for researchers to work within distinct research traditions and to simply opt for certain designs as a matter of default. Such a tendency is understandable in that different designs require different skills and training, and in a practical sense researchers are thus not free to choose among research designs. Nonetheless, short of justifying their design choices in light of the range of possible options, at the very least researchers should address the impact of their choices on the certainty of their conclusions. As this article shows, this entails a consideration of the direct impact on the prospects of causal inference of the four core choices involved in research design and of the indirect impact, due to their requirements and contributions vis-à-vis theory and data, of these choices.

A second point is the need to creatively construct bridges across research traditions. Although it is common for certain traditions to be presented as inherently superior to others and the standard against which other traditions should be measured, this article has shown that all traditions are characterized by certain strengths and weaknesses and that it is thus more accurate and useful to think in terms of the tradeoffs involved in working within different traditions. An implication of this assessment, then, is that greater effort should be made to capitalize on what are clearly the complementary strengths of different traditions (compare Tables II–IV).

Efforts at bridging, whether carried out through multiple studies on the same question or mixed designs that combine multiple designs within a single study, are very demanding. Thus, although it is common to point out that multiple studies in the context of a shared research program offer a way of combining different designs, such combinations are only effective inasmuch as research programs are organized around clearly specified concepts and questions and are advanced, at least to a certain extent, through explicitly coordinated teamwork. In turn, the effective use of mixed designs requires a level of methodological sophistication, as well as theoretical and substantive knowledge, that is rare. Nonetheless, the high payoffs associated with the use of mixed methods make these options strongly recommendable.

See Also the Following Articles

Experiments, Overview • Explore, Explain, Design • Longitudinal Cohort Designs • Observational Studies • Quasi-Experiment • Sample Design • Survey Design • Time-Series–Cross-Section Data • Validity, Data Sources

Further Reading

Blalock, H. (1991). Are there really any constructive alternatives to causal modeling? *Sociol. Methodol.* **21,** 325–335.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Brady, H. E., and Collier, D. (eds.) (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield/Berkeley Public Policy Press, Lanham, MD.

Campbell, D. T. (1988a). Factors relevant to the validity of experiments in social settings. In *Methodology and Epistemology for Social Science: Selected Papers. Donald T. Campbell* (E. Samuel Overman, ed.), pp. 151–166. University of Chicago Press, Chicago [Original work published 1957].

Campbell, D. T. (1988b). Degrees of freedom and the case study. In *Methodology and Epistemology for Social Science: Selected Papers. Donald T. Campbell* (E. Samuel Overman,

ed.), pp. 377–388. University of Chicago Press, Chicago [Original work published 1975].
Campbell, D. T. (1999). Relabeling internal and external validity for applied social scientists. In *Social Experimentation* (D. T. Campbell and M. J. Russo, eds.), pp. 111–122. Sage, Thousand Oaks, CA [Original work published 1986].
Collier, D. (1993). The comparative method. In *Political Science: The State of the Discipline II* (A. W. Finifter, ed.), pp. 105–119. American Political Science Association, Washington, DC.
Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs.* Jossey-Bass, San Francisco.
Freedman, D. A. (1991). Statistical analysis and shoe leather. *Sociol. Methodol.* **21,** 291–313.
Goldthorpe, J. H. (2000). *On Sociology: Numbers, Narratives, and the Integration of Research and Theory.* Oxford University Press, Oxford, UK.
Good, P. I. (2000). *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.* Springer, Berlin.
Kagel, J. H., and Roth, A. E. (eds.) (1995). *The Handbook of Experimental Economics*. Princeton University Press, Princeton, NJ.
King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry. Scientific Inference in Qualitative Research.* Princeton University Press, Princeton, NJ.
Mahoney, J. (2000). Strategies of causal inference in small-*N* research. *Sociol. Methods Res.* **28**(4), 387–424.
McDermott, R. (2002). Experimental methods in political science. *Annu. Rev. Political Sci.* **5,** 31–61.
Oehlert, G. W. (2000). *A First Course in Design and Analysis of Experiments.* Freeman, New York.
Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin, Boston.
Smith, H. L. (1990). Specification problems in experimental and nonexperimental social research. *Sociol. Methodol.* **20,** 59–91.
Tashakkori, A., and Teddlie, C. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Sage, Thousand Oaks, CA.
Webb, E. J., Campbell, D. T., Schwartz R. D., and Sechrest, L. (2000). *Unobtrusive Measures*. Revised Edition. Sage, Thousand Oaks, CA.
Western, B., and Jackman, S. (1994). Bayesian inference for comparative research. *Am. Political Sci. Rev.* **88**(2), 412–423.

Research Ethics Committees in the Social Sciences

Søren Holm
Cardiff University, Cardiff, United Kingdom

Louise Irving
University of Manchester, Manchester, United Kingdom

Glossary

funder(s) Organization(s) providing funding for the study through contracts, grants, or donations to an authorized member of the employing and/or care organization.
institutional review board U.S. term for research ethics committee.
research ethics committee An official or semiofficial body that has as its main task the assessment of the ethical and/or legal probity of research projects involving human research subjects before these research projects are initiated.
Strategic Health Authority—UK A body established by the National Health Service to oversee health matters for the population of a defined area.
subjects Persons agreeing to take part in the study. In some legal and regulatory documents the term "participants" is used instead. The term participants implies a more equal relationship between researcher and those being researched. This may not always reflect the true state of affairs.

Social science research projects involving human research subjects in some circumstances have to be reviewed by a research ethics committee (REC). A REC is any official or semiofficial body that has as its main task the assessment of the ethical and/or legal probity of research projects involving human research subjects before these research projects are initiated. Such bodies exist in many jurisdictions and in many research organizations, although their name varies from context to context (e.g., institutional review boards in the United States). They should be distinguished from bodies dealing with questions of scientific fraud and misconduct and from bodies whose main involvement in research ethics is the writing of guidelines for researchers or for RECs.

History and Current Structure of Research Ethics Committees

History

The current system of research ethics committees (RECs) has its origin in two separate historical developments, one internal and one external to the academic world. The closest institutional precursors to RECs are the formal and informal systems that most academic departments and research organizations have traditionally had for the evaluation of the research projects of doctoral students and junior members of staff. Senior members of staff have traditionally assessed the projects of junior staff before they could be initiated. This assessment has often concentrated on methodological issues, but it has also included assessment of the objective of the research and of any ethical or legal problems. In some settings, this assessment was formalized through, for instance, a dissertation committee that had to approve all doctoral projects.

The other major historical root of current RECs is the public and political response to "research scandals" in which instances of unethical research conduct have been exposed. Such scandals have occurred regularly, perhaps most prominently in biomedicine but also in the social sciences and psychology. Examples include Stanley Milgram's obedience studies in psychology and Laud Humphrey's "tearoom trade" study in the social

sciences. Stable features of these scandals have been harm to research subjects, violation of the rights of research subjects, deception of research subjects, exploitation of vulnerable or marginalized groups, and in general a lack of attention to the interests of research subjects.

Such scandals have, over time, led both to the tightening of legal rules and guidelines for the ethical conduct of research and to the establishment of RECs in order to ensure the following of rules and guidelines. It is impossible to give a complete account of this succession of scandals here, so those that have been most important for the formation of RECs are discussed.

The Nuremberg (Nürnberg) code from 1948 was developed as part of the judgment against the doctors involved in the horrific medical and biological research conducted on concentration camp prisoners in Nazi camps during World War II. This declaration was the first to require consent from research subjects as a necessary part of any research project.

In the 1960s, a number of publications surveyed research practice and found many examples of research conducted without any consent of research subjects or involving deception. This led the World Medical Association to put forward the Helsinki Declaration in 1964, regulating medical research and introducing the general requirement of informed consent for research. In 1966, the U.S. Surgeon General issued instruction to all research institutions receiving federal grants that they had to have institutional review boards assess all federally funded research. At approximately the same time, a number of professional associations in the social sciences put forward explicit research ethics guidelines (the American Sociological Association in 1969 and the American Anthropological Association in 1967). Despite the regulatory developments in the late 1960s introducing RECs or REC-like bodies, there has been a continuing stream of new research scandals, leading to tighter control and more explicit guidelines.

Current Structure and Formal Role of RECs in Major Jurisdictions

In the United States, RECs are called institutional review boards (IRBs). Each institution that is supported by a department or agency subject to federal policy, and that conducts research involving human subjects, must have an IRB to review and approve research. All projects in such institutions have to be reviewed, whether or not the individual project receives federal funding. If an institution does not comply, all federal funding may be removed. It is possible for institutions to use the IRBs of other institutions or an outside commercial IRB provider. Federal policy requires that IRBs have at least five members with varying backgrounds to promote complete and adequate review of research activities conducted by the institution. The membership of the IRB must include one scientist, one nonscientist, and one person not affiliated in any way with the institution. The nonaffiliated member should have knowledge of the local community, and if there is a large minority population, he or she should be representative of that particular minority. An IRB should not be composed entirely of men or of women. Experts may be called in as nonvoting advisers. IRBs, which regularly review research involving vulnerable subjects such as children or those incapacitated, must consider inclusion of a member with the appropriate expertise.

In the United Kingdom, social science research can fall under the remit of a REC in two different ways. It can be conducted in the National Health Service (NHS) or on a population derived in some way from the NHS (e.g., former patients identified through NHS files), or it can be conducted at a university or other institution of higher education having a REC.

All research into health and social care that involve the NHS is subject to independent review by NHS RECs. These committees are established by the geographically defined Strategic Health Authorities or directly by the Department of Health. Their remit is to evaluate all research involving human subjects in the NHS whatever the methodology or profession of the researcher. Their legal basis is binding guidance from the Department of Health. The appointment of REC members is by a publicly advertised, open process and members are appointed for a fixed term, usually 5 years. There must be at least 7 members and no more than 18 members. The members should have a sufficiently broad range of experience and expertise to assess the scientific, clinical, and methodological aspects of research proposals. At least 3 members must be independent of any organization in which research under ethical review is likely to take place. The lay members must constitute at least one-third of the membership, and they must be independent of the NHS. At least half the lay members must never have been either health or social care professionals and must never have carried out research involving human participants, their tissue or data. For both lay and expert members, there should be a balanced age and gender distribution, and effort should be made to recruit black, ethnic minority, and disabled members.

Most UK universities have established institutional RECs to evaluate research with human subjects. These RECs have no clear legal basis, but submission of all projects involving human subjects is usually compulsory for all staff and students as part of their employment or study conditions. Because these RECs are established by the individual universities, there are no common rules

concerning membership, but most have some kind of lay representation.

The Council of Europe has agreed to a human rights convention covering a range of areas in biomedical ethics, including research ethics. The Convention for the Protection of Human Rights and Dignity of the Human Being with Regard to the Application of Biology and Medicine has been ratified by most European countries, and its provisions on research are thus binding in these countries. These provisions cover all research in the health care field, including social science research. The main provision is article 16, which states that

> *Research on a person may only be undertaken if all the following conditions are met:*
>
> *i. There is no alternative of comparable effectiveness to research on humans;*
>
> *ii. The risks which may be incurred by that person are not disproportionate to the potential benefits of the research;*
>
> *iii. The research project has been approved by the competent body after independent examination of its scientific merit, including assessment of the importance of the aim of the research, and multidisciplinary review of its ethical acceptability,*
>
> *iv. The persons undergoing research have been informed of their rights and the safeguards prescribed by law for their protection;*
>
> *v. The necessary consent as provided for under Article 5 has been given expressly, specifically and is documented. Such consent may be freely withdrawn at any time.*

A more comprehensive additional protocol to the convention, explicating and extending the provisions on research, is being prepared.

In their work, RECs will often look beyond the formal legal rules and take account of research ethics guidelines promulgated by different professional organizations. Such guidelines exist both at the international level and at the national level.

At the international level, the relevant guidelines include the Helsinki Declaration of the World Medical Association covering research in the health care setting and the Council for International Organizations of Medical Sciences (CIOMS) guidelines covering medical research in developing countries. No generally recognized international ethics guidelines exist for social science research.

At the national level, influential guidelines have been prepared by the Social Research Association and the British Sociological Association in the United Kingdom and by the American Sociological Association, the American Anthropological Association, and the American Psychological Association in the United States. All the professional codes give guidance on access to subjects, the acquisition of informed consent, the right of subjects to privacy and confidentiality, the reputation of the profession, and obligations to sponsors and colleagues.

Purpose and Function of RECs

RECs have three major purposes. They are intended to protect research subjects/participants from harm and from violations of their rights, to ensure that the quality of research projects is sufficient to warrant the "use" of human subjects, and to add public legitimation to research projects.

Protection of Research Subjects

In most cases, there is a fundamental inequality in the (social) power, status, and knowledge possessed by researchers and their research subjects. This enables researchers to exploit research subjects in various ways. Researchers may harm research subjects directly during the research process or indirectly through the way the results of the research are presented. This possibility of harm exists in all kinds of research involving human subjects. There are no research methods in which the possibility of harm to the subjects is completely removed.

In their evaluation of harm, RECs will seek to ensure that the risk and magnitude of harm are not so great that they make the project ethically problematic, and that the risk and magnitude of harm are appropriately balanced by the benefits the project can generate. There is thus both an absolute threshold for allowable harm and a relative balancing against potential benefits. The benefits in question can be benefits to the research subjects, but they may also be benefits to the research population or to society.

Researchers may also violate the ethical and legal rights of subjects, even in cases in which they do not directly harm the subjects. These rights include rights to privacy, self-determination, physical integrity, and control of personal data.

If the researchers have the permission from the research subjects to perform the research in the form of a fully informed and voluntary consent, and the research is conducted according to the permission given, then there is no rights violation involved since all the rights mentioned previously are waivable. A REC will therefore seek to ensure that the description of the research project given to potential research participants is comprehensive and truthful in order to make any consent given fully valid.

Scientific Evaluation of Projects

As part of the approval process, RECs evaluate the scientific quality of the research projects submitted. This function of RECs is more controversial than the

function of protection of research subjects. It is sometimes claimed by disgruntled researchers that RECs should not assess scientific quality and/or are not qualified to assess scientific quality.

However, it is important to note that the assessment of scientific quality can be seen as a necessary component in the overall evaluation of the ethical status of a project. The major ethical justification for "using" human subjects in research (e.g., for using their time) is that the research project in question can generate valuable knowledge. If a research project is badly designed and is therefore unlikely to generate any kind of addition to knowledge, it is *prima facie* unethical. Scientific review is therefore part of a comprehensive ethical review.

How can the scientific quality of a wide range of projects be reviewed in RECs that typically have a limited number of members and therefore a limited range of expertise concerning research methodologies? First, it is important to note that the task of the REC is not to decide whether a given project has the best possible methodology and research design but only to decide whether the methodology is good enough. This is still not an uncontentious evaluation, but it is clearly less contentious than trying to decide on the best methodology. RECs can generally solve this task either by using external experts for projects for which no members are expert or by inviting the researchers to explain their methodology and its strengths and weaknesses.

RECs as Public Bodies Legitimating Research

Although researchers often view RECs only as an obstacle to be overcome, it is important to acknowledge that RECs also legitimate and thereby facilitate research. The right to academic freedom and to free choice of research topic (where it exists) is only a negative right protecting the researcher from interference from outside bodies; it is not a positive right giving the researcher permission to involve other persons as research subjects. RECs are a way of giving ethically legitimate research a form of quality approval, and being able to point to REC approval can help researchers in situations in which potential research subjects have an a priori negative attitude toward research.

Research involving human subjects is a social practice that relies on social acceptance for it to continue and flourish. This social acceptance has to encompass both the goals of the activity and the way the activity is conducted. Research and development is always an optional goal. It is not incoherent or irrational to think that no more research should be performed, as long as one is willing to also accept that no more progress will be made. The RECs probably have only a minor role to play in explaining the general goals of research to the public, but they do have potentially very important roles to play with regard to the social acceptance of the goals and conduct of specific projects.

If we view RECs as institutions within a democratic framework that at the same time regulates and legitimizes research, we can become clearer about the roles of the RECs. When an REC approves a project, it is not a neutral administrative act—it is also an implicit endorsement of the project and its qualities.

RECs and the Evaluation of Research in Other Countries

Developed countries often sponsor research in developing countries, either research directed at the research needs of the developing country in question or research directed at general research questions not specifically of interest to the developing country in question. Many research funding agencies require REC approval from two RECs, one in the funding country and one in the country in which the research is to take place. This creates possibilities for conflict and raises two important questions. Why require REC approval in the funding country? and What standard should be applied?

Often, agencies require dual approval because they want to ensure that they cannot be accused of performing research in developing countries that could not be performed in the home country for ethical or legal reasons. This is obviously a laudable aim, even if it contains a large element of self-interest, but it does not fully answer the second question of what standards to apply.

Standards of consent, confidentiality, anonymity, data protection, and reporting that may be suitable in the context of a developed country may not fit another cultural context. Such a standard may be ill fitting in many ways. It may not take into account different ideas about the proper limits of self-determination or privacy, and it may in certain areas be too confining and in others too permissive. How is the REC to know this? In many cases, fairly detailed knowledge is needed of the culture in question to be able to assess the ethical appropriateness of a specific research project, and it can be very difficult for a REC to get this information.

A number of national and international bodies have tried to analyze these problems and develop more precise guidelines. This activity has been particularly intense in the biomedical field, in which some AIDS research in poor countries has been seen as exploitative. A U.S. report and a UK report have reached similar conclusions that are also echoed in the latest revision of the World Medical Association's Helsinki Declaration. The main principle is that although the precise way in which informed consent is obtained and documented may vary from setting to

setting, research that is performed in developing countries has to adhere to the same ethical standards as research in the sponsoring country, and it has to be of some benefit to the country in which it is performed.

Common Problems in the Evaluation of Social Science Research by RECs

A number of problems commonly occur in the evaluation of social science research by RECs. In RECs that mostly deal with biomedical research there is often a lack of knowledge about social science research methods, and the professional members of such RECs (often doctors and other health care professionals) may believe that the biomedical research paradigm, with its emphasis on quantitative research, is the only valid paradigm. This can lead to problems for qualitative researchers.

Other problems occur regularly even in RECs that are used to assessing social science projects. The three most common problems concern research in sensitive areas, research involving deception, and observational research in spaces on the border of private and public space.

Research in sensitive areas can be difficult to evaluate either because it is controversial or because it raises questions about the role of the researcher when he or she is gaining knowledge about illegal actions performed by research participants or is indirectly participating in these illegal activities. RECs should never disallow valid research projects simply because they are controversial, but this has occurred in many instances. This problem is especially likely to occur if the REC in question is institutional and therefore has a potential conflict of interest between the unbiased evaluation of the research project and the interests of the institution in maintaining its reputation.

Research involving deception in inherently problematic because it is impossible for the research subjects to give true informed consent if they are deceived about the purpose or procedures of the research. In some jurisdictions deceptive research is prohibited, but in those in which it is allowed RECs have a difficult task in deciding how to balance the interests of the research subjects against the knowledge likely to be produced by the research. Classic examples of deceptive research for which the ethics of the research is still being discussed are Stanley Milgram's obedience studies and Laud Humphrey's "tearoom trade" study.

Observational research can raise problems for RECs in cases in which it is performed without the consent or knowledge of some of the persons being observed, especially if they are observed in spaces that are on the border between public and private space. It is generally accepted that by performing actions in a public space a person has relinquished his or her right to privacy concerning these acts, and that any observation of persons in a public space for research purposes is therefore *prima facie* justified. However, the demarcation between public and private space varies between cultures (and between age cohorts of individual cultures) as well as between different agents in specific social settings. A hospital ward, for instance, may be conceptualized as a public space with regard to the nurses and other health care professionals working there but as a private (or at least semiprivate) space with regard to the individual patients. A REC asked to assess a project involving observation of nurses washing patients would therefore have to make a number of difficult decisions concerning consent from nurses and patients and the proper limits of observation in this setting.

Future Trends in REC Function

Like all mechanisms for public control, RECs have to develop as societies and research practices develop. A number of current trends in societal development are likely to influence the future configuration and function of RECs: the reconfiguration of autonomy in postmodern societies, the recognition of the full implications of multiculturalism, the increasing commercialization of research, and the perceived need to control compliance with REC decisions.

There is no doubt that the current system of research ethics regulation, including the REC system, is built on modern ideas of the ideal, rational decision maker weighing harms and benefits according to a rationally construed ordering of preferences. In postmodern society, this ideal is being increasingly undermined. Although additional weight is being given to the rights of individuals to organize their lives as they choose, it is also recognized that they may do so in ways that do not conform to ideals of rational decision making. RECs will therefore have to struggle to develop ways of protecting this new kind of postmodern decision maker, with his or her own constructed set of values and preferences.

Most RECs function in multicultural societies, but until now they have mostly functioned based on the values of the majority culture. As the full implications of multiculturalism are developed and absorbed into public life, RECs will also have to change their composition and function. A major discussion for the future will be whether REC membership should be representative of the different cultures involved in projects evaluated by the REC or whether the values and ideas of these cultures should be incorporated into the function of the REC in some other way. Representative membership is an attractive idea but may be practically impossible to achieve in areas where the number of relevant cultural groups is large.

Research is being increasingly commercialized. This happens both directly when research is initiated, directed, and funded by commercial organizations and indirectly when academic institutions become more focused on commercial exploitation of the intellectual property generated by their employees. RECs will have to respond to this trend. The response is likely to include much closer scrutiny of the relationship between the researcher and the funder, with particular focus on the researcher's independence and on whether the funder can bar publication of the research if the results are not favorable. More focus is also likely to be placed on how research subjects are treated in regard to the distribution of the economic and other benefits flowing from the intellectual property generated through the research.

A final trend that is already evident in RECs dealing with biomedical research is a move to giving RECs a role not only in the initial ethical assessment of projects but also in the control of whether the projects are carried out in an ethical manner. This is likely also to happen for other kinds of RECs. When research "scandals" are unveiled, a standard response is "We must have stricter regulation," but it is doubtful whether this is a correct and useful response. Many of the scandals concern projects that have either never been approved by a REC or are conducted in breach of the approved protocol. It is unclear why and how stricter regulation can help in such cases. The more reasonable response seems to be to punish the transgressors and to ensure better control in the future so that no unapproved research can be conducted and breaches of the approved protocols can be detected and rectified.

See Also the Following Articles

Ethical Issues, Overview • Human and Population Genetics

Further Reading

Beauchamp, T. L., Faden, R. R., Wallace, R. J., and Walters, L. (eds.) (1982). *Ethical Issues in Social Science Research*. Johns Hopkins University Press, Baltimore, MD.

Bulmer, M. (2002). The ethics of social research. In *Researching Social Life* (N. Gilbert, ed.), 2nd Ed. Sage, London.

Homans, R. (1991). *The Ethics of Social Research*. Longman, London.

Kimmel, A. J. (1988). *Ethics and Values in Applied Social Research*. Sage, Newbury Park, CA.

Lee, R. J. (1993). *Doing Research on Sensitive Topics*. Sage, Newbury Park, CA.

National Bioethics Advisory Commission (2001). *Ethical and Policy Issues in International Research*. National Bioethics Advisory Commission, Bethesda, MD.

Neuman, L. W. (2000). *Social Research Methods: Qualitative and Quantitative Approaches*. 4th Ed. Allyn & Bacon, London.

Sieber, J. E. (1993). *Planning Ethically Responsible Research*. Sage, London.

Residential Segregation

Michael J. White
Brown University, Providence, Rhode Island, USA

Ann H. Kim
Brown University, Providence, Rhode Island, USA

Glossary

checkerboard problem Geographic units are related to one another in space but tabular segregation indices assume independence of parcels that can be problematic.
composition independence Invariance to the group fractions in the overall region.
decomposability A property of a segregation index in which aggregation across geographic units or population groups is possible.
parcel A unit of geography such as a census tract or block group.
residential segregation The differential distribution of social groups across geography.

Residential segregation is the differential distribution of social groups across geographic space. Social groups can be demarcated by a variety of traits including, but not limited to, racial and ethnic classifications, religion, language or nativity, gender, and family structure. However, residential segregation is most commonly measured for racial and ethnic groups across urban neighborhoods. Most broadly, segregation may be seen as the differential distribution of social traits across a set of units such as census tracts, occupational categories or schools.

Introduction

Population distribution in terms of residential patterns provides a window on the social organization of groups. The extent of concentration and segregation reveals the structure of group relations in society and frames social interaction. It also provides a snapshot of a number of social processes at work such that at a given point in time, segregation indices summarize the outcome of discrimination, self-selection, governmental intervention, and differential residential patterns incidental to compositional processes. Segregation indices have become so well established that they are now cited regularly in policy documents and news media.

History and Development of Segregation Measurement

Residential segregation has been one phenomenon measured extensively by social scientists and discussions of the importance of residential segregation permeate 20th century writing in urban sociology. Efforts to develop a summary statistic for urban residential patterns extend back over the past half-century and have overlapped with efforts to measure population distribution and population diversity more generally.

For residential segregation methodology, a watershed was reached in 1955 with the publication of a methodological analysis by Duncan and Duncan. This well-traveled article reviewed six major indices, pointing out their mathematical properties, their strengths and weaknesses, and their relation to the segregation (Lorenz) curve. Presumably as a result of that article and the Duncans' own empirical research, the dissimilarity index (D) became the workhorse of segregation analysis. In the 1970's, a methodological debate broke D's hegemony, particularly as it was noted that D failed several key

statistical and substantive criteria and did not necessarily capture all of the information that would be present in a map shaded according to demographic composition.

New measures of segregation were developed, sometimes without awareness of traditional measures. For instance, alternative measures were developed to capture racial segregation across schools. The geography literature also independently spawned indices and a literature of spatial autocorrelation. In other studies of social group segregation, parallel methodological debates arose. These can be most commonly found in occupational differentiation and income segregation. Multiple indices of income and industrial inequality have been developed to measure the distribution of resources in the population of workers, households, and firms.

Although the methodology of segregation is employed most often in ethnic residential settings, it also appears in other applications. Residential segregation has been studied for socioeconomic status, nativity, language, and less often for life cycle and family structure traits. In some applications, investigators have examined an added dimension within the residential pattern, such as racial residential segregation within broad socioeconomic groupings.

For residential segregation, many researchers have adopted a strategy of reporting multiple indices of residential segregation in the aftermath of challenges to the dissimilarity index. Recently, the U.S. Census Bureau and other researchers have reported as many as 19 segregation indices of metropolitan areas for decennial censuses from 1980 to 2000.

Common Properties

All residential segregation measures try to capture the pattern of neighboring in some way, most commonly providing a normed summary statistic describing the degree of residential differentiation. Typically, indices make use of data cross-classified by social category and census geography, such as the census tract.

Perhaps the easiest starting point for the concept of segregation is the two-dimensional tabulation (see Fig. 1). From this sort of two-dimensional tabulation we can generate most of the residential segregation indices in widespread use. Such arrays arise directly in census tabulations. For instance, in the 2000 U.S. census public release data, e.g. SF1 and SF3, the "groups" above appear for race, Spanish Origin and ancestry categories. The "parcels" available are units of geography, such as census tracts or block groups. In general, any data that can be arrayed in this form—most specifically but not exclusively that for which the "group" is a socioeconomic trait and the "parcel" is a geographic unit—lend itself to segregation measurement of the type described below. In principle

	Group A	Group B	Group C
Parcel 1			
Parcel 2			
Parcel 3			
Parcel 4			
Parcel 5			

Figure 1 Two-dimensional tabulation for segregation measurement.

then, the very techniques used for summarizing two-dimensional tabulations could be carried over to residential segregation measurement. This simple fact is often underappreciated in the segregation literature, which seems to have grown up independently of the general literature in social statistics. Explicit spatial measures (below), which take into account the absolute or relative location of the parcels, require a somewhat different approach.

A good index is meaningful statistically and from the social scientific standpoint, if it can handle three or more groups simultaneously, reflects map validity, and allows comparisons over time and space. Various criteria have been nominated to evaluate any index of residential segregation. Although this list is not exhaustive, it includes:

1. A measure of association. The interpretation is a proportion of variance explained measure.
2. Normed. For example, bounded between zero and unity, where 0 indicates no segregation (no relationship between parcel and group membership; and 1 indicates complete segregation (complete predictability) of group from parcel).
3. Decomposability. A preferred measure would have a property of being able to be aggregated across geographic units or population groups.
4. Composition Independence. Some investigators argue that a preferred segregation would be invariant to the fraction minority in the overall region. Others have a preference for composition-dependent measures.
5. Social welfare criteria. Scale invariance, the principle of transfers, system size invariance and the Lorenz criterion have evolved in the income inequality literature and have been brought into the segregation literature (see James and Taeuber; Reardon and Firebaugh; Schwartz and Winship).

Few measures meet all of the desiderata and some analysts would forgo a property for theoretical reasons.

At this point we turn to a selective review, concentrating on measures that have appeared most often in the

literature. First we treat nonspatial indices, measures that can be calculated directly from a parcel-by-group tabulation. Second, we treat explicitly spatial measures. These may arise from either (a) tabulated data augmented by some coordinate or other spatial information identifying each parcel or (b) individual data with point coordinates (x–y or longitude–latitude) and associated attributes of the person or household at that location. Of course, geographic aggregation of spatial point data into parcels can generate (a) from (b).

Conventional Tabular Measures

In this section, we describe five of the more preferred and commonly implemented indices, their features, and their applications. This can be calculated from the standard parcel-by-trait calculation above. We begin with the most widely used and historically important index.

Index of Dissimilarity (*D*)

$$D = \frac{1}{2}\sum_{i=1}^{I}\left|\frac{n_{ij}}{N_j} - \frac{n_{ik}}{N_k}\right|, \tag{1}$$

where j, k are distinct groups; n_{ij} is the population of group j in the ith parcel; n_{ik} is the population of group k in the ith parcel; N_j is the total population of group j; and N_k is the total population of group k.

The workhorse of residential segregation indices, the index of dissimilarity, is the most widely used measure to compare the levels of residential segregation of racial and ethnic groups within urban areas and across them. It is calculated by taking half the sum of the absolute difference between the proportions of each group in each parcel.

Strengths and weaknesses of the dissimilarity index are well known. Strengths are several. First it has longevity; it has been in use since the early 20th century, and it formed the basic index for many early studies of residential segregation. As a consequence, there exists a long time series of dissimilarity, measures: racial residential segregation for cities, some occupational segregation indices and the like. Second, it meets several key desirable properties. It is normed with 0 indicating a proportionate distribution of each group in each parcel and 1 indicating that no groups share a parcel. It is not sensitive to group size (composition invariant); that is, doubling the number of persons of one group in each parcel (and thus increasing the fraction in the city) leaves the index unchanged. The dissimilarity index also benefits from an intuitive verbal interpretation: the fraction of one group that would have to relocate to produce an "even" (unsegregated) distribution.

The index also suffers from several key weaknesses, which have been described at length. The dissimilarity index also does not satisfy all of the welfare approach criteria, perhaps most seriously failing in the principle of transfers. Redistribution of persons across parcels that both exceed (or fall below) the mean does not change the index value. This is an undesirable payoff criterion in the eyes of many, in that redistribution away from the most extreme parcels may not change D's value. The dissimilarity index also fails for decomposition. Perhaps the most disadvantageous aspect of the dissimilarity index for contemporary segregation studies is its limitation to dichotomies. The two-group limitation was of little consequence when cities were literally seen in black and white. While efforts have been made to adjust D for three or more groups (see the work of Sakoda and of Reardon and Firebaugh), these are not widely used and have been generally superceded by other measures.

Still, ease of calculation, a high level of correlation with other indices and ease of interpretation have given the dissimilarity index considerable staying power. Finally, some authors have identified dissimilarity to be the key indicator of one dimension of segregation, namely evenness, which we discuss below.

Entropy Index (*H*)

$$H = \frac{(E^* - \bar{E})}{E^*}, \tag{2a}$$

$$E^* = (-1)\sum_{k=1}^{K} P_k(\ln(P_k)), \tag{2b}$$

$$\bar{E} = (-1)\sum_{i=1}^{I}\frac{n_i}{N}\sum_{k=1}^{K} P_{ik}(\ln(P_{ik})), \tag{2c}$$

where P_k is the proportion of group k in the population; P_{ik} is the proportion of group k in parcel i; n_i is the total population in parcel i; N is the total population (note, we define $P_{ik}(\ln(P_{ik})) = 0$ for $P_{ik} = 0$).

Introduced to the segregation literature by Theil and Finizza to measure the extent of racial segregation across schools, it is also known as the information theory index. The entropy statistic (E) gives the demographic diversity of each parcel. Thus, using the interior summation of formula (2c) above, it is possible to calculate the diversity of each individual neighborhood with respect to race or ethnic groups. Diversity is maximized (to the value of $\ln(K)$) when all groups are present in equal proportions ($p_i = p_j$). In contrast, units with only one group present would have a value of zero. This feature demonstrates the connection between neighborhood diversity and metropolitan segregation.

The entropy index (H) is bounded from 0 to 1. Values closer to zero reveal the average diversity of categories to

be similar to the total diversity level (E^*), indicating less segregation. The maximum value of unity for complete segregation is attained when every parcel contains only one subgroup. The entropy index is the weighted average deviation of each category's diversity from the total diversity, standardized by the total diversity.

The entropy index offers several positive attributes. Most notably it is readily decomposed across both geography and groups. Theil and Finizza's original work made use of this property to describe racial segregation within and between school districts. Because K is not limited to 2, the entropy index has been increasingly used to handle multiple group situations. For instance, metropolitan areas might be compared on the degree to which they are generally segregated with respect to Anglos (non-Hispanic Whites), Blacks, Latinos, Asians, and others simultaneously. Current applications of this sort are found in the work of Fischer, who looks at multiple ethnic groups and income segregation simultaneously for 60 U.S. metropolitan areas, and in the work of Reardon, Yun, and Eitle in their examination of school segregation in over 200 U.S. metropolitan areas.

One weakness of this index is that it does not strictly adhere to composition invariance (i.e., doubling the number of one group will change the index value), although this seems not to be a major issue in practical applications. Also, it does not offer quite as intuitive an interpretation as dissimilarity, but the entropy index can be seen as a proportional reduction of error measure for nominal variables.

Exposure Indices (*P*)

Isolation Index (within group)

$$P_{jj} = \sum_{i=1}^{I} \left(\frac{n_{ij}}{N_j}\right)\left(\frac{n_{ij}}{n_i}\right) \tag{3a}$$

Interaction index (across group)

$$P_{kk} = \sum_{i=1}^{I} \left(\frac{n_{ij}}{N_j}\right)\left(\frac{n_{ik}}{n_i}\right) \tag{3b}$$

where n_{ij} is the population of group j in parcel i; n_i is the population in parcel i; N_j is the total population of group j; and n_{ik} is the population of group k in parcel i.

The first of the two expressions describes the hypothetical probability that a member in subgroup j will meet another member of his/her subgroup. Hence, it has been labeled the isolation index, where greater values suggest higher levels of isolation and segregation. The second expression, P_{jk}, assesses the probability of exposure of one group to a second and greater values on this index reflect higher levels of exposure between members of two different groups or less segregation. The isolation and interaction indices for a subgroup will sum to one if there are exactly two groups in the area or if interaction indices are estimated for all additional groups.

Exposure indices (sometimes denoted P^*) have gained a considerable following in the residential segregation literature. This may be for several reasons. First, the indices are readily calculated from available census tabular data. Second, these indices have an intuitive interpretation as "exposure," and equivalently, the probability of encountering a member of one's own or another group in one's neighborhood. Perhaps the most debatable trait of the exposure family of indices is that of its sensitivity to composition. Exposure indices fail the composition-independence criterion mentioned above. It can be shown that these indices are bounded by the overall composition of the wider unit, e.g., metropolitan area, of which they are a part. Thus $P_j \leq P_{jj} \leq 1$ and $0 \leq P_{jk} \leq 1$. The composition-dependent nature of the exposure index can be seen by some as a singular deficiency. (One can norm the index for composition, as has been discussed in the literature, at which point it becomes equivalent to the correlation ratio.) Alternatively, some argue that the sensitivity to composition is sociologically meaningful. Metropolitan ethnic exposure indices do exhibit a higher correlation with overall ethnic composition than dissimilarity or entropy.

Spatial Measures

Spatial Proximity (*SP*)

$$P = \sum_{k=1}^{K} \frac{\left(N_k^2\right)\left(P_{kk}\right)}{\left(N^2\right)\left(P_{00}\right)}$$

$$P_{kk} = \left(\frac{1}{N_k}\right)^2 \sum_{i=1}^{I}\sum_{j=1}^{J} \left(n_{ik}\right)\left(n_{jk}\right) f\left(d_{ij}\right) \tag{4}$$

$$P_{00} = \left(\frac{1}{N}\right)^2 \sum_{i=1}^{I}\sum_{j=1}^{J} \left(n_i\right)\left(n_j\right) f\left(d_{ij}\right)$$

where n_{ik} is the population of group k in parcel i; n_{ij} is the population of group k in parcel j; $f(d_{ij})$ is a function of distance between parcels i and j; n_i is the population in parcel i; n_j is the population in parcel j; N_k is the total population of group k; and N is the total population.

The spatial proximity index was presented by White to address the "checkerboard" or map validity problem of other measures of residential segregation. In a slight revision of it, White norms the measure in [0, 1] and it is presented above. Despite their reference to the distribution of subgroups across geographic areas, these other indices (dissimilarity, exposure, entropy) do not account for any of the relative spatial position of the parcels.

If $f(d_{ij}) = d_{ij}$ then P equals the proportion of spatial variance explained, and the measure has a parallel to

those used in the geography literature. If $f(d)$ takes on another function, then P can represent the proportion of variation in that aspect of neighboring captured by the function. While this discussion is in terms of explicit geographic distance, in any tabulation for which the cells could be connected by some (social) distance measure, this calculation could be applied. Other spatial approaches have been devised, and as we discuss below, there is opportunity for further development along these lines.

Other Measures

The number of other potential segregation measures (and the associated literature on inequality generally) is too large to enumerate in detail. We mention only a few and their links to this literature. Several of these—often termed concentration or centralization indices—describe the relative location of (ethnic) groups among tracts within an (metropolitan) area, but these have not seen wide application in the literature. The measurement of income inequality can be shown to be related to residential segregation, especially through the segregation or Lorenz curve and associated inequality indices, such as the Gini index or the family of Atkinson measures. The interested reader is referred to other methodological summaries for discussion of properties of these measures. (Formulas are given in the work of Iceland, Weinberg, and Steinmetz, of White, and of Massey and Denton among others.)

Selecting a Measure: Dimensions of Segregation Measurement

Taken together, the literature has grown to identify 20 or more indices. The expansion of the number of competing indices has spawned a related series of investigations to analyze the "dimensions" of segregation, with an eye toward choosing an index representative of each dimension. Massey and Denton, for instance, used factor analysis to identify five dimensions of segregation. Massey, White, and Phua confirmed this for segregation statistics about a decade later. The dimensions generally identified (and the indices classified within them) include (following Iceland, Weinberg, and Steinmetz):

Evenness: *Dissimilarity, Entropy* (*or Information Theory*), Gini, Atkinson
Exposure: *Isolation, Interaction*, Correlation Ratio
Concentration: Delta, Absolute Concentration, Relative Concentration
Centralization: Absolute Centralization; Relative Centralization
Clustering: *Spatial Proximity*, Absolute Clustering, Relative Clustering, Distance Decay Interaction, Distance Decay Isolation

Several of these indices (in italics) have been described above. Although space does not permit a full discussion of these measures, some attention to their objectives and interpretation is useful.

Evenness measures generally adjust for composition. That is, dissimilarity, Gini, and entropy are calculated in such a way that the ethnic composition of the city overall does not directly influence the value of the index. This would be preferred in most measures of association, thus freeing the summary statistic of the influence of the marginal distribution. Note that this dimension contains the companion measures used to assess income inequality. The Gini index has long been the normed [0, 1] standard for an index of income inequality. It can also be used in the residential segregation case, which can be seen by plotting the associated segregation curve as the cumulative fraction of one group against the cumulative fraction of another. (Duncan and Duncan give a full exposition of this.) The Atkinson family of indices does much the same, but allows one to weight different segments of the segregation curve, and in so doing, makes explicit the value of reducing inequality in certain portions of the distribution.

Exposure indices have grown in popularity, at least in sociological circles. As discussed above, the isolation and interaction indices are influenced by population composition, and some investigators see this as preferable, since the hypothetical probability of encountering someone is the behavioral concept pursued by these indices. The correlation ratio (sometimes called eta-squared) can be interpreted as the interaction index adjusted for composition. Once adjusted in this way, it is very highly correlated with evenness measures, further indicating the differing assumptions in each group played by overall city composition. The correlation ratio can also be interpreted as a measure of association for a binomial variable. Each of these exposure indices is defined in such a way that only a pairwise comparison can be made at one time.

The several concentration indices generally calculate the distribution of a group with respect to the population densities that apply for the units of residence. Thus, a minority group would be maximally concentrated if its members resided in highest density neighborhoods. As described by Iceland and others, there are three variants of the calculation, but all three require information on the geographic extent of residential units. Somewhat parallel to this, measures of centralization capture the degree to which the group of interest resides in the neighborhoods located most closely to the center of the city. For example, a level of relative centralization approaching unity indicates that minority group members

are almost all located closer to the center of the city than majority group members. Again, this pair of measures requires explicit geographic information. Iceland, Weinberg, and Steinmetz argue that the current urban pattern (multiple nucleation) makes measures of centralization increasingly outmoded.

Clustering measures explicitly incorporate some aspect of relative spatial location, although it need not be oriented to the central business district or residential density (as above). Again, several indices exist within this group and the spatial proximity measure is one of these. As expressed above, it offers the interpretation as a normed measure. Other measures—absolute clustering, relative clustering—have parallel interpretations, although their specifications differ. The distance decay interaction and distance decay isolation indices operate similarly to the exposure indices, save that the value of the contribution across parcels is weighted by a function of distance presumed to capture the social importance of space.

Avenues for Segregation Measurement in the 21st Century

The changing demography of urban areas immediately translates into significant conceptual and methodological challenges for new research on residential segregation. The lively index debate of past years often focused on basic mathematical properties and related substantive criteria, such as the principle of transfers. The compromise has often been to calculate many indices and identify where they agree or just let the user choose. We can identify a few factors that are likely to provide pressure for deeper thinking about index choice and revisiting some of these methodological issues. Currently, technology of data dissemination and analysis all favor work with more complex demographic structure and virtually any index specification. Whereas once the choice of an index may have been guided by practical concerns about the burden of calculation, new work will not be so hampered.

The most consequential development for segregation methods is the increasing ethnic diversity of urban settings. Fueled by international migration, cities in high-income societies have come to accommodate a wider array of ethnic groups, and even in moderate income societies, ethnic diversity (fueled by both internal and international migration) leads to much the same outcome. Furthermore, the salience of residential segregation for other traits (family composition, occupation) also argues for the development and use of measures that can handle three or more groups simultaneously. Measures developed for segregation between 2 groups have been further developed to incorporate more groups (see the work of Reardon and Firebaugh and of Sakoda) but they remain more commonly applied to pairwise computations. Others, such as entropy and spatial proximity, can better manage polytomies, and it is likely the desire to analyze multigroup segregation—seen in the emerging literature on multiethnic cities—will generate more use of these particular indices.

Second, and a somewhat related development is the issue of the value of a "reference" group. If one argues that there is a clear comparison group, the dichotomous calculations of segregation from that group make sense. Thus, in the conventional analysis of ethnic segregation one might examine segregation from Canadian charter groups (English and French origin) or in the United States from the Anglo (non-Hispanic white) population. Once that reference group orientation is abandoned, then the analyst needs a way to summarize the relative position of groups to one another and/or a summary measure of segregation across all groups. Some data reduction techniques, such as multidimensional scaling or cluster analysis, could prove helpful in showing the relative segregation of groups from one another. Lest it be seen as an issue limited to ethnic diversity, the same problem would arise for the analyst interested in summarizing the segregation pattern across any nominal variable.

A third exciting challenge for segregation analysts is the incorporation of more geographic detail into the calculations and indices. Several of the indices already mentioned above travel somewhat down this road, but to be sure, most have a rather crude operationalization of urban geography. The desire to develop more geographically sophisticated measures would suggest that indices should incorporate more socially meaningful functions of space. Dating back to the 1980s, several analysts have made progress in involving geography. Measures that involve more work with neighborhood contiguity, functions of distance, and topographic irregularities might find increasing numbers of adherents in years to come.

A fourth challenge is seen in the need to "control" for other traits in segregation analysis. For instance, the analyst would like to calculate the amount of segregation between a pair of (or larger set of) ethnic groups after adjusting for the effect of income. While some work on this has been done (confirming a high level of racial segregation within income classes), a full multivariate set of controls is harder to introduce. Multilevel and microlevel datasets present some opportunities here.

A fifth challenge is that of making comparisons across space and time. Although still less well established than the unemployment rate or the Gini index of income equality, segregation measures have become a key social indicator. The challenge, then, is to begin reworking segregation analyses, exploiting improved indices and data management. Comparable work can be conducted for other settings, including developing countries, thus improving scholars' ability to make international

comparisons. With respect to time, the assessment of segregation trends is complicated by changing social identities, particularly in race and ethnic origins. This complication is further compounded by modifications of geographic boundaries. Population change within geographic units from one point in time to another results in changes in the classification and boundaries of larger urban regions such as metropolitan areas as well as of parcels such as census tracts. From one census to another, new metropolitan areas emerge while others surrender this status, new areas become contained within existing metropolises and census tracts are amalgamated or further divided.

Finally, the technology of data analysis, data management, and calculation will help analysts surmount these challenges. Large data tabulations for multiple levels of geography and time periods are being placed readily in the public domain by statistical agencies. The painstaking "hand" calculation of indices of two generations ago, or the arduous computer manipulation of data files, have been replaced by a circumstance in which most any index can quickly be calculated across multiple geographies for a wide array of social traits. Given the social importance of residential segregation by race, ethnic group, or other socioeconomic trait, it is likely that the future will see more sophisticated indices calculated for a wider array of social groups.

See Also the Following Articles

Clustering • Demography • Urban Studies

Further Reading

Duncan, O. D., and Duncan, B. (1955). A methodological analysis of segregation indexes. *Am. Soc. Rev.* **20,** 210–217.

Fischer, M. J. (2003). The relative importance of income and race in determining residential outcomes in U.S. urban areas, 1970–2000. *Urban Affairs Rev.* **38,** 669–696.

Grannis, R. (2002). Discussion: Segregation indices and their functional inputs. *Soc. Methodol.* **32,** 69–84.

Iceland, J., Weinberg, D. H., and Steinmetz, E. (2002). *Racial and Ethnic Residential Segregation in the United States: 1980–2000,* Series CENSR-3, p. 151. U.S. Census Bureau, Washington, DC.

Jakubs, J. F. (1981). A distance-based segregation index. *Socio-Econ. Plan. Sci.* **15,** 129–136.

James, D. R., and Taeuber, K. E. (1985). Measures of segregation. *Soc. Methodol.* **15,** 1–32.

Massey, D. S., and Denton, N. A. (1988). The dimensions of residential segregation. *Soc. Forces* **67,** 281–315.

Massey, D. S., White, M. J., and Phua, V.-C. (1996). The dimensions of segregation revisited. *Soc. Methods Res.* **25,** 172–206.

Morgan, B. S. (1982). The properties of a distance-based segregation index. *Socio-Econ. Plan. Sci.* **16,** 167–171.

Reardon, S. F., and Firebaugh, G. (2002). Measures of multigroup segregation. *Sociol. Methodol.* **32,** 33–67.

Reardon, S. F., Yun, J. T., and Eitle, T. M. (2000). The changing structure of school segregation: Measurement and evidence of multiracial metropolitan-area school segregation, 1989–1995. *Demography* **37,** 351–364.

Sakoda, J. M. (1981). A generalized index of dissimilarity. *Demography* **18,** 245–250.

Schwartz, J., and Winship, C. (1980). The welfare approach to measuring inequality. *Sociol. Methodol.* **11,** 1–36.

Theil, H., and Finizza, A. J. (1971). A note on the measurement of racial integration of schools by means of informational concepts. *J. Math. Soc.* **1,** 187–193.

White, M. J. (1986). Segregation and diversity measures in population distribution. *Popul. Index* **52,** 198–221.

Wong, D. W. S. (1999). Geostatistics as measures of spatial segregation. *Urban Geogr.* **20,** 635–647.

Response Bias

Timothy R. Graeff
Middle Tennessee State University, Murfreesboro, Tennessee, USA

Glossary

apathy bias Bias that results from respondents not caring about the survey or deciding to respond quickly or randomly simply to finish the survey.
counter biasing statement A statement that is given at the beginning of a question and assures respondents that all responses are acceptable and appropriate.
interviewer bias Bias that results from the characteristics of an interviewer causing respondents to censor or alter their answers, or from an interviewer's reactions to a respondent's answers encouraging certain types of answers and inhibiting others.
leading questions Questions that contain words or phrases that suggest or imply a certain answer.
random sampling error A type of survey error that is the result of taking a random sample that includes only a subset of the population.
randomized response questions A series of two questions in which the answer to the first question (known only to the respondent) determines which of two subsequent questions the respondent answers: either an innocuous question or a potentially sensitive question.
reverse-scored items Survey questions that are purposely phrased negatively (reverse phrasing of the majority of other items on the survey).
social desirability bias Bias that occurs when respondents answer based on what is perceived as being socially acceptable, and not the respondent's true state.
uninformed response bias Bias associated with respondents answering questions about which they have no knowledge or experience simply because they feel obligated to respond.

Response bias is a type of survey error that results from respondents intentionally or unintentionally biasing, changing, censoring, or otherwise misrepresenting their true opinions, thoughts, and beliefs when answering survey questions. As a result, respondents' answers to survey questions do not accurately represent their true state. Response bias can result from the topic of the question, the data collection procedures and the characteristics of interviewers, the order of questions on a survey, the phrasing or wording of survey questions, and even from people simply forgetting, wanting to be nice, wanting to appear informed and not ignorant, or wanting to please the interviewer. Because there are no statistical measures of the amount of response bias for a survey, and increasing the sample size generally has no effect on reducing response bias, it is incumbent upon the researcher to identify and anticipate response bias so that the sources of the bias can be eliminated or changed prior to conducting the study and collecting the data. To identify sources of response bias, researchers should consider the reasons why a respondent would lie or report answers to survey questions that do not accurately reflect their true state.

Surveys and Error

The purpose of conducting a survey research project is to measure a characteristic (parameter) of a population of interest. Unfortunately, researchers cannot be 100% confident that the calculated sample statistic (e.g., average age of people in the sample) is exactly the same as the true population parameter (the true average age of people in the population). The difference between the calculated sample statistic and the true population parameter is called survey error, of which there are many sources. Response bias is a type of survey error that results from respondents intentionally or unintentionally biasing, censoring, or otherwise misrepresenting their true opinions, thoughts, beliefs, and responses to survey

questions. For example, many people lie about their age, income, weight, daily activities, and even the types of products they buy. As a result, survey results very often do not accurately estimate true population parameters.

Comparing Response Bias to Random Sampling Error

To understand response bias it is helpful to first understand what response bias is not. Response bias is not random. In fact, many of the issues associated with response bias and strategies for dealing with response bias differ greatly from those associated with random sampling error. Random sampling error is the result of taking a random sample that includes only a subset of a population. Every time a different sample of people from a population is taken the calculated sample statistic will be a different value because different people are being included in the sample each time. Because of random sampling error, sample statistics are just as likely to underestimate the population parameter as they are to overestimate the population parameter. Even though it can never be eliminated (short of taking a complete census of every member of the population), there are statistical measures for estimating and accounting for the amount of random sampling error (plus or minus) that is associated with a sample statistic. The standard error of the mean is used to estimate the amount of random error associated with an average, and the standard error of the proportion is used to estimate the amount of random error associated with a proportion.

$$SE_{\text{mean}} = S/\sqrt{n} \qquad SE_{\text{proportion}} = \sqrt{p(1-p)/n},$$

where S is the sample standard deviation, p is the sample proportion, and n is the sample size. Notice that random sampling error can be reduced by increasing the sample size (the denominator in each equation).

In contrast, response bias is a source of error that usually affects all respondents in a similar manner, resulting in error that is consistently in one direction or the other. With response bias, the difference between the population parameter and the sample statistic can result from the topic of the survey questions; respondents' desire to fit in, appear prestigious, or please the interviewer; and can even result from problems with the data collection methodology, techniques, or procedures. Data collection procedures and the topic of survey questions systematically affect respondents in a manner that causes them to lie or misrepresent their true state. For example, if male university students are interviewed by an attractive female interviewer, they will most likely overestimate how often they exercise to impress the interviewer. Unfortunately, there are no statistical measures of response bias that are comparable to standard error measures of random sampling error. Further, increasing the sample size will generally have no effect on reducing this response bias. Unless the data collection procedures are changed, a larger sample will only result in a greater number of males lying about how often they exercise. The important differences between response bias and random sampling error are summarized in Table I.

Response bias can be reduced by changing the data collection procedures and methodologies. For example, anonymous mail surveys might be best for measuring how often male students exercise. Unless the biasing effect of the attractive female interviewer is removed, males will tend to overestimate how often they exercise. Because there are no statistical measures of the amount of response bias, and increasing the sample size has no effect on reducing response bias, it is imperative that researchers identify and anticipate response bias to eliminate the sources of such bias prior to conducting the study and collecting the data.

Types of Response Bias

Response bias occurs when respondents do not answer truthfully or their answers are not representative of how they really behave, think, or feel. They intentionally or unintentionally misrepresent their true state. To identify potential response bias, ask, "Why would someone lie, misrepresent, bias, or censor their answers?" There are many potential answers to this question. Sometimes response bias is simply due to the topic of the question. Other times response bias occurs because of the data collection procedures and the characteristics of interviewers. Response bias is also caused by the order of questions on a survey and the phrasing of survey questions. And yes, response bias can also result from people simply forgetting, wanting to be nice, wanting to appear informed and not ignorant, and even wanting to please the interviewer.

Social Desirability Bias

People naturally want others to view them favorably with respect to socially acceptable values, behaviors, beliefs, and opinions. Thus, answers to survey questions are often guided by what is perceived as being socially acceptable. For example, even if a person does not donate money to charity, they might report that they have donated. Donating money to charity is the socially acceptable behavior. Social desirability bias can affect responses to questions about whether or not people spank their children, whether or not they recently purchased any fur coats, or even whether or not they voted in recent elections. Research on topics about which there are socially

Table I Comparing Response Bias to Random Sampling Error

	Response bias	*Random sampling error*
What causes the error?	Problems with data collection procedures, question wording, the nature of the survey procedures, characteristics of the interviewer, respondents wanting to please the interviewer, etc.	Randomly selecting members from the population. Each time a sample is taken different people will be included in the sample, and thus, a different sample statistic will be calculated.
What is the effect of the error on sample statistics?	Systematic. Because of response bias, the sample statistic tends to either overestimate or underestimate the true population parameter consistently and systematically across all respondents.	Random. The sample statistic is just as likely to overestimate the population parameter as it is to underestimate the population parameter.
Can the error be estimated statistically?	No. The amount of response bias cannot be estimated statistically.	Yes. The amount of random sampling error can be estimated as the standard error of the mean (when calculating an average) and the standard error of the proportion (when calculating a percentage). $SE_{\text{mean}} = S/\sqrt{n}$ $SE_{\text{prop.}} = \sqrt{p(1-q)/n}$
How can the error be reduced?	Changing the data collection procedures and methodology to reduce or eliminate the source of the response bias. Increasing the sample size has no effect on reducing the amount of response bias.	Increasing the sample size (the denominator in the standard error equations) will reduce the amount of random sampling error in the estimates.

acceptable behaviors, views, and opinions is very susceptible to social desirability bias.

Social desirability bias is by far the most studied form of response bias. Social desirability bias can result from (1) the nature of the data collection or experimental procedures or settings, (2) the degree to which a respondent seeks to present themselves in a favorable light, (3) the degree to which the topic of the survey and the survey questions refer to socially value-laden topics, (4) the degree to which respondents answers will be viewed publicly versus privately (anonymously), (5) respondents' expectations regarding the use of the research and their individual answers, and (6) the extent to which respondents can guess what types of responses will please the interviewer or sponsor of the researcher.

Social desirability bias is often viewed as consisting of two factors, self-deception and impression management. Self-deception refers to the natural tendency to view oneself favorably. Self-deception has been linked to other personality factors such as anxiety, achievement, motivation, and self-esteem. Impression management refers to the situational dependent desire to present oneself in a positive light. This can manifest itself in the form of false reports and deliberately biased answers to survey questions.

There is no standard statistical procedure for measuring the amount of social desirability bias across varying situations, contexts, and survey topics. Nonetheless, researchers have developed scales, such as the Marlowe–Crowne 33-item Social Desirability Scale, and shorter versions of the scale, to identify and measure the presence of social desirability bias in survey results. When a social desirability scale is added to a survey, significant correlations between the social desirability scale and other survey questions indicate the presence of social desirability bias due to respondents' desire to answer in socially desirable ways. Low correlations between the social desirability scale and other survey questions suggest the lack of social desirability bias.

Unfortunately, such social desirability scales cannot be used as standardized measures of social desirability bias across varying situations, settings, data collection procedures, and research topics. Along with the obvious disadvantage of making the survey longer, such social desirability scales often contain questions that respondents might perceive as inappropriate and unrelated to the fundamental purpose of the survey.

Other attempts to measure and reduce social desirability bias include the use of a pseudo lie detector called the bogus pipeline. With a bogus pipeline, researchers tell respondents that the data collection procedures and/or measurement apparatus are capable of identifying when someone is lying. Thus, respondents are more motivated to respond truthfully. Of course, such procedures would be inappropriate for many social research procedures, techniques and data collection methodologies.

The amount of social desirability bias in a survey can vary by (1) mode of contact (anonymous versus

face-to-face interviews or signed surveys), (2) differences in a respondent's home country and culture (respondents from lesser developed countries have been found to be more likely to respond to personality surveys in a manner consistent with existing cultural stereotypes), and (3) the amount of monetary incentive provided to the respondent (respondents receiving larger monetary incentives have been found to exert greater effort in completing a survey and were more likely to respond in a manner that was favorable toward the survey sponsor).

Acquiescence/Yea-Saying Bias

With acquiescence bias a response is based on respondents' perceptions of how they think the researcher wants them to respond, leading to potential demand effects. Respondents might acquiesce and respond favorably toward the idea of a new product because they think that is what a market researcher wants to hear. The manner in which interviewers react to respondents' answers can also give respondents cues regarding what the interviewer wants to hear. The way interviewers provide encouragement to respondents can affect their sense of what they are supposed to do and how they are to respond. Some respondents can also be biased by a natural tendency toward positive responses. Many people exhibit a tendency to agree with statements presented to them rather than disagree. Such tendencies toward positive responses can even vary across cultures.

Prestige Bias

Many people will bias their responses to make themselves appear more prestigious in the eyes of others. Males might overestimate their income if the interviewer is an attractive female. People naturally overestimate social status and occupation, and underestimate their weight, age, and extent to which they suffer from certain medical problems.

Threat Bias

With threat bias, a response is influenced by the extent to which respondents feel threatened by negative consequences associated with certain answers. People might lie about whether or not they have ever shoplifted because of the potential legal ramifications of admitting to shoplifting.

Hostility Bias

With hostility bias, a response is biased due to feelings of anger on the part of the respondent. If respondents feel forced to complete a long survey, or if the interviewer or data collection procedures cause them to get mad, they might deliberately fabricate answers in a retaliatory effort to get even with the researcher. Any affective responses to the research or the data collection procedure itself can introduce confounding variables and additional sources of bias into the research results.

Sponsorship Bias

With sponsorship bias, a response is influenced by respondents' perceptions of the person or organization sponsoring the research. If respondents know who is sponsoring the research, they can often figure out how the research sponsor wants them to respond, leading to demand effects. People might underestimate the amount of red meat they eat on a survey conducted by the American Heart Association. Alternatively, people might overestimate the amount of red meat they eat on a survey conducted by the National Cattlemen's Beef Association. This is similar to acquiescence bias in that people often answer survey questions based on how they think the researcher wants them to respond. This is likely for research measuring low involvement situations or behaviors.

Question Order Bias

Beliefs, feelings, or attitudes that are made salient because of questions asked early in a survey can influence responses to subsequent questions. Questions about teenage delinquency and other social problems associated with young people asked early in a survey can influence responses to subsequent questions about spanking, child discipline, and corporal punishment. Also, the order in which a series of questions are asked on a survey can affect responses to the items. Extreme evaluations (e.g., strongly positive or strongly negative) of the first few items in a list might influence evaluations about subsequent items in the list. Responses to early questions can be used as an anchoring point against which responses to subsequent questions are compared.

Extremity Bias

Many people do not like using the extreme ends of scales and do not like feeling that they are at the extreme, compared to others. For example, if income categories are set up as: less than $20,000; $20,000–$30,000; $30,001–$40,000, etc., a person who earns $19,000 might bias their response by choosing the second category so that they are not in the bottom income category. When asked about the amount of time spent on personal matters while at work, people will naturally not want to be in the top category. One solution to this problem is to include additional response categories that are at the extreme ends of the scale—even if you do not expect anyone to use them. For example, make the income categories: less

than $10,000; $10,000–$15,000; $15,001–$20,000; $20,001–$25,000; and so on. This way someone earning $19,000 is more likely to honestly report their income (check the third income category) without having to be in the bottom income category.

Interviewer Bias

Characteristics of an interviewer can cause people to bias their responses. An attractive interviewer might cause males to bias their responses related to income, weight, or frequency of exercising. However, the effects of interviewer characteristics are minimized in telephone interviews compared to in-person interviews simply because the respondent cannot observe the interviewer. An interviewer's reactions to a respondent's answers can actually encourage certain types of responses and inhibit others. Based on an interviewer's smile, frown, or even the raise of an eyebrow, people can determine what the interviewer wants to hear (what would please them), and then give them those answers. If a person gives their opinion on a controversial topic and the interviewer responds with, "*Hmmm . . . , that's interesting*," the respondent can quickly determine that their answer was not expected or somehow out of line with normal responses. The respondent might censor or bias subsequent answers. Interviewer training can reduce this problem.

Memory Bias

People forget things over long periods of time, and evaluations and opinions can easily change over time. Memory for events can be biased by telescoping errors. Backward telescoping occurs when respondents report that recent events occurred further back in time. Forward telescoping occurs when respondents report that distant events occurred more recently than they really did. Further, evaluations can also change over time. A person might not feel as strongly about the quality of a restaurant meal (either positively or negatively) the more time that lapses between when they ate the meal and when they are asked about it. Evaluations (both positive and negative) can become more moderate as more and more time lapses. What is *terrific* today might be evaluated as only *pretty good* a week from now. Similarly, what is *terrible* today might be evaluated as *not that bad* a week from now. If asking for evaluations, beliefs, perceptions, or attitudes toward events, minimize the amount of time between the person's experience of the event and when they are asked questions about it.

Apathy Bias

This results from respondents who do not care about the survey or who decide to respond quickly or randomly simply to finish the survey. This can happen with very long surveys and those for which respondents have not developed a commitment to complete the survey.

Uninformed Response Bias

Respondents often answer questions about which they have no knowledge or experience simply because they feel obligated to respond. They do this because they do not want to appear uninformed or ignorant. Offering, and making sure that respondents are aware of, a *Don't Know* or *No Opinion* option on the survey tends to reduce this bias. However, many people will still answer questions about which they are uninformed even when they are told it is acceptable to respond with *Don't Know.*

Uninformed responses are especially troublesome in that efforts to increase response rates can have inadvertent negative effects on the validity of research results. Stimulus factors designed to increase item response rates can lead to higher uninformed response rates. Uninformed respondents can be pressured to provide meaningless answers to survey questions. Further, if the purpose of the research is to measure beliefs, attitudes or perceptions, respondents' familiarity with similar sounding attitude objects can also guide their responses. Thus, respondents often unintentionally give uninformed responses simply because they are mistaken about the object being evaluated.

Reducing Response Bias

As listed above, there are many sources of response bias. By being aware of these, researchers can plan their survey procedures to minimize the likelihood that these sources of response bias will have significant effects on their survey results. Strategies for reducing these types of response bias include:

1. Assure respondents of anonymity of responses and privacy with respect to data related to their individual responses (to reduce social desirability bias and threat bias).
2. Whenever possible, use anonymous survey data collection procedures and consider procedures that do not require an interviewer (to reduce social desirability bias, prestige bias, and interviewer bias).
3. Avoid revealing the purpose of the research, the sponsor of the research, or the source of the survey (to reduce acquiescence bias and year-saying bias).
4. Make the survey short, interesting, and easy to complete. Try to get respondents committed to completing the entire survey. Use prompters to help respondents work their way through the survey, such as *the next section will be easier; thank you for your help with*

those questions, please answer a few more questions; or *there are only a few more questions remaining to answer* (to reduce hostility bias and apathy bias).

5. Carefully consider the order of the survey questions and the possible response categories. Try to ask more general questions earlier in the survey, and ask questions about more specific issues, people, events, places, or ideas later in the survey (to reduce question order bias and extremity bias).

6. Reduce the amount of time between a respondent's experience of an event and their responses to questions about that event (to reduce memory bias).

7. Consider using reverse scored items on the survey. Most survey questions are phrased positively. However, some researchers purposely reverse the phrasing of some items so that they are phrased negatively to increase the chance that respondents will read all of the questions, decreasing the likelihood of acquiescence bias, apathy bias, and straight line (column) responding (e.g., apathetically circling a column of Strongly Answer answers). For example, questions one, two, and four below are phrased positively, and question three is phrased negatively. If respondents answer strongly agree (SA) to all four questions, this indicates that they did not carefully read all four questions. They might have assumed that all four questions were phrased positively—leading to a straight line (column) of Strongly Agree answers (see Table II).

A respondent's answer to a reverse scored question must be converted by subtracting the answer's scale value (*X*) from the total number of scale values plus one. In this example, if SD were coded as 1, and SA were coded as 5, a respondent's answer to question 3 would be converted as (6 – X) to place all four scales in the same direction.

However, using reverse scored (reverse worded) items is not without its own limitations. Recent research has demonstrated that the mixture of positive and negative phrased items can lessen a scale's internal consistency. Negative worded items often show lower reliability and weaker item-to-total correlations than positive worded items. When a scale is subjected to factor analysis, reverse scored items often load on a separate factor, thus eliminating the unidimensionality of a scale designed to measure a single construct. Research has also demonstrated that such problems often arise when researching respondents from subcultures, such as ethnic and racial minorities. Differences in cultures and traditions can lead to varying patterns of responses to negatively worded questions, especially when survey questions are translated into languages that employ different methods of representing negatives and contradictions. Thus, including reverse scored items can be particularly problematic for cross-cultural research and surveys conducted in foreign cultures.

Table II Adding a Reverse Scored Item to a Survey

1. I am very articulate.	SD	D	N	A	SA
2. I am very comfortable talking with people I have recently met.	SD	D	N	A	SA
3. I often have difficulty striking up a conversation with people I have just met.	SD	D	N	A	SA
4. I have very good verbal communication skills.	SD	D	N	A	SA

8. Make respondents aware that they can answer any question with *Don't Know* or *No Opinion*. Include questions that measure respondents' level of knowledge about a topic in addition to their attitudes and opinions about a topic (to identify and reduce uninformed response bias).

Response Bias and Question Wording

The wording of survey questions themselves can also lead to response bias. Leading questions are those that contain words or phrases that suggest or imply a certain answer. They lead people to respond in a biased manner. Such response bias is not due to respondents purposely biasing their answers, but rather to the particular way in which a survey question is worded. Consider the following question: "*More and more people are coming to accept cell phone usage in public as socially acceptable behavior. Do you feel that using cell phones in public is socially acceptable behavior?*" Many respondents will respond with *yes* because the question implies that many other people feel it is acceptable behavior. To avoid response bias due to question wording:

1. Avoid suggestions and implications in the question. Asking, "*How pleased are you with the efficient manner in which the Governor dealt with the State's financial problems?*" will lead to biased responses because the question implies that the Governor dealt with the problems efficiently.

2. Avoid emotionally charged words. Asking about possible solutions to the terrorism *crisis* will result in different responses than asking about solutions to the terrorism *problem*, which in turn will lead to different responses than asking about solutions to the terrorism *situation*. Compared to a situation, a crisis usually demands quicker and greater attention.

3. Do not provide reasons or justifications for responses in the question. Let respondents come up with their own reasons for how they answer. Asking, "*Do you support an increase in the state income tax to raise money to help fund education so that the children in our state will be better*

prepared to succeed in life?" encourages positive responses because of the justification given in the question—to help children be better prepared to succeed in life.

4. Use a counter biasing statement. A counter biasing statement is given at the beginning of a question and assures respondents that all responses are acceptable and appropriate. Students might naturally overestimate the amount of time they spend studying. However, including a counter biasing statement such as, "*Some students spend very little time studying, others spend a great deal of time studying, how many hours do you study in a typical week*?" lets them know that all answers are acceptable, even those that indicate very little studying.

Randomized Response Questions

Another strategy to avoid response bias is to use randomized response questions. Randomized response questions consist of two questions in which the answer to the first question (known only to the respondent) determines which of two subsequent questions the respondent answers—either an innocuous question or a potentially sensitive question. Randomized response questions are used when asking about potentially embarrassing, threatening, or otherwise sensitive issues. They reduce response bias by allowing the respondent to answer potentially sensitive questions honestly, because the researcher does not know which of two questions the respondent answered. For example, all respondents are asked an initial question. Respondents do not reveal their answer to the initial question. *Is the last digit in your social security number even or odd?* If it is an even number, they are to answer question A. Alternatively, if it is an odd number, they are to answer question B.

A: *Is the last digit in your home phone number between 5 and 9?*

B: *Do you spank your children?*

___Yes___ No

Respondents are assured anonymity and are more likely to answer truthfully because only they know which second question (A or B) they have answered. Based on simple probability theory, the percentage of respondents who spank their children can be determined as follows: $P(\text{yes}) = [P(\text{question A}) \times P(\text{yes to A})] + [P(\text{question B}) \times P(\text{yes to B})]$.

To illustrate, 35% of the respondents answered *yes* to the second question. Since the probability of answering question A (the last digit in their SSN is even) and the probability of answering question B (the last digit in their SSN is odd) are both 50%, and the probability of having a phone number that ends in a digit between 5 and 9 is also 50%, the percentage of respondents who spank their children (X) is calculated as:

$$0.35 = [0.5(0.5)] + [0.5(X)]$$
$$0.35 = 0.25 + 0.5(X)$$
$$0.10 = 0.5(X)$$
$$0.20 = X$$

Thus, 20% of the respondents spank their children.

Response Bias and Research Conclusions

All that researchers ever really know is what they measure—a person's response to a survey question, the box a person checks on a survey, or the numbers and words they write on a survey. The question that must be asked is, "Why did we observe this particular measurement or these particular differences in measurement across groups or time periods?" There are two possible answers to this question, which refer to two possible explanations for observed measurements.

1. The observed measurement is a real effect. The observed measurement is an accurate estimate of the population parameter. The observed measurements reflect accurate differences in population parameters across groups or time periods.

2. The observed measurement is the result of error. The observed measurement is not an accurate estimate of the population parameter. The observed measurements do not reflect accurate differences in population parameters across groups or time periods.

To illustrate, consider the case of a survey designed to measure the amount of time that parents of elementary school children spend reading to their children and working with their children on school work. The results revealed that parents spend an average of two hours every night reading to their children and working with them on school work. One explanation for this result is that it is an accurate estimate of the population parameter. That is, parents really do spend an average of two hours every night reading to their children and working with them on school work. However, an alternative explanation is that the survey result is due to error. In addition to random sampling error (where the survey result is just as likely to either overestimate or underestimate the true parameter), a significant amount of response bias is likely in this situation. Parents will naturally want to overestimate the amount of time they spend reading to their children and working with them on school work if the survey is administered by the school district (social desirability bias, prestige bias, acquiescence bias, sponsorship bias).

Statistical tests allow for the possibility that observed measurements are due to either real effects or error. The p-value of a statistical test indicates the probability

that an observed measurement is due to error, and not a real effect. But, such p-values indicate only the probability that the observed measurements are due to random sampling error. Even if a statistical test reveals a very low p-value (e.g., $p < 0.01$) indicating that there is less than 1% chance that the observed results are due to random sampling error, the observed measurements could still be the result of systematic response bias that cannot be estimated statistically. Further, the margin of error for a survey (e.g., $\pm 3\%$ for a sample size of approximately 1000) represents the amount of random sampling error that is associated with an estimate. It does not include any error that might be present in the results due to systematic response bias. Increasing the sample size will reduce the amount of random sampling error, but will have no effect on reducing response bias. A larger sample size will not reduce the natural tendency for all parents to overestimate the amount of time they spend reading to their children and working with them on school work.

Conclusions from the results of social research must always be evaluated by recognizing the potential for response bias in the results. Unfortunately, no statistical procedures are available for measuring the amount of response bias in a survey. Researchers must examine the entirety of the data collection procedures and methods to identify potential sources of response bias being introduced into the results.

See Also the Following Articles

Interviews • Survey Questionnaire Construction • Surveys

Further Reading

Ballard, R. (1992). Short forms of the Marlowe–Crowne social desirability scale. *Psychol. Reports* **71,** 1155–1160.

Bradburn, N. M., and Sudman, S. (1988). *Polls and Surveys: Understanding What They Tell Us.* Jossey–Bass, San Francisco.

Chenhall, R., and Morris, D. (1991). The effect of cognitive style and sponsorship bias on the treatment of opportunity costs in resource allocation decisions. *Account. Organ. Soc.* **16,** 27–46.

Converse, J. M., and Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire.* Sage, Newbury Park, CA.

Crowne, D. P., and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* **24,** 349–354.

Fowler, R. J., and Mangione, T. W. (1990). *Standardized Survey Interviewing: Minimizing Interviewer Related Error.* Sage, Newbury Park, CA.

Graeff, T. R. (2002). Uninformed response bias in telephone surveys. *J. Business Res.* **55**(3), 251–259.

Graeff, T. R. (2003). Exploring consumers' answers to survey questions: Are uninformed responses truly uninformed? *Psychol. Market.* **20**(7), 643–667.

Greenberg, B. C., Adbula, A. L., Simmons, W. L., and Horvitz, D. G. (1969). The unrelated question in randomized response model, theoretical framework. *J. Am. Statist. Assoc.* **64,** 520–539.

Groves, R. M., and Kahn, R. L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews.* Academic Press, New York.

Herche, J., and Engelland, B. (1996). Reverse-polarity items and scale unidimensionality. *J. Acad. Market. Sci.* **24**(Fall), 366–374.

Holbrook, A. L., Green, M. C., and Krosnick, J. A. (2003). Telephone versus face to face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Pub. Opin. Quart.* **67**(1), 79–126.

James, J. M., and Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Pub. Opin. Quart.* **54**(3), 346–361.

Jones, E. E., and Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychol. Bull.* **76,** 349–364.

Keillor, B., Owens, D., and Pettijohn, C. (2001). A cross cultural/cross-national study of influencing factors and socially desirable response biases. *Int. J. Market Res.* **43**(1), 63–85.

King, M. R., and Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychol. Market* **17**(2), 79–103.

Morwitz, V. G. (1997). It seems like only yesterday: The nature and consequences of telescoping errors in marketing research. *J. Consumer Psychol.* **6,** 1–29.

Peterson, R. A. (2000). *Constructing Effective Questionnaires.* Sage, Thousand Oaks, CA.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys.* Academic Press, New York.

Steenkamp, J-B. E. M., and Burgess, S. (2002). Optimum stimulation level and exploratory consumer behavior in an emerging consumer market. *Int. J. Res. Market.* **19**(2), 131–150.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science* **185,** 1124–1131.

Risk and Needs Assessments

Christopher Baird
National Council on Crime and Delinquency, Oakland, California, USA

Glossary

base rate The rate at which an event occurs for the entire population studied.
equity The quality of being just, fair, and impartial.
interrater reliability The extent to which different raters or decision makers come to the same conclusion.
needs assessment A formal process ensuring that specific areas of potential needs are assessed periodically for every case referred for services or supervision.
risk assessment A formalized process used to classify individuals into different groups based on observed rates of a specific behavior or outcome experienced by cases with similar profiles.
validity The degree of accuracy attained in measuring what the system is designed to measure; the extent to which a system or premise produces the desired results.

Introduction

In every human service field, there is a critical need to determine the proper level of intervention required to ensure the safety and well-being of the individual served and/or the safety of society in general. Risk assessment, by definition, attempts to identify those cases in which a subsequent negative event (crime, child maltreatment, or disease) is most likely to occur. Social services also recognize that there is an ethical mandate to ensure that decisions that have substantial impact on children, families, and the community are just, consistent, and based on appropriate criteria. Formal risk and needs assessments are designed to accomplish these objectives.

Beginning in the mid-1970s, the use of formal risk and needs assessment instruments in social service agencies expanded rapidly in the United States. Today, these instruments are commonly found in the fields of adult corrections, juvenile justice, and child welfare. In addition, recent efforts to create similar indices for domestic violence initiatives and programs designed to prevent other social problems have shown considerable promise. The increase in the use of risk and needs assessments was driven largely by recognition for the need for more efficient, consistent, and valid decision making in human service agencies.

Frontline workers in social services are asked to make extremely difficult decisions, but there is often significant variance in the skill levels, education, and experience of staff that make these decisions. Consequently, decisions regarding case openings and closings and the level of care, supervision, or custody to be provided have long been criticized as inappropriate, inconsistent, or both. Research has clearly demonstrated that decisions in social service agencies vary significantly from worker to worker, even among those considered to be experts in a particular field. As pressure to make critical decisions in a timely fashion increases, so does the potential for error. Inappropriate decisions can be costly or even tragic, resulting in serious injury or death when agencies fail to properly assess risk.

Risk assessment systems used in social services are formalized methods that provide structure and criteria with the expectation that these assessments will increase the reliability and accuracy of decision making. A variety of systems for estimating risk have been developed over the years. Some experts have expressed concern that the theoretical and empirical support for these systems is inadequate. Because different methods of risk assessment development have been employed over the past two decades, risk assessment procedures vary on a number of dimensions, and the task of comparing one to another is quite complex. Generally, however, there are two basic

types of risk assessment systems. In consensus-based or expert systems, workers assess specific client characteristics identified by the consensus judgment of experts and then exercise their own clinical judgment about future risk. Actuarial systems are based on an empirical study of the relationship between case characteristics and outcomes. The study identifies items/factors with a strong association to observed behaviors and constructs an actuarial instrument that workers score to identify low-, medium-, or high-risk cases.

A substantial number of "head-to-head" tests have been conducted in a number of disciplines during the past five decades. In 1993, Dawes eloquently summarized the results of this research as follows:

> *In the last 50 years or so, the question of whether a statistical or clinical approach is superior has been the subject of extensive empirical investigation; statistical vs clinical methods of predicting important human outcomes have been compared with each other, in what might be described as a "contest." The results have been uniform. Even fairly simple statistical methods outperform clinical judgment. The superiority of statistical prediction holds in diverse areas, ranging from diagnosing heart attacks and predicting who will survive them to forecasting who will succeed in careers, stay out of jail on parole, or be dismissed from police forces.* (p. A40)

In 1996, Grove and Meehl presented an exhaustive review of the literature citing 136 studies ranging from the 1928 Illinois Parole Board Study to the present, concluding that a "great preponderance of studies favor the actuarial approach." They did, however, find a few (8) studies in which clinical judgment outperformed actuarial models.

It is now generally recognized that actuarial or research-based risk assessments are superior to "expert systems." In addressing the efficacy of actuarial systems, three issues are of critical importance: validity, reliability, and equity.

Measuring Validity: Classification versus Prediction

Human services have long struggled with how the validity of risk instruments is best determined. Validity of decision-making systems has traditionally been measured by the degree to which "predictions" about case outcomes are realized. Ruscio (1998) defines validity of child welfare risk instruments in the following manner:

> *The efficacy of your decision policy can be examined through the use of a simple fourfold classification table crossing the optimal outcome for each child (kept at home vs placed into care) with the decision that is reached. There are two types of correct decisions, or "hits," that are possible: True positives are decisions that place children into care when appropriate, and true negatives are decisions that keep children at home when appropriate. There are also two types of incorrect decisions, or "misses," that are possible: False positives are decisions that unnecessarily place children into care, and false negatives are decisions that fail to place children into care when placement is necessary. Based on this classification table, the effectiveness of a decision policy may be evaluated in several ways. For instance, one could determine how many of the decisions to place a child into foster care were correct (true positives divided by the sum of true and false positives); how many children who optimally should have been kept in the home actually were (true negatives divided by the sum of true negatives and false positives); or how many placement decisions, overall, were correct (the sum of true positives and true negatives divided by the total number of cases).* (p. 148)

Although calculation of false positives, false negatives, and the overall percentage of correct predictions is useful in many settings, it may not be the best method for gauging the efficacy of a risk assessment system when the probability of success/failure is substantially different than 50–50. When events are relatively rare, they are inherently difficult to predict. In such instances, simply assuming an event will not occur may produce more predictive accuracy than any attempt to determine where or when occurrence is likely. For example, if recidivism is reported in only 20% of cases released from juvenile correctional agencies, then simply predicting no case opened to services will have subsequent maltreatment reported produces an 80% "hit rate." Obviously, such a prediction, although highly accurate, is of little value. (In essence, the "sensitivity" of the prediction is 0.8, but the specificity—correct identification of those who do fail—is zero.) A valid and reliable risk assessment system may improve the hit rate marginally, but it is possible such a system could result in a higher percentage of false positives and false negatives and still provide the agency with quality information about the relative probability of subsequent maltreatment. Consider the scenario in which a child welfare population ($N = 100$) has a subsequent maltreatment rate of 15%. A risk assessment identifies 25% of the population as "high risk," which, for this example, is equated with a prediction of subsequent maltreatment. Actual versus predicted outcomes are presented in Table I.

In the previous example, an overall hit rate of 82% is attained (3% lower than that attained when all cases are predicted to succeed) with a rate of false positives (subsequent maltreatment) of 56% and false negatives of 5.3%. Despite the high proportion of false positives, cases that were rated high risk experienced maltreatment at a 44% rate, whereas only 5.3% of those rated at lower risk levels had subsequent maltreatment reported. The

Table I Actual vs Predicted Outcomes

	Predicted outcomes	
Actual outcomes	*No subsequent maltreatment*	*Subsequent maltreatment*
No subsequent maltreatment	71	14
Subsequent maltreatment	4	11

ratio of "failures" in the high-risk group to failures in the low-risk group is more than 8:1. Such results help agencies identify which families are more likely to abuse or neglect their children. In addition, 11 of the 15 cases (73.3%) in which subsequent maltreatment occurred were correctly identified (a relatively high rate of specificity).

Many fields, such as juvenile justice, medicine, and adult corrections, have largely abandoned the idea that risk assessment is an exercise in prediction. Instead, terms such as base expectancy rates have replaced discussions of false positives and false negatives. In corrections, for example, high risk does not equal a prediction of failure: In fact, in most correctional systems, more high-risk cases succeed than fail. Instead, high risk simply denotes inclusion in a group of offenders with significantly higher historical rates of recidivism than other groups.

The field of medicine offers similar examples. In cancer research, it is common practice to identify characteristics of malignancies and surrounding tissue and to classify patients as high, moderate, or low risk based on the observed rates of recurrence within a specified time period. A designation of high risk of recurrence does not equate with a "prediction" that the cancer will recur. In fact, most medical professionals carefully avoid making such predictions. As treatment options expand and improve, recurrence-free survival rates have increased to the point where, if false positives and negatives were to be minimized, the best "prediction" for high-risk cases would be "no recurrence." Still, knowing that cases with similar characteristics have experienced a recurrence rate of 10, 25, or 45% helps the doctor and patient select the most appropriate treatment plan.

Conceptually, the use of false positives and false negatives to evaluate risk assessment systems creates another dilemma. Although outcomes are often dichotomous (an event will either occur or not occur), most risk assessment models assign cases to at least three different risk levels (low, moderate, and high). If efficacy is based on predicting an outcome, it must be asked what prediction is being made for cases at intermediate risk levels: Is the designation "moderate risk" a prediction that subsequent maltreatment will or will not occur? We submit that it is neither, but simply the recognition that these cases "recidivate" at higher rates than some and at lower rates than others. Knowing this allows workers to establish appropriate service plans, just as similar information permits doctors and patients to decide on a particular course of action.

Therefore, in evaluating the relative efficacy of risk assessment systems, it is imperative to be very clear about expectations. The terms prediction and classification are often used interchangeably but really connote different expectations. Prediction is more precise than classification. According to "Merriam Webster's" definition, prediction "declares in advance on the basis of observation, experience, or scientific reason." To predict accurately in any field is difficult; to accurately predict human behavior is especially complex because many factors contribute to determining how individuals will act. Classification, on the other hand, is simply "a systematic arrangement in groups or categories according to established criteria." Although accurate prediction would greatly benefit human services and society, it has not proven feasible. The goals of risk assessment are much more modest; it is simply meant to assign cases to different categories based on observed rates of behavior.

New definitions of purposes of risk assessment have emerged in recent years. Silver and Banks (1998) state that "traditional measures of 'predictive accuracy' which carry with them the assumption that dichotomous decisions will be made, have little utility for assessing the potency of a risk classification model" and that "the primary utility of a risk classification model is in providing a continuum of risk estimates associated with a variety of conditions which can be used to guide a range of decision making responses." *(p. 8)*

Two factors are important in measuring the potency of risk assessment instruments: the relative differences in outcomes between risk groups and the actual distribution of cases across risk classifications. In essence, the validity of risk assessment instruments should be measured by the degree to which subgroupings of a meaningful size are identified and the degree to which different rates of "failure" are found for each subgroup.

Actuarial or research-based risk assessment instruments have demonstrated the ability to accurately classify cases to different risk levels in several different fields. Figures 1 and 2 represent results obtained in child welfare and adult corrections, respectively.

As the figures illustrate, well-designed risk instruments can effectively identify groups with very different rates of failure (as defined by each field). In Fig. 2, for example (a risk assessment system used in Nevada), high-risk cases are 6.5 times more likely than low-risk cases to have a new conviction or revocation while on community supervision. This information helps agencies define the level of supervision required to protect the community and help ensure a higher success rate for inmates transitioning to living in the community.

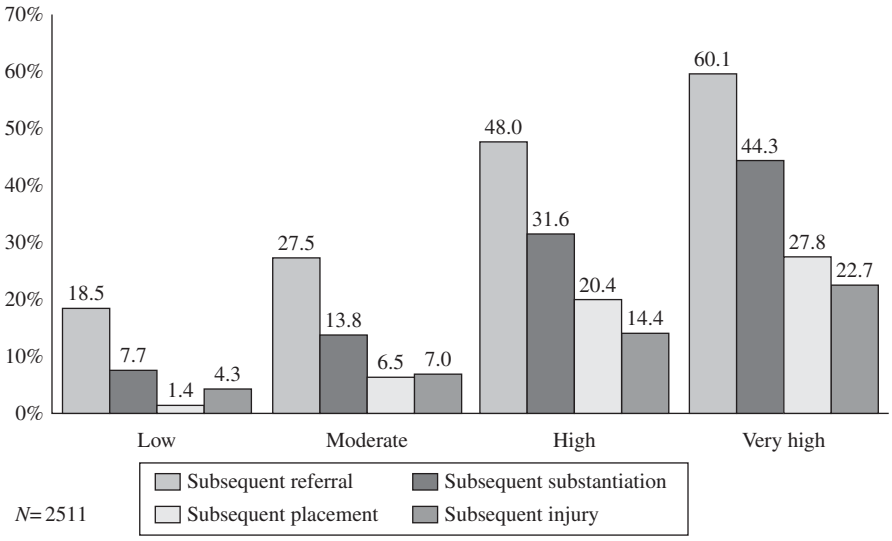


Figure 1 California risk assessment outcomes by risk level.

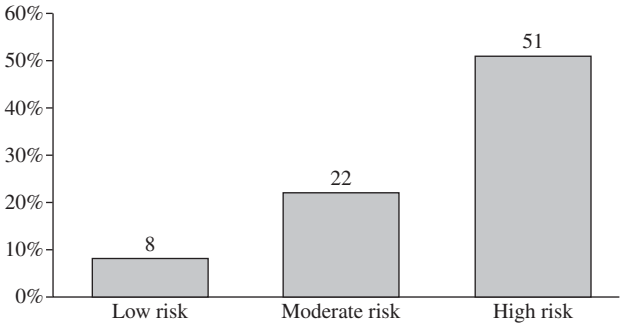


Figure 2 Nevada probation and parole: Subsequent revocation or new felony conviction by risk level.

Reliability and Equity

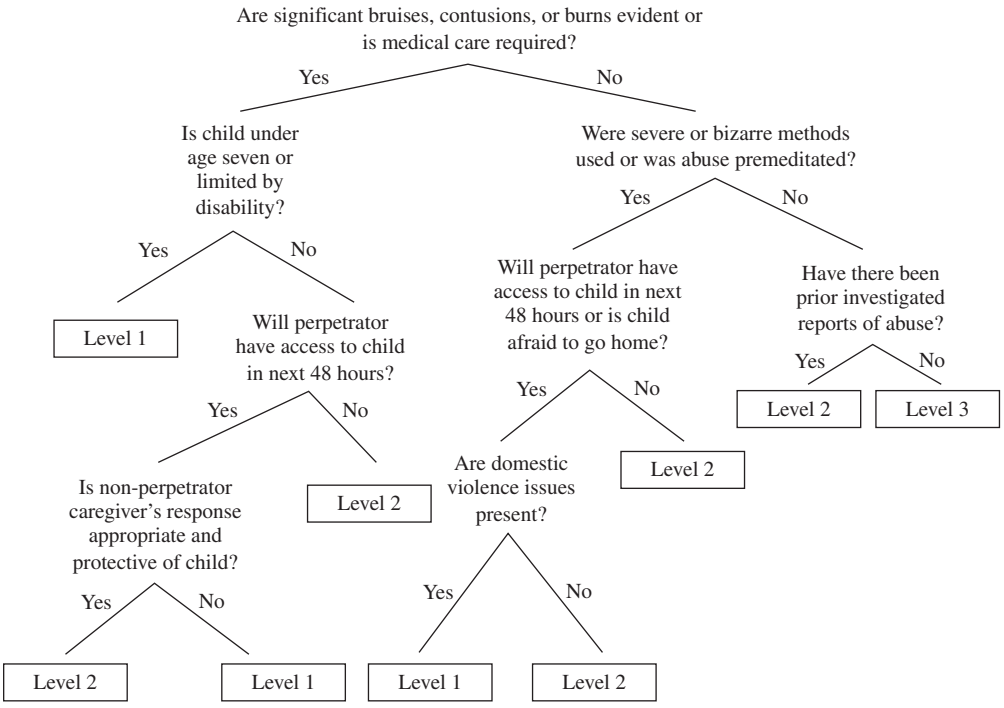
Obtaining valid estimates of the risk represented by each case is obviously critical to social services. To be valid, risk instruments must also be reliable and equitable. As noted earlier, decision making in social service agencies is often related more to who makes the decision than to the characteristics of the case. To enhance consistency, or interrater reliability, a variety of mechanisms, most of which include risk assessment, have been introduced into the decision-making process. In some instances, different measures are combined to guide decisions. For example, sentencing guidelines sometimes consider both risk of reoffending and the severity of the current offense; supervision or service requirements are sometimes established by combining risk and needs assessments in a "service matrix."

In risk assessment, it has been firmly established that a high level of reliability can be attained when risk factors are simple, objective, and well defined. It is also evident that reliability will suffer unless individual risk factors are combined in some fashion to derive a specific risk designation. Two methods for arriving at such a designation are common: a decision tree format (Fig. 3) and an additive index (Table II).

An early study demonstrating why ratings of risk factors must be systematically combined to arrive at a risk classification was conducted by Margaret Blenkner in 1954. Three expert clinical social workers were asked to perform several clinical assessments of 47 clients using information recorded at intake to a private social service agency. After reading each file, making the assessments, and recording them on a data-collection form, each clinician was asked to make a prognosis for future casework success (the cases were closed and outcomes had been established previously by other clinical judges).

In subsequent analysis, five of the assessment measures these clinicians had recorded demonstrated a strong relationship with case outcome, and Blenkner added them together to create a summary score. Although the summary scores demonstrated a very high correlation with case outcomes, the prognoses of the three clinical judges proved unrelated to outcomes and to one another. Meehl observed that

> *Apparently these skilled case readers can rate relatively more specific but still fairly complex factors reliably enough so that an inefficient mathematical formula combining them can predict the criterion; whereas the same judges cannot combine the same data "impressionistically" to yield results above chance.* (p. 108)



Level 1 requires a face-to-face response within 24 hours.
Level 2 requires a face-to-face response within three working days.

Figure 3 Minnesota response priority: physical abuse.

The Blenkner study illustrates why many observers believe clinical judges do poorly in a predictive setting; they differentially select and weigh information about the subject. Despite the fact that the variables in Blenkner's formula had been developed by clinicians and required clinical skill to observe, clinical judges performed very poorly when asked to predict case outcome.

The best measures of interrater reliability are quite simple. First, it is important to determine the rate at which independent raters agree on a risk designation. Next, it is important to adjust "percent agreement" to account for "chance agreement." This is done by applying a statistical measure (usually Cohen's kappa). Although there is no definitive threshold that designates an acceptable level of reliability, kappas <0.3 generally indicate very weak reliability. Although researchers vary on what is considered an adequate kappa (depending on the types of questions posed, the number of potential responses, etc.), a kappa of 0.6 is generally deemed acceptable.

The third measurement issue critical to risk assessment is equity. Equity is a major issue in societies in which minorities are often treated differently by the justice and social service systems. In the United States, for example, there is substantial minority overrepresentation in the juvenile justice, adult corrections, and foster care systems. Inequities in these systems reverberate throughout society, resulting in serious social ills at nearly every level of the social structure. Hence, it is essential that risk assessment systems demonstrate that all groups in a society are treated equitably. To ensure equal treatment, risk assessments must be tested independently on cases of each race/ethnicity. The number of cases assigned to each risk level, as well as observed "failure rates," should be essentially equal for all groups, or it must be clearly evident that differences are due to factors other than race/ethnicity.

Well-designed risk instruments in many social service settings can help control bias and sharply reduce minority overrepresentation. Table III presents child welfare data from Michigan. Essentially, equal proportions of African Americans and Whites are classified to each risk level. The results from Michigan are similar to those from several other jurisdictions in the United States that use actuarial risk instruments. What is clear from data presented here is that proper use of validated risk assessment instruments can result in more equitable decision making and substantially reduce disparity in case disposition.

Table II Nevada Parole and Probation Risk Assessment

Select appropriate answer and enter associated weight in score column. Total all item scores to get total risk score.

	Score
1. Number of residence changes in last 12 months	
a. None	0
b. One	1
c. Two or more	2
2. Current employment	
a. Satisfactory full-time employment (or equivalently occupied) for 1 or more years	0
b. Employed (or occupied) less than full-time/full-time <1 year	2
c. Unsatisfactory employment/unemployed/unemployable	3
3. Alcohol usage problems	
a. No serious problem	0
b. Serious problem/impairs function	1
4. Other drug usage problems	
a. No use	0
b. Some use, no severe disruption of functioning	2
c. Frequent abuse, serious disruption of functioning	4
5. Companions/peer relationships	
a. Good support/influence	0
b. No adverse relationships	1
c. Associations with occasional negative results	2
d. Associations completely negative	4
6. Age at first arrest (adult or juvenile)	
a. 25 or over	0
b. 20–24	2
c. 19 or younger	4
7. Number of prior periods of probation/parole supervision (adult or juvenile)	
a. None	0
b. One or more	3
8. Number of prior probation/parole revocations	
a. None	0
b. One or more	4
9. Number of gross misdemeanors/felony convictions	
a. None	0
b. One	1
c. Two or more	3
10. Convictions or juvenile adjudications for (check all that apply and then score total score total number checked): ___ Theft ___ Burglary ___ Auto theft ____ Robbery	
a. None	0
b. One or two	2
c. Three	3
d. Four	4
11. Number of prior jail sentences	
a. None	0
b. One or more	2

Total Risk Score	
Minimum (0–7)	
Medium (8–16)	
Maximum (17–34)	

Table III Percentage of Families at Each Risk Level in Michigan

Risk Level	*Whites, N = 6651 (%)*	*African Americans, N = 5296 (%)*
Low	10.5	11.3
Moderate	30.7	30.0
High	45.1	46.0
Very high	13.7	12.7

Needs Assessment

In addition to determining the level of risk represented by cases entering the social service system, it is equally important to ensure that case needs are appropriately assessed and corresponding service plans are developed and implemented. Needs assessment is often a companion piece to risk assessment that is used to systematically identify critical issues and help workers plan effective

Table IV Family Strengths and Needs Assessment

Case Name: _____ Date: ___/___/_____

Case Number: _____ Referral Date: _____/_____/_____ Initial Reassess #: 1 2 3 4 5

	Score	Scored
1. Substance abuse		
a. No evidence of problem	0	
b. Abuse creates some problems in family OR caregiver in treatment	3	
c. Serious abuse problem	5	_____
2. Emotional stability		
a. No evidence or symptoms of emotional instability or psychiatric disorder	0	
b. Moderate problems that interfere with functioning	3	
c. Problems that severely limit functioning	5	_____
3. Intellectual ability		
a. No evidence of limitations in intellectual functioning	0	
b. Somewhat limited intellectual functioning	2	
c. Intellectual ability severely limits ability to function	3	_____
4. Health		
a. No known health problems that affect functioning	0	
b. Moderate disability/illness; impairs ability to care for child(ren)	2	
c. Serious disability/illness; severely limits ability to care for children	3	_____
5. Parenting skills		
a. No known/minimal deficits in parenting skills	0	
b. Needs improvement in parenting skills	3	
c. Repeated displays of abusive, neglectful, or destructive parenting patterns	5	_____
6. Environmental		
a. Family has adequate housing, clothing, and nutrition	0	
b. Physical environment presents potential hazards to health or safety	2	
c. Conditions exist in household that have caused illness or injury	3	
d. Family is homeless	4	_____
7. Support systems		
a. Family has available, and uses, external support system(s)	0	
b. Resource limited or have some negative impact or caregiver reluctant to use	2	
c. Caregiver unable to access internal or external resources (skill deficits)	3	
d. Resources unavailable or have major negative impact	4	_____
8. Financial		
a. Family income sufficient to meet needs and is adequately managed	0	
b. Income limited but is adequately managed	1	
c. Income insufficient or not well managed; unable to meet basic needs/responsibilities	2	
d. Family is in financial crisis—little or no income	3	_____

continues

Table IV *continued*

9. Education Literacy	
a. Basic education and functional literacy skills 0	
b. Caregiver marginally educated or literate; creates some problems 1	
c. Functionally illiterate; creates major problems 2	_____
10. Family interaction	
a. Developmental roles/interactions appropriate 0	
b. Moderate communication or behavior problems and/or some inappropriate role functions 2	
c. Serious family dysfunction in communication or behavior patterns, personal boundaries, attachment, and roles 4	_____
11. Child(ren) characteristics	
a. No known emotional, behavioral, intellectual, or physical problems 0	
b. Minor problems, but little impact on functioning 1	
c. Problems in one or more areas that sometime limit functioning 2	
d. One child has severe/chronic problems that result in serious dysfunction 3	
e. Children have severe/chronic problems that result in serious dysfunction 4	_____

Child(ren)'s problem areas (check all that apply):

____ Substance abuse	____ Health/handicap	____ Emotional stability	____ Exceptional education needs
____ School behavior/truancy	____ Support system	____ Intellectual ability	____ Life/social skills
____ Sex abuse issues	____ Assaultiveness	____ Status offending	____ Delinquent behavior
		____ Peers	**Total score** _____

The primary needs of the family are:	Needs level
1. _____	_____ Low (0–10)
2. _____	_____ Medium (11–20)
3. _____	_____ High (21–54)

service interventions. In essence, needs assessment serves the following purposes:

- It ensures that all workers consistently consider each case's strengths and weaknesses in an objective format when assessing need for services.
- It provides an important case planning reference for workers and first-line supervisors that eliminates long, disorganized case narratives and reduces paperwork.
- It provides a basis for monitoring whether appropriate service referrals are made.
- The initial needs assessment, when followed by periodic reassessments, permits case workers and supervisors to easily assess change in case functioning and thus monitor the impact of services on the case.
- It provides management with aggregated information on the issues clients face. These profiles can then be used to develop resources to meet client needs.

Needs assessments can be used at either the individual or the family level, depending on the mission and goals of each social service agency. An example of a family-based needs assessment used in a child welfare agency is presented in Table IV.

Summary

Formalized risk and needs assessments can significantly improve decision making in social services and help agencies target resources to cases that need them most. These assessments increase consistency among workers and help control bias that is evident in systems in which decisions are made by individuals with varying levels of education and experience. They offer a proactive and efficient means for improving service delivery and outcomes.

See Also the Following Articles

Reliability Assessment • Validity Assessment

Further Reading

Baird, S. C., Ereth, J., and Wagner, D. (1999). *Research-Based Risk Assessment: Adding Equity to CPS Decision Making.* Children's Research Center, Madison, WI.

Baird, S. C., and Wagner, D. (2000). The relative validity of actuarial- and consensus-based risk assessment systems. *Children Youth Services Rev.* **22**(11/12), 839–871.

Children's Research Center (1999). *A New Approach to Child Protective Services: Structured Decision Making.* National Council on Crime and Delinquency, Children's Research Center, Madison, WI.

Dawes, R. (1993). Finding guidelines for tough decisions. *The Chronicle of Higher Education*, A40, June.

Meehl, P. (1954). *Clinical versus Statistical Predication: A Theoretical Analysis and a Review of the Evidence.* Minneapolis, University of Minnesota Press.

Murphy-Berman, V. (1994). A conceptual framework for thinking about risk assessment and case management in child protective services. *Child Abuse Neglect* **18,** 193–201.

Rossi, P., Schuerman, J., and Budde, S. (1996). *Understanding Child Maltreatment Decisions and Those Who Make Them.* Chapin Hall Center for Children, University of Chicago, Chicago.

Ruscio, J. (1998). Information integration in child welfare cases: An introduction to statistical decision making. *Child Maltreatment* **3,** 143–156.

Silver, E., and Banks, S. (1998). Calibrating the potency of violence risk classification models: The Dispersion Index for Risk (DIFR). Paper presented at the annual meeting of the American Society of Criminology, November 1998, Washington, D.C.

Sample Design

W. Penn Handwerker
University of Connecticut, Storrs, Connecticut, USA

Glossary

autocorrelation Correlation of case residuals, ordinarily among cases adjacent in time or space.
case A member of a population, the primary sampling unit on which one makes measurements.
confidence interval (limits) Interval identified by specific upper and lower bounds, or limits, that contains the population parameter a specific proportion of the time (e.g., 95%) for the same variable measured for all possible samples of a specific size drawn from a specific population at a specific time.
enumeration (listing) unit Spatially discrete sampling unit that contains sets of cases.
ethnography, ethnographic analysis Identification and description of cultures based on analysis of behavioral and cognitive similarities among cases.
parameter The value of a variable that characterizes a population.
point estimate One's single best guess about a parameter of interest.
population The set of cases to which one generalizes sample findings.
sampling distribution The frequency distribution of all possible statistics of a given kind (e.g., means) calculated from measurements of a specific variable made on all possible samples of a specific size drawn from a specific population at a specific time, the average variability in which is summarized by a number called a standard error rather than a standard deviation.
sampling frame A list of all cases (members of the population) either individually or by enumeration unit.
statistic The value of a variable calculated from sample data that constitutes a point estimate of a parameter.

Sample design refers to the means by which one selects the primary units for data collection and analysis appropriate for a specific research question. These units may consist of states, cities, census enumeration districts, court records, cohorts, or individuals. Irrespective of the kind of unit, data are always collected at specific times and places about a specific set of cases (a sample) that comprises a selected subset of a larger set of cases, times, and places (a population). Answers to research questions thus take the form of inferences from samples to populations. A useful sample design warrants the conclusion that one's inferences are both accurate and appropriately precise.

Research Questions Can Target Variables or Cultures

Decisions about sample design depend on the research question. Some research questions call for answers that come from the analysis of variables:

- What is the percentage of students at Humboldt State University in fall 1987 who believe that we need to preserve wilderness areas?
- Has the degree to which Barbadian couples share child care and household responsibilities changed between 1950 and 1980?
- Does traumatic stress experienced in childhood increase the risk of depression in adulthood?

Other questions call for answers that come from the analysis of cases, or ethnographic analysis:

- Is there a pattern of behavior that constitutes a culture of drug use or a cultural model of what constitutes sexual behavior?

Sampling and Inference for Variables

The set of cases, times, and places from which we sample is a population (or universe). Populations are characterized

by values called parameters. Each question posed earlier identifies a specific parameter and the pertinent population. For example, the population in the first question consists of all students enrolled in Humboldt State University in the school year 1987–1988; the parameter is the percentage who believe we need to preserve wilderness areas. Instead of the number of students enrolled in a university, the population might consist of the people who live in St. Johns, Antigua, or a cohort of women living in Brazil. Instead of a percentage, the parameter might be an incidence rate (e.g., the incidence of child abuse), an average (e.g., the average number of children born to women by age 50), or another univariate statistic. Answers to questions such as these ordinarily come from synchronic observational studies that employ a cross-sectional design. Accurate and precise answers depend on samples of sufficient size that employ random selection criteria.

The population in the second question consists of all the couples living in Barbados between 1950 and 1990; the parameter is a measure of historical differences in the variable "the degree to which Barbadian couples share child care and household responsibilities." Instead of the couples who live in a specific geographical region during a specific historical period, the population might consist of all countries in the world from 1960 to 1990 or court records for the past 12 months. A measure of historical differences remains the parameter of interest, but the variable that may or may not change may be infant mortality or the proportion of drug use charges made against people from ethnic minorities. Answers to questions such as these come from diachronic studies that employ a retrospective or prospective panel or time series design. Accurate and precise answers depend on samples of sufficient size that employ random selection criteria.

The population in the third question consists, potentially, of all people at all times and places; the parameter is a function that maps "traumatic stress experienced in childhood" onto "depression in adulthood." Answers to questions such as these come from a variety of both synchronic and diachronic research designs applied to specific sets of people at specific times and places. Accurate and precise answers depend on samples of sufficient size that employ random selection criteria.

Sampling and Inference for Ethnographic Analysis

The population in the fourth question, like the third, consists potentially of all people at all times and places; the parameter, however, is a construct that summarizes behavioral and cognitive similarities among a set of people. This shift in the meaning of the parameter means that answers to ethnographic questions come from the analysis of autocorrelation among cases.

Cultures consist of recurrent patterns of behavior rationalized by shared domain-specific schemes. They exert effects because the sensory inputs they generate constitute an environment to which people must respond. The effects of cultures come from how people respond to ecological contingencies that influence the consequences of behavior. In short, people construct the cultures in which they participate, and make them evolve, through social interaction. The socially constructed properties of cultures means that any one person who knows about a particular culture participates with other experts in its construction and evolution. Cultures thus inescapably embody spatial and temporal autocorrelation. What one cultural participant does or tells you will correspond closely to what another cultural participant does or tells you. The errors you make in predicting what one cultural participant will do or say will correspond closely to the errors you make predicting what another cultural participant will do or say.

This means that a random sample of people does not constitute a random sample of culture. The culture of an individual consists of configurations of cognition, emotion, and behavior that intersect in multiple ways the cultures of other individuals. Hence, random samples of individuals will yield a random sample of the intersecting configurations of cognition, emotion, and behavior (i.e., the cultures) in a population. However, random samples (defined by case independence) of cultural phenomena (which necessarily contain case dependence) cannot exist: they constitute mutually exclusive alternatives. To identify and describe cultures, ethnographic analysis aims to accurately characterize spatial and temporal autocorrelation. Like answers to the third question, answers to ethnographic questions come from a variety of both synchronic and diachronic research designs applied to specific sets of people at specific times and places. Accurate and precise answers depend on samples designed to actively search for cultural variation that comes from specific forms of variation in life experiences. Sample size depends on the degree of similarity among cases.

Selection Criteria Provide the Ingredients for Sample Designs

Cases can be selected on the basis of one or more of six criteria:

1. Availability
2. Fulfilling a size quota
3. Random (or known probability) selection
4. Case characteristics
5. Presence in specific enumeration units
6. Presence along transects or at specific map coordinates

All samples that utilize random (or known probability) selection are called probability samples. If one does not employ random selection, one produces one of four different forms of nonprobability samples.

Nonprobability Samples

If you select a predetermined number or proportion of cases with specific case characteristics, or from specific enumeration units, transects, or sets of map coordinates, you produce a quota sample. If you select cases on the basis of case characteristics to acquire specific forms of information, you produce a purposive (judgment) sample. If you select cases simply because they will participate in your study, you produce an availability (convenience) sample. If cases become available because one case puts you in contact with another, or other cases, you produce a snowball sample.

Probability Samples

Probability samples are distinguished from nonprobability samples because the former exhibit known sampling distributions that warrant parameter estimation with classical statistical tests (e.g., chi-squared, *t* test, and *F* ratio). By convention, we identify parameters with Greek letters, such as β (beta), α (alpha), ε (epsilon), ρ (rho), and σ (sigma). Samples, in contrast, yield statistics. By convention, we identify statistics with Latin letters and words (e.g., *b*, median, percentage, and mean). Each statistic constitutes a point estimate of a parameter, which is one's single best guess about the value of the parameter.

Statistics constitute point estimates of parameters because samples of populations cannot perfectly replicate the properties of the populations from which they derive. Every sample yields different findings, and every statistic contains three sources of error (construct, measurement, and sampling). Construct error derives from trying to measure a construct that imperfectly fits the culture or cultures found in the population studied. Measurement error derives from imperfections in the means by which a value is assigned to an observation from a set of possible outcomes. To the extent to which significant construct and measurement errors can be ruled out, the difference between a specific statistic and the population parameter constitutes sampling error in that specific sample. Measurements of the same variable made on a large number of samples of the same size drawn from the same population exhibit a characteristic sampling distribution of errors around the parameter. Some statistics underestimate the parameter, whereas others overestimate the parameter.

Sampling errors may reflect chance or bias. Sampling errors that derive from chance exhibit characteristic distributions. Many such sampling distributions (the family of *t* distributions and the normal distribution) are symmetrical and are summarized by a mean of 0 and a standard deviation of 1. The average amount of error in a sampling distribution is called the standard error rather than standard deviation to distinguish sampling distributions from the frequency distributions of the variables studied in social science research.

Although some statistics underestimate the parameter and others overestimate it, when cases are selected independently and have the same probability of inclusion in any one sample, sampling errors come solely from chance. When this condition applies, the sampling distribution of all possible statistics reveals that most statistics come very close to the parameter, and the average amount of sampling error is 0. With statistics that exhibit a normal sampling distribution, for example, 68% of all sample statistics fall within ± 1.00 standard errors of the parameter, and 95% of all sample statistics fall within ± 1.96 standard errors of the parameter.

Small samples contain large amounts of sampling error because randomly selected extreme values exert great effects. Large samples contain small amounts of sampling error and thus estimate parameters very precisely. Sample precision is measured by the size of confidence intervals. Accurate samples yield confidence intervals that contain the parameter a given proportion (usually 95%) of the time. Statistical test findings apply to samples of all sizes because they incorporate into their results the degree of sampling error contained in samples of different sizes. Confidence intervals for small samples are wider than confidence intervals for large samples, but statistics from both large and small samples estimate parameters equally accurately.

This generalization holds only for statistics from samples that are reasonably unbiased. Unbiased samples are those in which all members of the population have an equal chance of selection. The only way to reliably obtain a reasonably unbiased sample is to employ the random selection criterion.

Simple Random Samples

Simple random samples (SRSs) constitute the reference standard against which all other samples are judged. The procedure for selecting a random sample requires two steps. First, make a list of all members of the population. Second, randomly select a specific number of cases from the total list. Random selection may rely on tables of pseudo-random numbers or the algorithms that generate uniform pseudo-random number distributions in statistical analysis software such as SYSTAT. One may sample with or without replacing cases selected for the sample back into the population. Sampling without replacement produces unequal probabilities of case selection, but these are inconsequential except with very small populations. More important, even SRSs overestimate the true

standard error by the factor, $\sqrt{N/(N-n)}$. Application of the finite population multiplier, $\sqrt{(N-n)/N}$, will produce correct standard errors. The importance of this correction increases as the ratio of sample size (n) to population size (N) increases.

Random Systematic Samples

Random systematic samples (RSSs) constitute a variation on SRSs in which random selection of a starting point is substituted for random selection of all cases. For example, to select an RSS of 20% of a population, randomly select a number between 1 and 5, make your first case the one with the randomly selected number, and select every fifth case thereafter. To select an RSS of 5% of a population, randomly select a number between 1 and 20, make your first case the one with the randomly selected number, and select every 20th case thereafter.

Periodicity in a list of population members introduces significant bias into RSSs. In the absence of periodicity, and with a known population size, to determine a sampling interval (k), divide the size of the population (N) by a desired sample size (n). RSSs produce unbiased samples when k is an integer. The bias introduced when k is not an integer is inconsequential with large populations. However, if you know the size of the population, the following procedure always yields unbiased estimates:

1. Randomly select a number (j) between 1 and N.
2. Express the ratio (j/k) as an integer and a remainder (m).
3. When m equals 0, select the case numbered k as your first sample element; when m does not equal 0, select the case numbered m as your first sample element.

Stratified, Cluster, Transect, and Case-Control Samples

All other probability samples incorporate SRSs or RSSs into the selection process. Stratified samples, for example, consist of a series of simple random or random systematic samples of population sectors identified by case characteristics (e.g., age, class, gender, and ethnicity) or combinations of characteristics (e.g., old and young women, and old and young men). Disproportionally stratified samples employ a quota criterion to oversample population sectors that might otherwise be insufficiently represented in the final sample. Cluster samples consist of samples in which cases are selected from SRSs or RSSs of enumeration units that contain sets of cases, such as households, hospitals, city blocks, buildings, files, file drawers, or census enumeration districts. Probability proportional to size samples are cluster samples in which the number of cases selected from specific enumeration units matches a quota proportional to the size of unit relative to the entire population. Transect samples consist of samples in which cases or enumeration units are selected from SRSs or RSSs of units that lie along randomly drawn transects or randomly selected map coordinates. Case-control samples consist of a set of purposefully (judgmentally) identified cases, a small set of which may be selected randomly, plus a set of randomly selected controls. This sampling procedure originated in epidemiology, in which cases are characterized by specific health conditions not experienced by controls. However, the procedure is readily generalizable by defining cases and controls by reference to a binary variable that distinguishes cases with a specific experience from controls without that experience.

How to Create Practical and Useful Sampling Designs

Practical sampling designs balance feasibility, cost, and power. What constitutes a useful sampling design varies with the properties of the parameter to be inferred and the data collection context.

Inferences about Variables

When one's research question calls for an inference about a variable's parameter, differences between parameters, or the parametric relationship between variables, accurate and precise answers depend on samples of sufficient size that employ random selection criteria. However, the primary types of such samples (SRS, RSS, stratified, cluster, transect, and case–control) vary dramatically in their feasibility, cost, and power for the issue at hand. For example, SRSs cannot be drawn in the absence of a complete list of primary sampling units. Such lists commonly do not exist; it may not be possible or cost-efficient to create one. If cases with a specific experience can be distinguished from controls without that experience, a case-control sample may be selected by SRS. Even when SRSs can be drawn, however, it may not be cost-efficient to search out and contact independently selected primary sampling units.

Primary sampling units almost always can be identified either within a spatially bounded region or by enumeration units. When the population occupies a spatially bounded region, samples based on transects or sets of map coordinates are efficient choices. When a comprehensive list of enumeration units can be assembled efficiently, cluster samples of one kind or another become both feasible and relatively cheap. However, both cluster samples and SRSs exhibit large standard errors compared to stratified samples. Stratification thus makes it possible to achieve the same power with smaller sample sizes.

Power refers to one's ability to precisely identify a parameter, to detect differences in a parameter over time or space, or identify the parametric influence of one variable on another, if the effect is real rather than due to chance. Sample design determines the population to which one can validly generalize, but you will waste your time if you don't put in the effort to select a sample large enough to estimate parameters with requisite precisions.

Power varies with the risk of a type I or α error that one is willing to accept, sample size, and the size of the effect that one wants to be able to detect. The probability of making a type II or β error—of not detecting a real relationship between variables—is 1-power. For a fixed sample and effect size, when α is lowered, β simultaneously rises. When one wants to rigorously avoid concluding, for example, that traumatic stress in childhood influences the risk of depression in adulthood when, in fact, it does not, one might set α at 0.01. But, how power goes up and β goes down varies dramatically with the size of the effect.

Figure 1 illustrates the interdependencies between sample size, power, and effect size, when α is 0.01 (assuming the standard errors of SRSs). Figure 2 illustrates the interdependencies between sample size, β, and effect size, when α is 0.01 (assuming the standard errors of SRSs). As sample size increases, the ability to detect a real relationship (power) increases, and the possibility that it will not be detected (β) decreases. However, the way in which power increases and β decreases varies dramatically with the size of the effect. If the real shared variance between variables is approximately 0.06 (a Pearson's r of 0.25), it would be missed approximately half the time even with a sample of 100 cases. If the real shared variance between variables is approximately 0.25 (a Pearson's r of 0.50), it would be missed only approximately 9% of the time with a sample of only 50 cases and 1% of the time with a sample of 75 cases. In contrast, if the real shared variance between variables is approximately 0.76 (a Pearson's r of 0.873), one could expect to miss it only approximately 3% of the time even with a sample of 10 cases and not at all with 15 cases. Decisions about sample size ordinarily seek to be able to detect the smallest important effect 80% of the time or better. Power analyses may be conducted by specialty software or by power routines that come with the major statistical software packages.

A useful balance of feasibility, cost, and power usually comes in the form of a multistage sample design. Table I shows a multistage design appropriate for a study of drug- and sex-related HIV risk behaviors among recent Latino migrants to the United States from Mexico and Central America. The context for such a study illustrates many of the difficulties that sample designs must resolve.

First, the size of the population is unknown, which eliminates the choice of an SRS. Second, it is unknown who among the population engages in drug- or sex-related

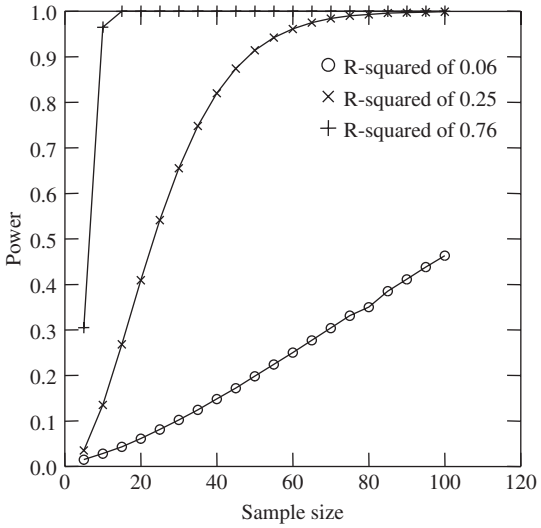


Figure 1 Relationship between power and sample size for effects of different sizes.

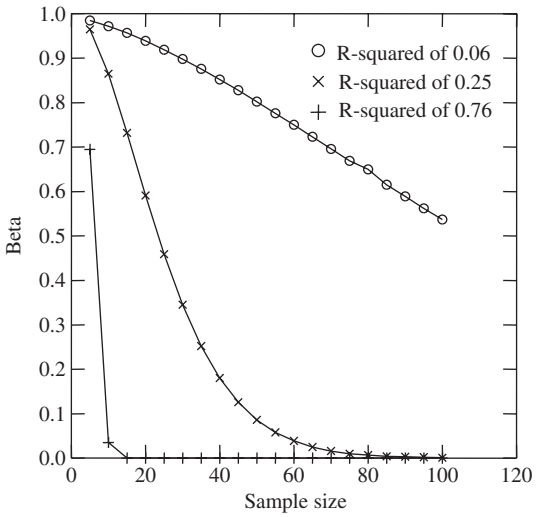


Figure 2 Relationship between beta and sample size for effects of different sizes.

HIV risk behavior, which eliminates the choice of a case-control sample. Third, the region in which migrants live is clearly delimited, but the target population of migrants may comprise a tiny proportion of the total number of people living within the region. Migrants frequently live in locations highly dispersed among the vast majority of the region's population; many may be effectively hidden from conventional enumeration units (e.g., households) because they change residence frequently. When these conditions apply, transect or map coordinate samples would constitute costly sample design choices. Conventional ways of thinking about cluster sampling do not apply.

Table I Multistage Sampling Design for a Cross-Sectional Observational Study of Drug- and Sex-related Risk Behaviors among New Latino Migrants

Stage I: Comprehensive list of enumeration units				
Stage Ib: Stratification of enumeration units				
Stage II: Random systematic sample of each kind of enumeration unit				
Stage IIb: Stratification by ethnicity and gender				
Ethnicity	*Mexican*		*Central American*	
Gender	*Men*	*Women*	*Men*	*Women*
No. of interviews	Contingent on power analysis	Contingent on power analysis	Contingent on power analysis	Contingent on power analysis
Stage III: Random systematic sample of primary sampling units				

Fourth, it remains possible to assemble without undue cost a list of the unconventional enumeration units that would include even those migrants who otherwise remain hidden. These units might include street locations, farms, bars, community agencies, churches, and significant time differences for each. If different types of locations and times attract cases with different characteristics, the comprehensive list of enumeration units may be usefully stratified into different kinds of units based on those characteristics.

Fifth, RSS is easy to apply and does not require a comprehensive list of primary sampling units (cases). When cases are distributed randomly, RSSs exhibit the same standard errors as SRSs. Stratification of enumeration units by case characteristics orders the cases with regard to the variables studied. With ordered cases, RSSs exhibit lower standard errors (greater power) than SRSs. Further stratification on the basis of ethnicity and gender may or may not be cost-efficient relative to the gain in power it would yield. In the absence of stratification, explicit measurement of internal validity confounds implements a posttest-only control group research design that substitutes for random assignment the explicit measurement of internal validity confounds. RSSs from each kind of enumeration unit and RSSs of cases from each randomly selected enumeration unit complete the multistage design.

An appropriate power analysis focuses on the objective of the proposed study to test hypotheses about circumstances that increase or decrease the likelihood of engaging in specific HIV risk behaviors. Given an alpha level of 0.05, a two-tailed test, and the assumptions of SRS, the analysis would indicate the sample size necessary to detect effects of specific independent variables 80% of the time, or the power of tests based on different sample sizes. Table II shows how power would vary for the study in question with variation in sample size, the ratio of the reference and response groups, and varying effect sizes using binary independent variables. A sample size of up to 1204 cases would be required to detect a 50% increase in the likelihood of a given risk behavior (or a 33% reduction in the likelihood of a given risk behavior) if the ratio of reference and response groups was 20/80. However, a 500-case sample would have good to excellent power to detect an odds ratio ≥ 1.76 (or ≤ 0.57) whether the ratio of reference and response groups approximates 60/40 or 80/20. A 600-case sample does not appreciably improve the power of these analyses. An argument that we could both anticipate effects of this size and that smaller effects would not be of clinical or substantive significance at the current time—or not worth the expense of doubling the sample size—warrants a total sample size of approximately 500 cases.

Table II Power for Logistic Regression Tests with Varying Sample Size, Ratio of Reference to Response Group, and Size of Effect (Odds Ratio) with a Binary Independent Variable

Sample split	*Odds ratio*	*Power, N = 400(%)*	*Power, N = 500(%)*	*Power, N = 600(%)*
80/20	2.07 or 0.48	82	89	94
	1.76 or 0.57	62	71	78
	1.50 or 0.67[a]	37	44	50
60/40	2.07 or 0.48	94	97	99
	1.76 or 0.57	78	86	92
	1.50 or 0.67[b]	50	59	67

[a] Sample size necessary to detect an odds ratio of 1.50 or 0.67 at a power of 80% is 204.
[b] Sample size necessary to detect an odds ratio of 1.50 or 0.67 at a power of 80% is 809.

By employing random selection criteria and sample sizes determined by a power analysis, the sample design in Table I allows accurate and reasonably precise estimates of parameters bearing on drug- and sex-related HIV risk behavior for a specific population of Mexican and Central American migrants to the United States. That the total population of cases came to be explicitly known only during the course of case selection and data collection does not bear on the validity of inferences from the sample to the population.

Inferences about Cultures

When a research question calls for an answer in the form of a construct that summarizes behavioral and cognitive similarities among a set of people, accurate and precise answers depend on samples designed to actively search for cultural variation that comes from specific forms of variation in life experiences. When generalizing about cases rather than variables, the meaning of power changes, and sample size depends on the degree of similarity among cases.

For example, in ethnographic analyses, power refers to the reliability and validity of inferences about the content of the behavioral and cognitive similarities among cases (the culture or cultures they share). Important work by Susan Weller has shown that estimates of both the reliability and the validity of those inferences come from the application of the Spearman-Brown prophesy formula to the average level of similarity among cases. If the average level of similarity is 0.50, 9 cases will yield a reliability coefficient of 0.90 and a validity coefficient of 0.95. Only 18 cases will yield a reliability coefficient of 0.95 and a validity coefficient of 0.97. If the average level of agreement is 0.60, only 12 cases are needed for the same level of reliability and validity. As the level of similarity increases to 0.70, 0.80, and 0.90, the number of cases (sample size) declines to 8, 6, and 3 cases, respectively. At an average level of agreement of 0.90, 3 cases yield a reliability coefficient of 0.96 and a validity coefficient of 0.99.

Sample designs for ethnographic analysis thus differ in important respects from sample designs for variable analyses. They do not require large sample sizes, and they do not depend on random selection. Useful sample designs for the study of cultures stratify the population by contrasting life experiences that may produce cultural differences; employ judgmental selection of key informants and critical cases; and select other cases based on their availability, either out of convenience or through a snowball procedure. Sample size for specific strata is set by quota, depending on the average level of agreement. Efficient sample designs track levels of agreement and expand sample sizes and change stratification criteria consistent with levels of agreement and identified cultural boundaries.

The multistage sample design in Table I may be usefully employed for an ethnographic study of the same population and the same topic, with the following important changes in procedure:

1. The list of enumeration units (stage I) may be assembled in the process of conducting informal and semistructured interviews or observations. Indeed, an ethnographic component to a research design allows one to assemble such lists for a later survey with which to make inferences about variable parameters, to assess the importance of stratifying such a list, and to assess and avoid construct errors that might otherwise find their way into a study's measuring instruments.
2. Purposive (judgmental) selection should substitute for RSS selection in stage II.
3. Selection of cases from specific enumeration units in stage III should employ a combination of purpose (judgment), availability (convenience), and snowball criteria, rather than RSS criteria.
4. Selection of cases must include selection of the social relations of those cases.
5. Data collection (i) begins with the purposeful or convenient identification of cases (and their social relations) and (ii) initiates an iterative process that results in the construction of the multiple stages shown in Table I.

Informal and semistructured interviews and observations are designed to actively search for sources of cultural difference. They elicit information on the adequacy of the initial stratification criteria. Identification of people who think and act differently leads to interviews with cases selected on the basis of new stratification criteria. Because people construct cultures and make them evolve, valid, reliable generalization is restricted to the population that exhibits those specific life experiences and to the immediate future. This makes it particularly important for ethnographic studies to explicitly measure potentially pertinent life experience variables.

Different research goals require different stratification criteria. Demeaning remarks directed at and restricted opportunities provided for members of ethnic minorities (e.g., Native Americans) by members of a dominant ethnic majority constitute two forms of traumatic stress experienced in childhood that may exhibit dramatic effects on later behavior. Table III shows a stratified sampling design for a retrospective study of continuity and change in

Table III Stratified Sampling Design for a Retrospective Study

	Age 20 in historical period X											
	1960s				*1980s*				*2000s*			
Gender	Women		Men		Women		Men		Women		Men	
Native (N)–nonnative (nN)	N	nN	N	nN	N	nN	N	nN	N	nN	N	nN
No. of interviews contingent on average level of similarity	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18

Table IV Stratified Sampling Design for a Case–Control Study

	Cases				*Controls*			
Gender	Women		Men		Women		Men	
Age	<20	>20	<20	>20	<20	>20	<20	>20
No. of interviews contingent on average level of similarity	4–18	4–18	4–18	4–18	4–18	4–18	4–18	4–18

the meaning of social interaction between members of majority and minority ethnic groups. People in their 60s in 2000 can tell what they remember about native–nonnative interaction in the 1960s, when they were in their 20s. People in their 40s in 2000 can tell what they remember about native–nonnative interaction in the 1980s, when they were in their 20s. People in their 20s in 2000 can tell what they remember about native–nonnative interaction during that historical period.

Table IV shows a stratified sampling design for a case-control study, a design widely applicable to outcomes evaluation research. Evaluation outcomes research tests the efficacy of interventions designed to induce specific forms of cultural change. Judgments about the efficacy of interventions require information on whether or not and the degree to which people who started with one culture ended with another. Cultural differences between participants (cases) and nonparticipants (controls) that cannot be explained by other potential internal validity confounds, such as gender and age, constitute evidence of a successful intervention.

How to Use Sample Design to Avoid Selection Bias

Selection bias alters the population to which one may validly generalize. It may make it impossible to answer one's research question. Unlike other aspects of sample design, the effects of selection bias do not vary with whether one's research question calls for an analysis of variables or cases (ethnographic analysis).

Nonresponse cases and missing data constitute important and, perhaps, the most common sources of selection bias. The severity of these sources increases as the level of nonresponse and missing data increases. However, solutions to these problems come from how a survey is implemented (including recruitment, training, and oversight of interviewers or observers), and from the design of data collection instruments. Solutions do not come from sample designs.

Sample design contributions to selection bias come from the inclusion or exclusion of important components of the population sampled. For example, health studies that draw on clinic samples miss all the cases who do not attend the clinic in question or, more generally, do not seek care during the study. Studies of entrepreneurship that exclude failed entrepreneurs can validly generalize only to successful examples. A study that seeks to evaluate trends will require a sample design that allows at least three data points, not just "before" and "after."

Final choices on a sample design must be based on careful examination of the possibility that a specific design might exclude an important subset of cases and how that exclusion may affect the findings, one's ability to generalize to one's target population, and even one's ability to answer the original research question.

See Also the Following Articles

Clustering • Confidence Intervals • Data Distribution and Cataloging • Ethnography • Randomization • Spatial Autocorrelation

Further Reading

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *Am. Soc. Rev.* **48,** 386–398.

Bernard, H. R. (2001). *Research Methods in Anthropology.* 3rd Ed. AltaMira, Walnut Creek, CA.

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (1994). *Distance Sampling: Estimating Abundance of Biological Populations.* Chapman & Hall, New York.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* 2nd Ed. Erlbaum, Hillsdale, NJ.

Handwerker, W. P. (2001). *Quick Ethnography.* AltaMira, Walnut Creek, CA.

Kish, L. (1963). *Survey Sampling.* Wiley, New York.

Levy, P. S., and Lemeshow, S. (1999). *Sampling of Populations.* 3rd Ed. Wiley, New York.

Schulz, K. F., and Grimes, D. A. (2002). Case-control studies: Research in reverse. *Lancet* **359,** 431–434.

Sudman, S. (1976). *Applied Sampling.* Academic Press, New York.

Weller, S. C. (1987). Shared knowledge, intracultural variation, and knowledge aggregation. *Am. Behav. Sci.* **31,** 178–193.

Sample Size

Chirayath M. Suchindran
University of North Carolina, Chapel Hill, North Carolina, USA

Glossary

censoring When individuals who are followed for a fixed duration of time do not experience the outcome of interest by the time the observation period ends. Such observations are called censored observations. For a right-censored observation, all that is known is that the time to the event occurrence exceeds the period of observation.

effect size The difference detected in the end point in a study; depending on the end point, the effect size may be means, regression coefficients, odds ratios, or hazards ratios.

intraclass correlation Used as a measure of homogeneity among elements of a cluster; can be viewed as the correlation among pairs of observations within a cluster.

Type I error An error that occurs when the experimental situation declares that the specified difference is real, when, in fact, this is not true. The probability of a Type I error is known as the level of significance.

Type II error In experimental studies, failure to detect the specified difference (the second kind of error). The power of a statistical test is then the conditional probability that the null hypothesis is correctly rejected when it is false (complement of the second kind of error).

A well-designed scientific study must determine, at the outset, the sample size; the sample must be large enough to provide an adequate number of observations such that the quantities of interest can be estimated with sufficient precision and that any difference of importance is likely to be detected. These determinations are based on sound statistical principles. The methods for determining the sample size depend on the goals of the study, the types of outcome measures, the planned mechanism of data gathering, and the tolerance in certain error levels. For example, the planned study may be observational or experimental. The planned data gathering may be through a simple random sample of individuals or through other complex sample design. Often, information is collected through complex sample surveys that involve stratification and several stages of clustering, and the quantities of interest may involve ratio and regression estimates. When the sampling scheme involves several levels, the sample size depends on the magnitude of variations at all levels. Intervention studies may involve many baseline measures before intervention starts and several postintervention measurements to determine the effect of intervention. In follow-up studies, it may also be important to adjust the sampling size for missing data, dropouts, and censoring.

Basic Principles

Sampling techniques are used either to estimate statistical quantities with desired precision or to test statistical hypotheses. The first step in the determination of the sample size is to specify the design of the study (simple random samples of the population, stratified samples, cluster sampling, longitudinal measurement, etc.). If the goal is statistical estimation, the endpoint to be estimated and the desired precision would be specified. The desired precision can be stated in terms of standard error or a specified confidence interval. If the goal is to conduct statistical testing, the determination of sample size will involve specifying (1) the statistical test employed in testing the differences in end point, (2) the difference in the end point to be detected, (3) the anticipated level of variability in the end point (either from previous studies or from theoretical models), and (4) the desired error levels (Type I and Type II errors). The value of increased information in the sample is taken into consideration in the context of the cost of obtaining it. Guidelines are often needed for specifications of effect size and associated

variability. One strategy is to take into account as much available prior information as possible. Alternatively, a sample size is selected in advance and the information (say, power or effect size) that is likely to be obtained with that sample size is examined. Large-scale surveys often aim to gather many items of information. If a desired degree of precision is prescribed for each item, calculations may lead to a number of different estimates for the sample size. These are usually compromised within the cost constraint. Sample size determinations under several sampling designs or experimental situations are presented in the following sections.

Simple Random Sampling

A simple random sample (SRS) is the simplest form of probability sample. As stated earlier, the goal of the study may be to estimate a quantity with a desired precision (defined as the variance or the deviance from the population mean) or to test a hypothesis about the mean. Each of the situations can be formally examined under the SRS scheme as follows. Assume that there is population of finite size N from which it is desired to draw a sample of size n. In the first scenario, the goal is to estimate the mean of a quantity with a desired variance V^2. An appropriate value of n can be determined by examining the theoretical value of the variance of the sample mean with the desired variance. From sampling theory, it is known that the sample mean $\bar{y}$, under simple random sampling without replacement has a variance $[(1-n/N)/n]S^2$, where S^2 is the element variance in the population. Equating the desired variance V^2 to the theoretical value, the desired sample size can be obtained as $n = n'/(1+n'/N)$, where $n' = S^2/V^2$. If the finite population correction can be ignored, the sample size will be exactly the ratio of the element variance to the desired variance. In another scenario, the precision is expressed differently in terms of margin of errors. The margin of error specification states the closeness of the sample mean to the population mean. Let μ denote the population mean; the aim is to estimate μ with a sample mean within a specified deviance. The specification is made as $P(|\bar{y}-\mu| \leq \varepsilon) = 1-\alpha$. In this specification, ε is the margin of error and α is the level of significance. Using the results on confidence intervals for sample means obtains an equation connecting the margin of error and sample size as follows:

$$\varepsilon = Z_{\alpha/2}\sqrt{1-n/N}\left(S/\sqrt{n}\right), \tag{1}$$

where $Z_{\alpha/2}$ represents the $(1-\alpha/2)$th percentile of the standard normal distribution. Writing $n' = Z^2_{\alpha/2}(S^2/\varepsilon^2)$, it can be seen from Eq. (1) that, in the case of sampling with replacement, the required sampling size will be n'. When sampling is done without replacement, the solution of Eq. (1) gives the required sampling size as $n = n'/(1+n'/N)$. In these examples, sample size calculations are specified as an estimation of the mean of the population with a specified error. In a third scenario, it is possible to specify the estimation of sampling size as a formulation of one sample test of a mean for which the null hypothesis is $\mu+\mu_0$ and the alternative hypothesis is $\mu+\mu_1$. With an assumed level of significance of α and power $1-\beta$, the required sample size can be obtained under the assumption of normality as follows:

$$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2}{(\mu_1-\mu_0)^2}S^2. \tag{2}$$

The various formulations of sample size calculations show the need to have some prior knowledge of the effect size $(\mu_1-\mu_0)$ and variability S^2 in carrying out the calculations. A number of suggestions have been made in the literature as to how to make reasonable guesses of these values. These include searching past studies that include similar variables or conducting pilot studies. Pilot studies that are too small are less dependable than expert guesses. A reasonable guess can be made concerning the variability in terms of the coefficient of variation. The formulas for sample size calculations can then be expressed in terms of coefficients of variation to obtain the required sample sizes. When the quantity to be estimated is a proportion (P), variance based on a binomial model can be utilized. The term $P(1-P)$ is not sensitive to changes in the middle range of P $(0.2-0.8)$, and generally, a reasonable guess of the value of P can be made.

Stratified Sampling

Stratification (or blocking) of the study population is often performed prior to sampling in order to increase the precision of the estimate of the quantity of interest. A procedure similar to the simple random sampling case requires knowledge of the variability within each stratum. Such information is seldom available. The concept of "design effect" has been introduced to simplify the calculations. The design effect (denoted as deff) is defined as the ratio of the variance of an estimate under a sampling plan to the variance of the same estimate from a simple random sample with same number of observation units. The sampling plan could be a stratified sampling or other complex sample designs. The design effect is a measure of the precision gained or lost by use of the more complex design instead of a simple random sample. If the design effect can be guessed, it is necessary to estimate the sample size using a simple random sample, as shown in the previous section, and multiply this sample

size by deff to obtain the sample size needed under the complex design. Thus, in order to estimate the population mean of a continuous variable with margin of error specified and use of stratified sampling, the required sample size n is obtained using a modification of Eq. (1) (ignoring finite population correction):

$$n = Z^2_{\alpha/2}\left(\frac{S^2}{\varepsilon^2}\right) \times \text{deff}. \quad (3)$$

The value of the design effect can be obtained from previous surveys or through pilot studies. Once the overall sampling size is determined, the allocation of the samples to strata must be considered. Two methods are generally proposed in the literature. In proportional allocation, the sampling units are allocated in proportion to the size of the stratum. When the variances of observations within strata are more or less equal across strata, proportional allocation is the best allocation for increasing precision. When the variances vary greatly across strata, an optimum allocation procedure is suggested. When the costs of sampling in each stratum are the same, the sample allocation in a stratum h is proportional to the product $N_h S_h$, where N_h is the size of the strata and S_h is the standard deviation of observations within the strata (Neyman allocation). One difficulty with the optimal allocation is that the correct information on the variability within the strata is rarely obtained.

In experimental situations in which the goal is to compare a treatment group with a control group, the allocation of the samples in each group can be simplified. Denote the mean and the variance of the first (treatment group) as μ_1 and σ_1^2 and the mean and the variance of the second group as μ_2 and σ_2^2. Also assume that the allocation sample to each group is made in a way such that $n_2/n_1 = r$. Note that in this case, the allocation to two groups is predetermined. Then the required sample size can be calculated as follows:

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/r)(Z_{\alpha/2} + Z_\beta)^2}{(\mu_1 - \mu_2)^2} \quad \text{and} \quad n_2 = rn_1, \quad (4)$$

where α is the desired level of significance and $1 - \beta$ is the power.

Cluster Sampling

Many surveys often employ cluster sampling, whereby sampling units are clusters of elements. In a one-stage cluster sampling, every element within a sampled cluster is included in the sample. In two-stage cluster sampling, a subsampling is done to select elements from the chosen clusters. Elements within a cluster may be very similar. A measure of similarity (or homogeneity) of elements within the cluster is provided by the intraclass correlation coefficient (ICC), which is defined to be the Pearson correlation coefficient of all pairs of observations within the cluster taken over all clusters. The ICC plays an important role in the calculation of sample size. For example, in a single-stage cluster sampling, when all clusters are of equal size, the design effect can be approximated as $1 + (M - 1) \times \text{ICC}$, where M is the size of the cluster. In this case, the number of clusters to be selected is calculated in two stages. First, determine the sample size as if the sampling is done under simple random sampling. Then multiply that sample size by the design effect. Once again, the ICC must be known to complete the calculations, but ICC is seldom known and has to be estimated through pilot studies or derived from the values obtained in similar surveys.

Many intervention studies use group-randomized design to examine the effect of an intervention. For example, school-based studies are often used in drug-use prevention studies. In these studies, schools are randomized to treatment and control groups. The question then arises as to how many schools must be selected for each group. In these trials, a school is considered as a cluster and the intraclass correlation is used as a measure of dependence among students within the school. If the goal is to test the difference in the means of a continuous outcome at a significance level α with a desired power $1 - \beta$, the number of schools (n) to be allocated for each group will be

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 2S^2[1 + (M - 1) \times \text{ICC}]}{M\Delta^2}, \quad (5)$$

where M is the size of the school, Δ is the hypothesized difference in mean of the treatment and control schools, and S^2 is the total element variance (which includes both within- and between-persons components of variance). In most situations, the cluster (school) size will not be equal. In such situations, the size M is replaced by an average of cluster sizes (usually a harmonic mean of the cluster sizes).

Repeated Measures Design

Experimental studies often involve repeated measurements of outcome measures. For example, for comparison of an intervention group with a control group, intervention studies make baseline measurements of outcome before the intervention and then repeat the measurements one or more times after implementation of the intervention. The sample size requirements depend on the type of hypothesis to be tested, the number of pre- and postintervention measurements, and the level of correlations among observations from the same individual. Under this scenario, sample size formulas

have been developed for three possible methods of analysis—namely, (1) a simple analysis using the mean of each individual's postintervention measures as the summary measure, (2) a simple analysis of each individual's difference between means of postintervention and preintervention measurements, and (3) using the preintervention measurements as a covariate in a linear model for comparing the intervention comparison of postintervention means.

Repeated measures data are considered as correlated observations and the generalized estimating equation (GEE) method is employed in analyzing such data. Several authors have discussed estimation of sample size when the GEE method is involved as a tool of analysis; one study provides an approach to estimate sample size for two-group repeated measures when the correlation structure among the repeated measures is unknown.

Follow-up Studies

Follow-up studies usually begin with assigning individuals to an intervention or control group; the individuals are then followed for a fixed period or until the event of interest occurs. The objective of this study design is to detect a change in the rate of occurrence of the event (hazard) in the intervention group in relation to that of the control group. In this study situation, it is possible that some individual observations will be censored; this means that some individuals may not experience the outcome of interest by the time the study is terminated. For censored observations, all that is known is that the time to the event exceeds the duration of observation. The desired sampling size is the minimum number of individuals required to detect a specified change in the hazards ratio. A simple formula has been developed to calculate the required sample size. Let P_I and P_C denote the proportion of individuals assigned, respectively, to the intervention and control group. Let the ratio of the hazard function of individuals in the intervention group to that of the control group be a constant denoted by Δ. Then the total number of individuals required for the study can be expressed as follows:

$$n = \frac{1}{d}\,\frac{(Z_\beta + Z_{1-\alpha})^2}{P_I P_C \log^2 \Delta},$$

where d is the proportion of individuals expected to experience the event of interest. As before, $Z_{1-\alpha}$ and Z_β denote $1-\alpha$ and β percentiles of the normal distribution. The determination of d requires some additional information. Let f denote the planned follow-up time. Often, there is some prior information available about the rate of event occurrence in the control group. Suppose that $S_C(f)$ denotes the probability that an individual in the control group does not experience the event by time f. Then the proportion of individuals in control group experiencing the event by the follow-up time f is $d_C = 1 - S_C(f)$. Thus, under the postulated hazards ratio Δ, the proportion of individuals expected to experience the event in the intervention group is $d_I = 1 - (1 - d_C)^{1/\Delta}$. Then $d = P_C d_C + P_I d_I$.

Epidemiologic Study Designs

Epidemiological studies often use study designs that require special formulas for sample size determinations. For example, in a case-control study, a sample of people with an end point of interest (cases) is compared to a sample without and end point of interest (controls). In this case, the sampling is performed with stratification according to the end point, which is different from the usual stratified sampling. A case-control design will lead to the calculation of an odds ratio as an approximate relative risk, and the sample sizes are determined using an odds ratio as the index of interest. To prevent confounding effects, matched case-control studies (in which the cases and controls are matched at a level of a potentially confounding factor) are sometimes used. Sample size calculations for such designs have been described. Other epidemiologic designs that require special formulas for sample size calculations include nested case-control studies and case-cohort designs.

Covariate Adjustments

Frequently, studies will have end points with regression adjustments for one or more covariates. Many of the sample size calculation formulas can be easily modified to take this situation into account. For example, consider a simple logistic regression situation for which the goal is to examine the relationships of a covariate with a binary response. The model setup is $\log[p/(1-p)] = \beta_0 + \beta_1 x$, where x is a covariate. The sample size determination is made to test the null hypothesis $\beta_1 = 0$. When x is a continuous covariate, the required sample size can be obtained as follows:

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{P^*(1-P^*)\beta^{*2}},$$

where P^* is the event rate at the mean of the covariate x and β^* is the effect size to be tested. Simple modifications are needed when the covariate is binary or when additional covariates are included.

Conclusion

Sample size determination is an integral part of any well-designed scientific study. The procedure to determine sample size depends on the proposed design characteristics including the nature of the outcome of interest in the study. There exists a vast amount of literature on the topic, including several books. The modern computer environment also facilitates determination of sample size; software designed exclusively for this purpose is available. Many of the procedures depend on the normality assumption of the statistic. Modern computer-intensive statistical methods give some alternative procedures that do not depend on the normality assumption. For example, many people working in this field of study now prefer to use bootstrap procedures to derive the sample size. Uncertainty in specifying prior information of effect size has led to Bayesian approaches to sample size determination. These computer-intensive procedures seem to have several advantages over many of the conventional procedures of estimating sampling size.

See Also the Following Articles

Age, Period, and Cohort Effects • Clustering • Population vs. Sample • Randomization • Stratified Sampling Types • Type I and Type II Error

Further Reading

Breslow, N. E. (1996). Statistics in epidemiology: The case-control study. *J. Am. Statist. Assoc.* **91,** 14–28.

Cohen, J. (1988). *Statistical Power for Behavioral Sciences.* Lawrence Erlbaum Assoc., Mahwah, New Jersey.

Ejigou, A. (1996). Power and sample size for matched case-control studies. *Biometrics* **52,** 925–933.

Elashoff, J. D. (2000). *NQuery Advisor. Release 5.0.* Statistical Solutions, Ltd., Cork, Ireland.

Frison, L., and Pocock, S. (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implication for design. *Statist. Med.* **11,** 1685–1704.

Hsieh, F. Y., Bloch, D. A., and Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statist. Med.* **17,** 1623–1634.

Joseph, L., Burger, R. D., and Belisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statist. Med.* **16,** 769–789.

Jung, S., and Ahn, C. (2003). Sample size estimation for GEE method for comparing slopes in repeated measurement data. *Statist. Med.* **22,** 1305–1315.

Kish, L. (1995). *Survey Sampling.* John Wiley, New York.

Lenth, R. V. (2001). Some practical guidelines for effective sample size calculations. *Am. Statistic.* **55,** 187–193.

Lenth, R. V. (2003). *Java Applets for Power and Sample Size.* Available on the Internet at www.cs.uiowa.edu

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized models. *Biometrika* **73,** 13–22.

Liu, G., and Liang, K. Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53,** 937–947.

NCSS. (2002). *Power Analysis and Sample Size Software*. Available on the Internet at www.ncss.com

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73,** 1–11.

Rochon, J. (1991). Sample size calculations for two-group repeated measures experiments. *Biometrics* **47,** 1383–1398.

Schoenfeld, D. A. (1983). Sample size formula for the proportional hazards regression model. *Biometrics* **39,** 499–503.

Shuster, J. J. (1990). *Handbook of Sample Size Guidelines for Clinical Trials.* CRC Press, Boca Raton, Florida.

Thomas, L., and Krebs, C. J. (1997). A review of statistical power analysis software. *Bull. Ecol. Soc. Am.* **78**(2), 126–139.

Troendle, J. F., and Yu, K. F. (2003). Estimation of sample size for reference interval studies. *Biometr. J.* **45**(5), 561–572.

Woodward, M. (1992). Formulae for sample size, power and minimum detectable relative risk in medical studies. *Statistician* **41,** 185–196.

Yafune, A., and Ishiguro, M. (1999). Bootstrap approach for constructing confidence intervals for population pharmacokinetic parameters I: Use of bootstrap standard error. *Statist. Med.* **18,** 581–599.

Zou, K. H., Resnic, F. S., Gogate, A. S., Ondategui-Parra, S., and Ohno-Machado, L. (2003). Efficient Bayesian sample size calculation for designing a clinical trial with multi-cluster outcome data. *Biometr. J.* **45**(7), 825–836.

Scales and Indexes, Types of

Thomas R. O'Connor
North Carolina Wesleyan College, Rocky Mount, North Carolina, USA

Michael T. Eskey
Troy State University, Troy, Alabama, USA

Glossary

anchoring When a respondent forms a frame of reference around some division point in a scale or index because of some strong opinion or perception of research purpose.
calibration The process of achieving precision and accuracy by breaking a complex phenomena down into coherent parts for measurement purposes.
cross-validation A test or series of pretests run on a scale or index to account for validity shrinkage and more accurate accounts of validity and reliability.
dimension (or domain) The important part of personality, behavior, feeling, opinion, attitude, or judgment that concerns the researcher and stands alone as a unitary portion of the whole. The standardization of the characteristic must be established to allow for the measurement of the value of the characteristic to be measured. For example, for numerical values it is units of measurement used for expressing this particular numerical value. For texts, it is the language of the text.
index A numerical scale used to compare variables with one another or with some reference number.
norming The reporting of expected scores among population subgroups by standardizing raw scores to reflect fairly normal distributions.
scale A cluster of items that taps into measurements developed on face validity and/or professional judgment of measuring what one intends to measure. The intent is to measure the relative degree, amount, or differences between variables.
triangulation Combining research methods to give a range of perspectives. It is often beneficial when designing an evaluation to incorporate aspects of both qualitative and quantitative research designs.

Scales and indexes (indices) are used by researchers when the objects of study are complex, usually social, phenomena, such as love or prejudice, or income, education, crime, or attendance. However, there is no reason to restrict the use of scales and indexes to feelings since personality traits, behavior, opinion, attitude, judgment, and many other concepts or variables in science lend themselves to measurement by scales and indexes. Many, if not most, researchers prefer not to make major distinctions between scales and indexes and to use the terms interchangeably. This article, however, emphasizes distinctions that can be made by reviewing the role of calibration in measurement, what scales and indexes look like, their usage and purposes, and current trends in social measurement.

Calibration in Measurement

Scales and indexes (indices), in addition to most kinds of summary measures or assessment instruments in the research process, are basically types of tests that consist of items in question format or those calling for a subjective judgment. Like all tests, they are intended to be scored, but there is a major difference between scaling and scoring a test. When a test is scored, the result is usually a number, such as 94 out of 100, and it is used primarily for ranking those who scored a 94 with those who scored lower or higher. When a test is scaled, the result is also usually a number, such as 5.562, but the scale score indicates how consistent or coherent the underlying

phenomenon of interest is measured for the person taking the test. The same is true for indexes, which often involve a collection of scale scores, because the related items in indexes, like scales, are designed to report on the unidimensional characteristics of an abstract, multidimensional concept. For example, a scale may be used to measure all the extroversion characteristics of personality, and an index may be used to measure crime by asking about involvement in property crime, violent crime, and vice crime, where these types of crime make up three unidimensional characteristics of the composite, multidimensional concept called crime.

Therefore, scales and indexes not only play a key role in the measurement part of the research process but also reflect on the conceptualization and operationalization stages. For measurement purposes, scales and indexes provide the benefit of calibration, a relatively easy way to achieve precision and accuracy, since complex aspects of a phenomenon are broken down into fairly discrete units. However, not all social phenomena lend themselves easily to calibration, and the researcher is often bound by how variables have been conceptualized (defined) and operationalized (to be measured). Some slack is always present between the construct (the way something is imagined or theorized to happen) and the concept (the way other researchers suggest something happens), and it is possible for unknown error, known error, and researcher bias to enter into the measurement process. Calibration in measurement provides an excellent way of safeguarding against these risks, but it must be done intelligently and thoughtfully, especially for the dependent variable, outcome, or effect since it is the complexity of social behavior that is to be explained. Because the construction of scales and indexes often relies on the intelligence and thoughtfulness of the researcher, they are frequently criticized, but it is up to the researcher to be rigorous and consistent in coding and scoring.

As part of a theory-driven research process, scales and indexes as methods of calibration provide an excellent way for researchers to bind the horizontal logic (the presumed flow of cause and effect) of their research design with the vertical logic (how accurately theory is reflected) of their research design. This is accomplished by the researcher thinking about the dimensionality of variables, and it involves establishing the level of data involved or testing for validity and reliability. Scales and indexes, like calibration in general, frequently capture the dimensionality of variables, reduce the complexity of phenomena, summarize scores in meaningful ways, and simplify data analysis. Hence, a danger or disadvantage to the use of scales and indexes is that they may oversimplify or overquantify phenomena, particularly complex human or social phenomena that a critic might argue does not lend itself well to measurement on a 5- or 7-point scale.

Overview of Scales and Indexes

The appearance of scales and indexes ranges from very simple graphic representations (e.g., of a thermometer) to designed questionnaire formats and fairly complicated grids of lines, circles, or objects for which responses are called for by making marks on those lines, circles, or objects. All scales have in common the measurement of a variable in such a way that it can be expressed on a continuum. Indexes are similarly constructed, except that multiple indicators of a variable or various aspects of a concept are combined into a single measure. Although both scales and indexes are composite measures, it may be useful to make further distinctions between scales and indexes for overview purposes.

A scale is an attempt to increase the complexity of the level of measurement of variables from nominal to at least ordinal, and possibly interval or ratio. Scales can be utilized to provide a more complex composite and/or indicator of phenomena (variables) to be compared. Aptitude, attitude, interest, performance, and personality tests are all measuring instruments based on scales. A scale is always unidimensional, which means that it usually has construct and content validity. A scale is always at the ordinal or interval level. Scales are predictive of outcomes (e.g., behavior, attitudes, or feelings) because they measure underlying traits (e.g., introversion, patience, or verbal ability). Scaling is the branch of measurement that involves the construction of an instrument that associates qualitative constructs with quantitative metric units. Scaling evolved out of efforts in psychology and education to measure "unmeasurable" constructs such as authoritarianism and self-esteem. Scaling attempts to do one of the most difficult of research tasks—measure abstract concepts. Scales are primarily used to predict effects, as Table I illustrates.

A great many scales can be found in the literature or in handbooks, and new researchers are well advised to borrow or use an established scale before attempting to create one of their own. However, a few researchers who develop scales are interested in improving current scales and utilizing technology to refine current variable measurements. Scales allow the level of measurement that permits an intensity, potency, or "pulling together" of a measurement of variables through the utilization of

Table I Example of Scale Items Measuring Introversion

I blush easily.
At parties, I tend to be a wallflower.
Staying home every night is alright with me.
I prefer small gatherings to large gatherings.
When the phone rings, I usually let it ring at least a couple of times.

numerous responses or indicators to measure what researchers purport to "measure." It is this intensity, potency, or coming together of behavior, attitudes, and feelings that the researcher calls a "trait," something inside the person that, it is hoped, is captured in scale construction. Clearly, scaling is about quantifying the mysterious mental world of subjective experience (the immeasurable) as it impacts on empirically observed phenomena (the measurable).

An index is a statistical indicator (data or variable that provides information or allows predictions) that provides a representative value of a dimension or domain of sets such as behavior, attitudes, or feelings by measuring or "indexing" these into a single indicator or score. Indices often serve as barometers for a given variable or interest and benchmarks (standards or comparisons) against which performance (attitude, aptitudes, interests, personality, etc.) is measured.

It is possible to use statistical techniques (e.g., factor analysis) to give a more robust construct validity (or factor weights), but it is difficult to employ indexes as multidimensional measures with the theory that statistics can help determine all the unidimensional characteristics of a multidimensional phenomenon.

Indexes are usually at the ordinal level of measurement; that is, they are rank ordered and directional. An index collecting data at the nominal level of measurement is called a typology (i.e., a simple classification of traits). Indexes can be predictive of outcomes (again, using statistical techniques such as regression), but they are designed mainly for exploring the relevant causes or underlying and measurable symptoms of traits (e.g., criminality, psychopathy, or alcoholism). Indexes primarily identify causes or symptoms, as Table II illustrates.

In sociological and criminological research, indexes are usually administered in the form of surveys or questionnaires and are typically found in the appendixes of published research articles. They also comprise a significant proportion of theses and dissertations. It is sometimes the case that a researcher does not know that he or she has developed an index until after publication of his or her research and subsequent adoption of portions of his or her questionnaire or survey as a commonly used index by other researchers, who often refer to it as a scale because they use the terms scale and index interchangeably.

Table II Example of Index Items Measuring Delinquency

I have defied a teacher's authority to his/her face.
I have purposely damaged or destroyed public property.
I often skip school without a legitimate excuse.
I have stolen things worth less than $50.
I have stolen things worth more than $50.
I use tobacco.
I like to fight.
I like to gamble.
I drink beer, wine, or other alcohol.
I use illicit drugs.

However, an index is quite different from a scale. An index typically involves the measurement of many more dimensions than a scale, and sometimes an index summarizes the combined scores for more than one phenomenon. Unlike scales, an index is not concerned with the intensity or "coming together" of attributes that make up a concept or variable. Rather, an index produces a more reliable indicator of a concept or variable by accumulating a number of responses or scores on attributes that, when taken together, conveniently grasp all the known possibilities in a cause-and-effect relationship. Consider Table II, in which there are items measuring defiance to authority, disrespect toward property, a dislike of rules, an inclination toward minor stealing, and an inclination toward major stealing, among other possibilities. Each of these could have separate motives or causes, and it is certainly possible for a delinquent to engage in defiance and disrespect without engaging in stealing, but it is the job of an index to exhaust the relevant possibilities. Scales can be more discriminating because one can easily devise a scale for defiance, another scale for disrespect, and so forth. As a general rule, if one has combined three or more (implicit or explicit) scales to measure different dimensions of the same concept, then one has constructed an index, if it appears that all the relevant possibilities are exhausted. Most researchers, however, prefer to say they are using scales since scales are regarded as superior to indexes because they convey more precise information about an individual concept or variable than do index scores.

Testing for Validity and Reliability

Scales and indexes are attractive to researchers because they usually have published calculations of validity and reliability. Whenever one uses a previously published scale or index, it is common to report these original calculations along with any current calculations in the present study. Part of the attraction may be due to the fact that almost all scales and indexes are checked or validated in a pretest situation. These procedures basically involve item analysis, and they are followed by further checks for validity and reliability of the whole instrument. A researcher should begin by applying some of the best practices of good questionnaire design toward the item and response patterns on pretest results. Do the items (questions) fit together in the most productive way, or do they overlap? Do the response patterns (answers) hint at ways to improve the measuring instrument? Table III presents a sample questionnaire item to demonstrate

Table III Sample Complex Questionnaire Item

12. On October 31, 1998, Sam robbed a bank while wearing a Halloween mask and carrying a gun. While speeding from the crime scene, Sam lost control of his Jeep Cherokee and ran into a telephone pole. When the police, who had previously received a bulletin about the bank robbery, arrived at the accident scene and saw the Halloween mask and bag of money in Sam's car, they immediately placed him under arrest for bank robbery, frisked him, and then asked him; "Where's the gun?" Sam replied that the gun was in his glove compartment. As police took him to police headquarters, Sam asked; "How many years am I going to have to do for the bank robbery?" Sam's lawyer has moved to suppress both statements because Miranda warnings were not given. Assuming that Sam's statements are suppressed, but would ordinarily have guaranteed him 15 years in prison with the statements included, what do you think his sentence should be?
    A. 0–2 years
    B. 3–6 years
    C. 7–10 years
    D. 11–14 years

Table IV Formula for Item Analysis

Difficulty index

$$\frac{\text{No. of people in best group who got item right} + \text{No. of people in worst group who got item right}}{\text{Total number of people in both groups}}$$

Discrimination index

$$\frac{\text{No. of people in best group who got item right} - \text{No. of people in worst group who got item right}}{(0.5)\ \text{Total number of people in both groups}}$$

how item analysis would be conducted on a single question.

First, the principles of good questionnaire design do not seem to have been followed with this example because the sentence stem in the question was not short, clearly stated, or well written. However, it is the kind of question one would find on an achievement test or exam in an academic environment, and this is the kind of question for which item analysis was designed. Note that there is an attempt to make all the responses equidistant from one another. It does not matter if the responses are A, B, C, D, 1, 2, 3, 4, or 0–2, 3–6, 7–10, 11–14, as long as the response choices are fairly equidistant and mutually exclusive. On an academic achievement test, care would be taken to ensure that there is only one correct answer and that distracters are kept to a minimum so that no more than 2% of respondents in the pretest situation are confused by any distracter. This 2% rule applies to responses for any one question and to all the questions that make up the scale or index. In other words, if more than 2% of respondents are thrown off from the pattern that most respondents follow, then one would question the content validity of the scale or index. Content validity, like face validity, is severely affected by researcher bias, and unless the test is being used in an academic achievement environment in which the instructor is certain about correct answers, the researcher investigating more complex social phenomena should be prepared to consider the item response patterns of respondents for validity purposes.

Coefficients of Difficulty and Discrimination

Item analysis of distracter patterns is usually followed by calculation of the difficulty level and the discrimination index. To calculate these, one must sort all the completed pretests in some rank order, such as from best to worse responses, in the judgment of the researcher. Then, one takes the top 27% of the best and the bottom 27% of the worst and works out the formulas shown in Table IV. The procedure is very similar to the Kuder–Richardson, or KR-20, coefficient, which is a type of split-half reliability.

The difficulty index will derive a number from 0.00 to 0.99; ideally, one hopes to obtain a number in the moderately difficult range (0.50–0.70). The discrimination index will derive a number from −1.00 to 1.00; ideally, one hopes to obtain a number in the twenties (0.20–0.29). Anything higher means that the researcher may be favoring what he or she regards as the best respondents. A zero, near-zero, or below zero score means that the researcher is rewarding chance, or guessing, since four responses equate to a 25% equal probability of getting an answer correct; the 27% best–worst dichotomy with this formula controls for this. There are tradeoffs between difficulty and discrimination, and the general rule is that as difficulty increases, discrimination approaches zero.

Calculating the difficulty and discrimination indexes provides some assurance that the researcher is not being arbitrary or biased. This is especially important if

and when the researcher decides to attach meaning or significance to the scale items in some sort of coding scheme, such as stating that those who chose longer punishments are more punishment oriented than those who chose shorter punishments. Otherwise, the researcher is left with a fairly empty claim that face validity depends solely on his or her arbitrary judgment.

Statistical Tests for Validity and Reliability

Establishing more advanced types of validity, such as construct, concurrent, and convergent validity, requires the researcher to take additional steps, often statistical. Construct validity refers to the degree to which inferences can be made from the operationalizations in the study to the theoretical constructs on which those operationalizations were based (what the researcher thinks can actually be measured). Concurrent validity is a method of determining validity as the correlation of the test with known valid measures. Validity may be assessed by the extent to which a test yields the same results as those of other measures of the same phenomenon. Construct validity is the extent to which the scores of an assessment instrument relate to other performances of the subjects according to some theory or hypothesis—that is, the extent to which a test measures "only" what it is intended to measure. The standard approach for construct validity is to examine the different subgroups of respondents and determine if there are any chi-square differences in the response patterns that might suggest that, externally, there are other indicators not included in the scale or index. Alternatively, researchers could argue that their multiple measures tap into all known domains or a dimension(s) of a concept by using discriminant function, factor, or cluster analysis, but these advanced statistical techniques are usually reserved for attempts to establish the predictive properties of the scale or index, which is the purpose of concurrent and convergent validity. The ability to forecast future events or behavior from a scale or index is certainly not guaranteed by statistical methods of calculating validity, and researchers would be well advised to supplement, or triangulate, their research design by including interviews, case record checks, or observational follow-ups on at least a subsample of respondents.

Reliability is typically calculated by manipulating the measuring instrument. Three primary techniques to demonstrate reliability are test–retest (administering the instrument again to the same group), multiple forms (disguising administration of the instrument to the same group), and the split-half technique (administering only half the instrument to a group at any one time). In using these methods, the researcher is not searching for the exact same pattern of responses but ones that are similar, stable, and consistent. Statistical methods such as the Kuder–Richardson coefficient and Cronbach's alpha calculate reliability; again, however, researchers would be well advised to not rely on statistics but to consider such threats to reliability as setting, history, interaction, and reactivity. Setting and history threats to reliability may occur because the scale or index was developed at a time when people were acutely conscious or aware of some social problem, and interaction and reactivity effects may cloud the stability and consistency of respondents' choices simply because they believe they are being studied. Some of the latter threats can be thwarted by incorporating lie scales or social desirability scales into the instrument, but it is more important for researchers to remember that it is not so much a matter of proving validity and reliability as it is a matter of reducing threats to validity and reliability.

Usage of the Most Common Types of Scales

Scales are typically either comparative or noncomparative. Comparative scales include the following techniques: paired comparisons, rank-order comparisons, constant sum scaling, Bogardus social distance scales, the Q-sort, and Guttman scales. Noncomparative scales include the techniques of continuous rating, Likert scales, the semantic differential, Stapel scaling, Thurstone scales, and multidimensional scaling. Other types of scales exist that defy categorization and rely on graphic representations, such as rulers, clocks, thermometers, or grids. It is probably best to begin with a discussion of comparative scales since it is most likely the case that no matter whether respondents are explicitly asked to compare something or not, they will undoubtedly do so because they are bound by their cognitive frames of reference, which often include comparisons between objects.

Comparative Scales

A paired comparison scale presents the respondent with two choices and calls for a preference. For example, the respondent is asked which color he or she likes better, red or blue, and a similar process is repeated throughout the scale items. Note that there are no scale properties within each item; that is, respondents are not provided with any scale other than the extreme choices they must make (e.g., red or blue) on each item. All the questions on such an instrument make up the scale. Scoring is accomplished by following the researcher's coding key for what each choice means. Although this technique appears to collect

nominal-level data, the scale score is considered to be at the ordinal level of measurement.

A second type of comparison scale involves rank-order comparison, which presents the respondent with a number of items, cues, or objects and asks him or her to rank them in order from first to last, from 1 to 10, or along some other rating category. For example, respondents may be asked to rate excerpts from political speeches and assign them a rank order from most conservative to most liberal. Each respondent's rankings are then tabulated to assign scale scores to different patterns of variation, which are normally treated as ordinal-level data.

A third type of comparison scale involves giving the respondents an imaginary constant sum or fixed amount of something, such as $100,000 to start a business, and then asks them to allocate this fixed amount among budget items such as salaries, benefits, equipment, and raw materials. To this might be added an item asking them what they would spend an additional several hundred thousand dollars on from a fixed list of items to purchase. The idea is to scale the respondent's choices each time, producing a number of subscales and, overall, producing a composite scale, in reference to a constant sum that may or may not be exactly the same for each question. This technique is generally considered to involve an ordinal level of measurement.

Bogardus social distance scales are named after Emory Stephen Bogardus (1882–1973), a sociologist who made contributions to the study of prejudice. A social distance scale measures degrees of tolerance or prejudice between social groups. For example, respondents would be asked if they want to allow certain foreigners to have citizenship, if they would want to live next door to certain foreigners, if they would want to be coworkers with certain foreigners, if they would want to be close friends with certain foreigners, and if they would want to be related by kinship in marriage to certain foreigners. Social distance scales are assumed to be cumulative (i.e., to have certain Guttman scale properties) and have the longest usage of any scales in sociology.

Q-sort scales are generally used for measuring the dimensions of complex attitudes, and there is a presumption that the technique measures true feelings. In addition, it is normally a fun instrument that rarely produces negative reactions in respondents. Q-sort is a comparative technique that requires respondents to sort various cues or items written on cards into a predetermined number of piles. The piles represent categories that range, for example, from most like their own attitude to least like their own attitude, and they comprise the scale. The cue cards typically contain short descriptors or adjectives, such as outgoing, happy, or sad, and it is customary not to exceed 140 items for respondents to sort. Q-sort scales can have as few as three piles, reflecting perceptions, for example, of the lower, middle, and upper class, or they can be complicated grids with a large number of rows and columns. Respondents are generally allowed to move the cards freely until they are satisfied with final placement.

Guttman scales are named after Louis Guttman (1916–1987), a statistician who made enormous contributions to social science measurement. Usage of a Guttman scale is sometimes called scalogram analysis. A Guttman scale is composed of a series of items for which the respondent cumulatively indicates agreement or disagreement. For example, a Guttman scale might ask if a person would tell a little lie (yes or no), then ask if he or she would tell a bigger lie (yes or no), then ask if he or she would lie for no reason at all (yes or no), and then ask if he or she would lie so much that he or she could not tell the difference (yes or no). It can be readily seen that Guttman scales are strictly unidimensional, and they are easily scored with the scale defined as the total number of questions or items passed or agreed with. Researchers should take care to ensure that the cumulative pattern is logical. Unpredictable response patterns, or random error, are controlled for by calculating a coefficient of reproducibility, which is the integer 1 minus the number of errors or unlikely responses divided by the number of responses. Items that produce greater than 80% agreement among all respondents should be discarded from the final scale, and items that reach a 90% coefficient of reproducibility indicate accuracy or scalability. Guttman scale scores have the advantage of being highly informative about where a respondent stands on each item since a scale score of 2 means that he or she agreed only to the first two items and not the ones that followed. They have the disadvantage of usually requiring the passage of time for a respondent to have done all the things asked about.

Almost all comparative scales are designed for use with fairly sophisticated, experienced, or opinionated respondents since it is the comparison process built into these scales that provides some safeguard against systematic error in the form of halo effects, generosity effects, extreme response effects, and contrast effects. To control for nonsystematic error, such as that driven by respondent fatigue or inattention, researchers are well advised to calculate the coefficient of reproducibility and conduct other checks for reliability, such as using multiple raters to generate composite ratings that can then be compared to individual ratings in a pretest environment.

Noncomparative Scales

Noncomparative scales are more likely to resemble a questionnaire presentation, and they expose the respondent to only two things to think about at a given time: the object of the sentence stem in the question or item and the translation of his or her subjective impression about that item into the number of rating divisions

provided. The respondents' frame of reference is not therefore unimportant, and they probably construct a frame of reference on the spot based on their assessment of the whole questionnaire. A number of books on questionnaire design as well as empirical literature indicate that respondents anchor their frame of reference on a neutral point or something they have a strong opinion about when they complete a questionnaire. Hence, researchers often take great care in thinking about the division points they provide raters in their noncomparative scales and indexes. Whether 5-, 7-, or 10-point divisions are used depends in large part on how successfully the researcher is able to thwart respondent anchoring as much as it involves providing reasonable divisions that represent a veridical map of subjective reality.

A continuous rating scale is a type of noncomparative scale that presents the respondent with a continuum, or line, on which to place slash marks in response to a question or item. The idea is to let respondents imagine for themselves what are the division points. Sometimes, the lines are labeled at each end, for example, by strongly disagree and strongly agree. Sometimes, the researcher provides hash marks along to line to help orient the respondent. Scoring is done by the researcher using a ruler, usually measured in millimeters, to record numerical, presumably interval-level, responses for each question or item. This technique has some appeal to those who use the confidence interval approach to statistical interpretation since the method of data collection and the method of data analysis are similar.

Likert scales are named after Renis Likert (1903–1981), an attitudinal researcher, and are the most widely used scales in social science. They provide ordinal data measurement. Generally, they consist of 10–15 questions, but they often begin with a pool as large as 100 questions. Normally, they have an even number of division points, such as strongly disagree, disagree, agree, and strongly agree, with no middle or neutral point representing no opinion. However, Likert scales with 4- to 9-point scales are frequently found in various fields of research, and 5-point divisions are perhaps most commonly used. Any even number of division points will eliminate neutral ground, and any odd number of division points will provide a neutral response option. Because Likert scales have such widespread use, an example of one is provided in Table V.

Because the example in Table V deals with self-esteem, it could obviously be extended to a longer series of items. In fact, researchers often brainstorm all the potential items that could comprise such a scale, as long as the items tap into a unidimensional concept of interest. To reduce the question pool, raters or judges can be used in a pretest environment to rank the items as more or less relevant to the concept of interest, the rankings of these judges can be correlated, and weak items with intercorrelations lower than 0.60 can be thrown out. A t test difference of means analysis can also be done between groups of judges to improve the discrimination value of a Likert scale. It is not infrequently the case that some questions or items are negatively worded, so reverse coding or scoring is needed for these items, with the idea being that such reversals downplay the risk of respondents faking or distorting their answers. The researcher typically uses a Likert scale to test for significant differences between the medians of comparable groups, in the two-sample case using the Mann–Whitney test, in the paired sample case using the Wilcoxon test, and with three or more samples using the Kruskal–Willis test. Well-constructed Likert scales have a tendency to produce rather high reliability coefficients, and the only problem researchers are likely to encounter is what to do with missing data when a respondent skips a question or item. When missing items occur, it is customary for the researcher to assign the average score calculated from whatever items have been answered, although this should not be done when there are too many missing items.

The semantic differential is a technique used to measure the meaning of an object to a respondent. Any concept, person, place, or thing can be presented, and respondents are asked to rate their subjective impressions on a series of bipolar, 7-point scales. Respondents would, for example, indicate their position among the dividing points between such bipolar endpoints as fair–unfair, good–bad, valuable–worthless, active–passive,

Table V Example of a Likert Scale

Strongly disagree	Disagree	Somewhat agree	Strongly agree	1. I feel good about myself.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	2. Most people think I'm a nice person.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	3. I always try to be who I am.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	4. I have a strong personality.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	5. I present myself well.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	6. I watch out for my self-interests.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	7. I am usually self-confident.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	8. I am proud of who I am.

and strong–weak. Researchers generally choose bipolar endpoints that are reasonably relevant to the object presented. For example, a good–bad bipolar endpoint might be useful when asking respondents to measure their impression of politicians, but a strong–weak bipolar endpoint might be appropriate for an impression of athletes. In other words, the bipolar endpoints are usually customized for the object under consideration. Numerical scoring with the semantic differential produces scale scores that usually require the researcher to assign some meaning to the bipolar directions, where the respondent chooses more favorable words or less favorable words. This ordinarily requires examination of the variation, or standard deviation, among the responses. If respondents are not using all available 7 points across the scale, researchers sometimes collapse the scale to 6 points or less to make the meanings more significant. Such collapsing of a scale should be reported in the research write-up, and it is normally considered an instrumentation threat to validity and reliability. Semantic differential scales have been found useful in cross-cultural research and where broad differences in vocabulary exist among a population sample. Semantic differential and Stapel scales are also widely used in the fields of market research and business.

A Stapel scale is a measure of attitude or impression that consists of a single adjective presented in the middle of an even-numbered range of responses, from −5 to 5. It is usually used when the researcher cannot think of any bipolar adjectives when constructing a semantic differential scale. For this reason, it is sometimes referred to as a unipolar version of the semantic differential. Respondents are sometimes, although infrequently, asked to report any words they think of that explain their numbered ratings. For example, if the unipolar adjective is "challenging" in relation to a university's reputation and the respondent choose −2, then the respondent might write down a word such as "boring" to better reflect what he or she meant by −2. This additional information captured by a Stapel scale can then be analyzed, become the basis for a semantic differential scale, or be used to refine future measurement.

Thurstone scales are named after Louis Thurstone (1887–1955), a psychologist who practically invented scaling at the interval level of measurement. Basically, a Thurstone scale is constructed so that different positions on the scale represent known attitudinal positions with meaningful distances between them. This is accomplished by using raters or judges in an extensive pretest environment. Working independently, the judges are asked to create as many statements about an object as possible. This produces a pool of statements. To get the judges started, the researcher often issues a fill-in-the-blank command, such as "criminals are _____." Sometimes, more than 100 statements are collected. Then, the judges are asked to rank each statement in the pool as either being unfavorably inclined toward the object or favorably inclined toward the object. These rankings are done on a scale of 1 to 11, with 11 being favorably inclined. Note that the judges are not allowed to express their own opinions but instead are objectively rating statements as more or less charitable toward the object. The median and interquartile range (75th percentile minus the 25th percentile) are then calculated from rankings for each statement. Items are selected for the final scale from statements that have equidistant medians and the smallest interquartile ranges within each median's category. This process produces a scale consisting of statements with known equidistant intervals between them, and it also includes statements that have the least amount of variability among judges. Weights can be assigned to some statements to adjust mathematically for a lack of perfect equidistance. Hence, a Thurstone scale discriminates around the intensity or potency of certain attitudinal positions that make up all the known high and low attitudinal positions toward an object. To administer a developed Thurstone scale, the researcher simply has to ask respondents to agree or disagree with the final set of statements, and the mean response is the scale score. Rarely are Thurstone scales encountered in the contemporary literature because the construction technique is considered cumbersome. Also, many researchers today are in the habit of considering ordinal as interval, especially with regard to the rather high reliabilities obtained with simpler methods, such as Guttman and Likert scaling.

Multidimensional scaling (MDS) is an extension of computerized factor analysis that reveals the underlying dimensions of a correlation matrix consisting of responses to a scale. For example, a scale is pretested, and the results are factor analyzed. However, instead of the researcher rotating the axes in an orientation that produces dimensions that can be easily explained (e.g., orthogonal and varimax rotation generally minimize to one or two dimensions), the researcher uses a function minimization algorithm that allows him or her to preset the number of dimensions (usually three) and let the computer program rotate in all sorts of orientations. A screen test is normally used to justify the number of dimensions to plot, and the scatterplot of distances between observed and expected similarities is called a Shepard diagram; Both are precursors to a graphical mapped output of the three dimensions. The researcher then examines this graphic for clusters of points, or configurations, that represent the similarities among scale items that uncover the underlying dimensions of the respondents' subjective impressions. MDS techniques work well with semantic differential-type measures and have proven useful with other scale refinements. They have the advantage of being compatible with multiple regression as a tool for regressing certain meaningful variables onto the coordinates for different dimensions.

Usage of the Most Common Types of Indexes

Indexes are different from scales in that two or more variables or the multiple yet related aspects of a dimension or domain are measured in an index. However, a more basic difference is that with an index, there is no assumption about measuring intensity, as would be done with a scale. For illustration purposes, Table VI shows this basic difference with regard to the way in which scoring is done.

Note in Table VI how the response to each item in the index can earn only one point. Also note how an increasing level of intensity is captured by the way scale items are presented, and that as many as seven points can be earned for each item in the scale. The inability of an index to capture information about the intensity of subjective impressions along unidimensional lines toward an object means that indexes would be more useful when research is on exploratory ground, particularly when multidimensional concepts or overlapping domains are involved. When the researcher cannot delineate all the known dimensions or domains of a construct, an index is appropriate because, in essence, the researcher is using proxies or indicators to guess at the fullness of the object. Thus, it is sometimes stated that an index collects causes and a scale predicts effects because an index taps into uncharted territory and a scale taps into known territory. It is the case, however, that common thinking about indexes is just the opposite. People have become accustomed to only seeing the indicator use of indexes, as with crime indexes and consumer price indexes that measure observable outcomes, behaviors, or events.

Truly useful indexes can be constructed for theory testing, elaboration, specification, or integration of theoretical variables. Some researchers continue to prefer calling them scales because they appear to be getting at elusive concepts, but this is only convergent validity, not predictive validity. An index can be validated or made reliable in most of the same ways as can scales, and it is even possible to construct indexes at the interval level with weighting schemes. Researchers constructing indexes are usually concerned with representativeness and generalizability, purposes that go beyond the quest for calibration in measurement, and these purposes are typically accomplished by standardizing and cross-validating a test. These procedures can also be followed for scales.

Table VI Comparison of Index versus Scale Measurement

Index:	Successful business plan + adequate venture capital + good corporate organization + successful public stock + satisfactory merger
Scoring:	1 point for each action
Scale:	(1) Successful business plan, (2) adequate venture capital, (3) good corporate organization, (4) successful public stock, (5) satisfactory merger
Scoring:	1–7 points for each action

Norming, Standardization, and Cross-Validation

A fully developed index is typically normed, standardized, and cross-validated. Norming involves examining the distribution of raw scores from the pretest respondents and doing within-group comparisons by age, race, and gender, for example, to search for signs of any adverse impact or gaps in what should theoretically be normal distributions. The number of cases selected for within-group analysis should reflect the proportions of those groups in the larger population. The raw scores are then converted to standard scores (z scores) or percentile points, and smoothing operations are performed on the actual scores across and within subgroups until normal distributions are approximated. Index items can also be deleted or revised at this point. The resulting index scores can then be said to have been normed on various population groups, and it is customary for researchers to report what the average score should be for these groups.

Standardization is the process of administering a revised index to a second sample of pretest respondents. Indexes are usually revised by rewording an item slightly, rather than by deleting items, because some aspects of a domain may be particularly difficult to write items for and it is better to try to tap that domain than to forget about it. The standardization sample generally involves respondents from a forensic population for which the researcher suspects attributes of the domain are common. For example, an offender population might be used to standardize a criminological index, but only to compare the distribution of scores between this offender population and the normed scores. Standardization here means the researcher has essentially used an experimental group to verify that the index is constructed so that approximately normal distributions can still be obtained, and it is customary for researchers to also report these respondents as a group upon which the instrument has been normed.

Cross-validation is the process of administering yet another pretest, this time to a sample other than a population for which the researcher suspects attributes of the domain are common—in other words, an ordinary sample of ordinary people. This sample should be used to calculate validity and reliability coefficients, and it should represent the reported properties of the index. Cross-validation is an important step because with only a standardization sample, a researcher will get artificially inflated coefficients, primarily because of criterion validity. It is more honest to let chance and validity shrinkage run

its course and report lower, if not average, coefficients from cross-validation testing.

The Dimensionality of Indexes

The least complicated indexes are unweighted; that is, each item counts for one point, or the items represent raw counts that occur naturally. An example of an unweighted index is the Uniform Crime Report (UCR Crime Index), which has been calculated by the Federal Bureau of Investigation since 1930 and uses raw counts of murder, rape, robbery, aggravated assault, burglary, larceny, auto theft, and arson. The UCR was designed to be a sort of moral barometer of the nation's mental health, at least insofar as counts of the most serious crimes in society are concerned. One way to improve the UCR is to create a weighted index that assigns extra points to some crimes and not to others based on surveying people's perceptions of seriousness. Another way would be to norm the index by accounting for offense rates or perceptions of seriousness by age, race, and gender. In any event, both unweighted and weighted indexes have the ultimate purpose of indicating, imaging, modeling, or profiling the potential for something based on the accumulation of point values that represent possible determinants, causes, or symptoms. Indexes usually contain many items that represent the behavioral domain because past behavior is a good predictor of future behavior, but indexes also cross domains and can include items that measure other dimensions, such as thinking and feeling, as Table VII illustrates.

Table VII presents portions of an inequity index derived from theories in criminology and social psychology about perceptions of effort-based injustice or unbalanced exchange relationships. This construct has played a role in many theories as a precipitating factor for the readiness to offend or the potential to engage in antisocial behavior. The inequity index consists of a series of three bundled items: The first item asks about the respondent's behaviors, or how much effort he or she put into something; the second item asks about the respondent's cognitions, or how much the respondent observed a reflection of his or her effort from others; and the third item asks about the respondent's feelings, or how much the respondent felt let down for his or her effort. For each bundle in the index, scoring proceeds by adding respondent values for the first item, subtracting values for the second item, and adding values for the third item. The domains of behavior, cognition, and feeling tapped by this instrument could just as easily have been measured by scales, but three separate scales would be required, the instrument would be unwieldy, and it would be doubtful whether the researcher was tapping into the inequity construct. The example should illustrate how particularly useful indexes can be for the measurement of theoretical constructs when the researcher is not interested in exhausting the dimensionality of any one domain but, rather, in exhausting all the overlapping dimensionality between domains.

Table VII Inequity Index Tapping the Behavioral, Cognitive, and Emotional Domains

Instructions: Think about the relationships you have with those closest to you, and place a number on the line in front of each item for how many times in the past year you have behaved, thought, or felt that way.
_____1. Made efforts to impress others with your intelligence
_____2. Were recognized for your intelligence
_____3. Felt unappreciated for your intelligence
_____4. Put a lot of effort into some project
_____5. Felt satisfied from working on some project
_____6. Felt unmotivated from working on some project
_____7. Tried to be sociable
_____8. Received comments on how sociable you were
_____9. Felt like others were not being sociable with you
_____10. Showed consideration for others
_____11. Received consideration from others
_____12. Felt like others were inconsiderate toward you

Trends in Scales and Indexes

The two most common trends in the field of measurement by scales and indexes are the habits of researchers to treat scales and indexes synonymously and to treat ordinal-level data collected by scales with high reliability as interval-level data for all practical purposes. These are unfortunate trends because there is a rich, unexplored usefulness for measurement by indexes, and although Thurstone-type procedures and weighting schemes are cumbersome, interval-level scales and indexes can be constructed.

Factor analysis also seems to be used frequently by researchers to identify the multiple dimensions of a concept. The logic of factor analysis is that sometimes there is utility in assessing the correlations between several related indicators in such a way so as to uncover the latent, abstract concept that those indicators measure. Factor analysis involves confirmation of whether or not those indicators are indeed measuring a single concept, and if they are, the group scores for those "factors" are called eigenvalues and treated by researchers as proxies for the dimensions.

Mathematical scaling or MDS is a fairly recent trend, and it seems to have found a place among data analysts steeped in the multiple regression tradition. MDS is a set of data analysis techniques that display the structure of distance-like data as a geometrical picture. It represents one way in which researchers are constantly working to revise and improve their scales and indexes. As previously stated, researchers are well advised to seek out existing

scales and indexes from a handbook, manual, or, in some cases, the researchers, at varying costs. At times, using a published scale or index will prevent reinvention of the wheel, and at other times there is no substitute for constructing, validating, standardizing, and cross-validating one's own scale or index. This area of measurement will likely remain vibrant for many years.

See Also the Following Articles

Education, Tests and Measures in • Likert Scale Analysis • Multidimensional Scaling (MDS) • Reliability Assessment • Thurstone's Scales of Primary Abilities • Validity Assessment

Further Reading

Bogardus, E. S. (1933). A social distance scale. *Soc. Social Res.* **3,** 265–271.

Brodsky, S., and Smitherman, H. (1983). *Handbook of Scales for Research in Crime and Delinquency.* Plenum, New York.

Buros Institute of Mental Measurements (2003). *The Fifteenth Mental Measurements Yearbook.* University of Nebraska Press, Lincoln. [Available at http://www.unl.edu/buros]

DeVellis, R. (2003). *Scale Development: Theory and Applications,* 2nd Ed. Sage, Thousand Oaks, CA.

Fowler, F., Jr. (1995). *Improving Survey Questions.* Sage, Thousand Oaks, CA.

Guttman, L. (1944). A basis for scaling qualitative data. *Am. Soc. Rev.* **9**(2), 139–150.

Labaw, P. J. (1980). *Advanced Questionnaire Design.* Abt, Cambridge, MA.

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* **140,** 1–55.

Netemeyer, R., Bearden, W., and Sharma, S. (2003). *Scaling Procedures: Issues and Applications.* Sage, Thousand Oaks, CA.

Osgood, C., Suci, G., and Tannenbaum, P. (1957). *The Measurement of Meaning.* University of Illinois Press, Urbana.

Scales: Social sciences (2003, October 3). In *Wikipedia: The Free Encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/Scale_(social_sciences) on November 1, 2003.

Thurstone, L. L., and Chave, E. J. (1929). *The Measurement of Attitudes.* University of Chicago Press, Chicago.

Trochim, W. M. (2000, June 29). Scaling. In *The Research Methods Knowledge Base*. Retrieved from http://trochim.human.cornell.edu/kb/scaling.htm on November 1, 2003.

Secondary Data

Marc Riedel
Southeastern Louisiana University, Hammond, Louisiana, USA

Glossary

aggregation Secondary data can be made available at various levels, for example, traffic accidents can be reported at the level of events with detailed information on participants. However, it can be aggregated so that the only information available is the number of traffic accidents per month or year.
imputation Statistical procedure that uses existing data to estimate the most likely values for data that are missing.
listwise deletion When cases are deleted from a data set if any variable for the case has missing values. For example, if a case had, in addition to three other completely reported variables, a variable of race and was missing the respondent's race or ethnicity, the entire case would be deleted from analysis.
objective indicators Data or information generated by an organization that is believed to represent the valid and reliable output of the organization.
pairwise deletion When cases are deleted from a data set only if information is missing for the variables being analyzed. For example, in comparing race to income, only those cases that had missing values on race and income would be deleted. Other variables with missing values would be retained.
rare events Events that are statistically uncommon when compared to a population at risk.
units Cases collected for the ultimate purpose of describing patterns; can be individuals, groups, or social artifacts.
values Characteristics describing objects and variables. For example, in a survey in which the variable was "gender," the values would be "male" or "female."

Secondary data is information that was gathered for another purpose. In his classic 1972 study of secondary analysis, Herbert Hyman stated that purpose is expressed when a secondary analyst "by an act of *abstraction* uses questions originally employed to indicate one entity to illuminate other aspects that a former analyst did not have in mind at all." [italics in the original] (p. 37). This article focuses primarily on quantitative secondary data.

Types of Secondary Data

There are three types of quantitative secondary data: surveys, official statistics, and official records. Depending on how the data are stored, each may be referred to as archival data. Surveys are characterized by data acquired through questionnaires or interviews and, generally, probability sampling. The research issues raised by secondary use of surveys are very different from issues surrounding official statistics and records. Many of the problems in the use of surveys focus on questions of analyses, whereas for official statistics and records, the major problems are with the characteristics of the data set itself.

Official statistics and records are collections of information maintained and made available in permanent form by organizations. Official records are collections of statistical data that are generated as an organizational by-product of another mission or goal. They are "official" in the sense that they are the records of a bureaucratic office; official statistics, on the other hand, are designed for public consumption. For example, police departments keep extensive data on criminal complaints, investigations, arrests, and characteristics of victims and offenders. Official statistics in the form of annual reports from the FBI on crime in the United States provide information that is accessible to anyone. Official records, on the other hand, are more difficult to access than official statistics because they are constructed primarily for internal use. Compared to official statistics, agencies that generate official records retain a proprietary interest in their use.

The proprietary issues range from legally protected information about individuals to a concern that research using official records may throw an unfavorable light on agencies.

Official records typically contain a greater amount of detail than official statistics. Official statistics are collected from many agencies and are meant to be disseminated widely. Hence, the data collected focus on a few elements that will be reported consistently and accurately. For example, the data that is reported in national mortality statistics is taken from death certificates that contain much less information than is available in the files of coroners or medical examiners.

For official records, the unit of analysis is usually based on the target of service delivery. Thus, welfare agencies collect information on individuals who apply for food stamps, employment agencies collect information on people looking for work, etc. Official statistics, on the other hand, make information available at higher levels of aggregation, such as the monthly or annual number of people applying for welfare or employment.

Finally, one advantage of official records is that the persons who originally completed the forms may be present in the agency and available to the researcher. In the absence of adequate documentation, the researcher is given the opportunity to learn how the information was gathered. By contrast, official statistics may be gathered from many different agencies or presented in a form that makes it difficult or impossible to discern the procedures.

The distinction between official statistics and official records is not meant to be a distinction between national and local data. Local and state agencies provide annual reports that qualify as official statistics and are not found in archives. Official records may be archived locally, but may not be generally available. The distinction is between types of secondary data that raise different issues for researchers wanting to use the data.

Uses of Secondary Data

Accessible and Inexpensive

Secondary data is both accessible and inexpensive. Several of the largest archives in the world are located in the United States. The largest is the Inter-University Consortium for Political and Social Research (ICPSR) located at the University of Michigan. The ICPSR archives hold more than 1700 studies divided into 26,000 files covering approximately 130 counties. While membership in ICPSR gives universities and other organizations access to services and data, a very large number of data files can be downloaded without cost from their Web site.

While archives provide access to a large number of data sets generated and maintained by federal and other agencies, many researchers contribute their data sets to the archives upon completion of their research. In addition, for researchers interested in records not found in archives, most agencies use computer systems to store data that can be used for research, providing access is given.

Secondary data archives have contributed greatly to the efficiency of individual researchers. In his book on secondary data, Riedel mentions how secondary data archives have "reduced the *schlepping* factor." Research a bare 25 years ago involved a great deal of "schlepping," a Yiddish word meaning a tedious journey or to carry slowly or awkwardly. Secondary data archives were small; if a student was lucky, his or her major professor would provide access to his or her funded data set for a thesis or dissertation.

Requests for data from organizations might result in receiving a reel of computer tape that must be mounted on a large mainframe computer, and the available software might or might not be compatible with the data. On the other hand, such requests might mean the researcher must spend months transferring information from paper forms to computer-readable formats. Analysis meant weeks and months of feeding batch programs to the mainframe, examining results, correcting errors, resubmitting jobs, and submitting new ones. Turnaround time of 20 to 30 minutes from submission to printout was considered rapid progress. All of this was in addition to the time needed to review research literature, develop research designs, construct indicators, interpret data, and report the results.

In a short 1963 article in the *American Behavioral Scientist*, Barney Glaser discussed how the independent researcher, someone who decides to do basic research with little or no resources, benefits from secondary data. With the emergence of inexpensive and accessible data sources, Glaser's suggestions are even more relevant today because the largest cost of doing research is data collection, not data analysis. Glaser discussed four types of independent researchers: team members, students, teachers, and people in administrative positions; here, students are briefly discussed.

Undergraduate and graduate students who want to engage in original research and/or are required to do a thesis or dissertation typically have few or no funds to support data collection and are unlikely, by themselves, to obtain any external funding in large amounts. Ideally, students can participate in the data-gathering efforts of well-funded research that might result in their access to primary data. More commonly, self-funding means making use of existing resources within a university that includes free use of computers for data processing, access to libraries, and being able to consult with faculty. The ready availability of a large number of data sets solves the expensive problem of data collection.

An opposing argument is that students need the learning experience of collecting their own data. It is difficult to see, however, what useful knowledge is gained by learning how to collect data with no funds and limited time. Students are sufficiently familiar with how poverty limits opportunities without imposing formal requirements. In addition, graduate students are increasingly expected to graduate with one or more publications to their credit. Being able to publish a thesis or dissertation is much more likely if higher quality data available from other research projects is used.

Difficult-to-Observe Events

Rare Events

Many phenomena are so statistically rare that secondary data are the only practical source of information. As a rule, the more serious the crime, the less frequently it occurs and the less frequently it can be observed directly. For example, in 2000, there were 891 murders in an Illinois population of almost 12.5 million people. Given the rarity and generally unannounced nature of that crime, information on murders must be gathered as a by-product of police investigation and apprehension.

Events Occurring in Settings Not Subject to Surveillance

Many events of interest to researchers occur in settings that are not routinely subject to surveillance. For example, crimes such as assaults, burglaries, robberies, and rapes occur in settings and ways that make identifying and apprehending offenders difficult. They only become known if the victim reports the crime to the police or if he or she happens to be a respondent in a crime victimization survey.

Of course, while settings are subject to routine surveillance is increasingly becoming an open question. For example in September, 2002, Madelyne Toogood turned herself in to the police because she was observed beating her child by a surveillance camera in a store parking lot. The fact that the video was carried on national television contributed to her turning herself in.

There are also events that occur in protected settings such as homes. For example, for decades women and children were physically assaulted and otherwise victimized in their homes without outside attention. Changes in legislation required reporting of child abuse and more careful accounting of spousal violence. Additionally, abused and battered women now have alternatives to remaining in an abusive relationships such as hotlines and women's shelters. Such changes led to the creation and storage of records that can be used for research.

Private Behavior with Public Consequences

People may engage in a variety of secret or routine behavior that comes to the attention of medical personnel, among others, who attribute very different interpretations to it and are frequently legally required to report it. Such information results in secondary data sources. Examples include wounds from weapons, repetitive patterns of injuries from automobile or industrial accidents, increases in certain types of infections or diseases, and the use of controlled substances. For example, the federal government supports a program called the Drug Abuse Warning Network (DAWN), which has two crucial drug indicators. The first is emergency department episodes in which persons are brought to a hospital emergency room with a drug-related reaction. The second source of information is obtained from medical examiners or coroners who report that one or more drugs caused or contributed to the death of persons.

Costly Data Collection

A major advantage of secondary data is that it makes information accessible that is too costly to collect under any other circumstances. There are several instances of secondary data sources that are costly to collect because of the scope of the information collected, because their utility stems from periodic repetition, or both. Perhaps the best example of secondary data of a repetitive extremely large-scale data collection that requires the resources of a national government is the U.S. Census. While the census is constitutionally mandated, it is used to do more than count people.

There are numerous costly national surveys of social indicators. The National Crime Victimization Survey, completed every six months by the Census Bureau and the Bureau of Justice Statistics, uses a nationally representative sample to provide details about crime victimization that exceed what is available from law enforcement sources. The General Social Surveys have been carried out annually (except for 1979, 1981, and 1992) by the National Opinion Research Center. The primary topics surveyed in national samples include socioeconomic status, social mobility, family, race relations, social control, sex relations, civil liberties, and morals.

Finally, the National Health Interview Surveys have been conducted since 1969 by the National Center for Health Statistics. By means of national sampling, it obtains information about the amount and distribution of illness, its effects in terms of disability and chronic impairments, and the kinds of health service people receive. Supplemental surveys collect data on such topics as AIDS knowledge and attitudes, child health care and immunization, dental care, and hospitalization.

Evaluation of the Functioning of Agencies

With respect to evaluating agencies, secondary data is used in two ways. The first views statistics and records as objective indicators of organizational effectiveness. Number of units sold, patients treated, cases resolved, and persons arrested are examples of how information from organizations is used to evaluate their performance.

Second, statistics and records can be used as an indicator of organizational processes. The records of organizations are viewed as representations of individual and institutional polices and practices. Arrests for illegal gambling, for example, represent police responses to public pressures to do something about crime more than they represent an indication of the actual amounts of what is defined as that crime. Similarly, a record of prosecutions for environmental pollution may show differential enforcement that reflects the extent of political contributions to the party in power.

Cross-Cultural or Transnational Research

Secondary data is essential to cross-cultural or transnational research not only because of the cost of collecting data, but because of the challenging organizational tasks of getting together a group of cooperating researchers and agreeing on a data collection instrument. ICPSR maintains a large collection of data sets comparing different countries. For example, *Political Participation and Equality in Seven Nations, 1966–1971* is a cross-national data set on political participation from Austria, India, Japan, the Netherlands, Nigeria, Yugoslavia, and the United States.

Comparisons of Different Times and Places

Secondary data can be used to compare changes in different places at the same time. Studies can be done comparing the amount of crime in different cities for the same time period. A study of trends illustrates how secondary data can be used to compare changes in the same place at different times. For example, the ongoing Consumer Expenditure Survey (CES) by the Bureau of Labor provides a continuous flow of information on the buying habits of U.S. consumers and also furnishes data to support periodic revisions of the consumer price index.

Limitations of Secondary Data

Indicators Not Available

The major disadvantage of secondary data is that indicators needed for the proposed concepts are not available in secondary data sets. An advantage of data collected by the investigator is that he or she specifies the concept of interest and is able to construct an indicator; the secondary data user has to use indicators that are available in the data.

There is no simple solution to the problem. In some instances, weak indicators can be found. For example, median income of a census tract has been used as an indicator of individual social class. While this has obvious weaknesses, it can be combined with other correlates of social class, such as years of education, and compared to the dependent variable. Where it is not possible to find an indicator that represents a good fit with the concept, it is possible to use several weak indicators and explore their adequacy as measures by additional analyses.

Unacceptable Levels of Aggregation

Data sets can be made available by cases or higher levels of aggregation. Where the data is available at the level of cases, it contains characteristics of the unit of analysis. For example, data on attitude surveys would be available at the level of respondents.

Data is sometimes aggregated by time intervals or by respondent characteristics. For example, the mean support for an issue is given for males in comparison to females. Where comparisons over time occur, there may be monthly or annual counts of the number of people applying for welfare, for example.

The data can be used as long as the unit of analysis in the present research is the same as the secondary data. Problems arise when the researcher wants to analyze data at a level lower than what is available. The higher level aggregation precludes creating composite variables at the level of individuals. For example, a data source may provide the monthly number of males and females, whites and African-Americans applying for unemployment. But it is impossible to examine white males, white females, African-American males or African-American females in relation to previous years of employment. Once cases are aggregated, they are "pulled apart" so that individual level comparisons are very limited.

Some analyses can be performed if there is information about the relationship between aggregated units. If one event tended to precede another event by a month, then it may be possible to compare months of occurrence to subsequent events one month later. For example, if comparisons are made between a criminal event and when the offender is arrested and only monthly data are available, it is possible to lag month of occurrence behind month of arrest in the analysis. Hence, the assumption is that most arrests occurred one month later than the month the crime occurred. While this is a reasonable use of aggregated data, it is no substitute for information that provides date of crime occurrence and date arrested for each offender.

Complex Data Structures

Secondary data is sometimes available only as complex data structures. One distinction is between rectangular, or flat, and hierarchal files. The simplest type is rectangular files, in which cases are the rows and variables are the columns, as seen in Table I. Hierarchal files can be illustrated by considering homicides as an event in which there may be any number of offenses, victims, and offenders. An incident record is created with administrative information and links to offense, victim, and offender records, however many of each there may be. The disadvantage is that it is more difficult to analyze hierarchal data sets because until recently most statistical packages assumed the use of flat files.

Inadequate Documentation

Clearly written and specific documentation is essential to using secondary data. But the clarity and specificity is a consequence of the amount of control over the data collection. When a national agency is acting as a clearinghouse in compiling data from local agencies, the national agency has little control over what they receive and relatively little quality control over data collection. On the other hand, when the agency has control over the data collection process, there may be initial detailed documentation, and it is revised and refined on the basis of the continuing data collection process. Thus, the U.S. Census has higher quality data than FBI national reports on crime, which are compiled from local police agencies.

A related problem is whether changes in the reporting system are adequately documented. It is important not only to know what changes occurred, but also to know how continuity in the data series is maintained. For example, the World Health Organization compiles mortality data from a large number of countries. Approximately every decade, representatives from participating countries gather to determine what changes need to be made in their publication the *International Classification of Diseases*. The question then becomes how the newly implemented changes in data collection procedures and definitions are related to previous decades.

The National Center of Health Statistics (NCHS) collects data for the United States according to the *International Classification of Diseases*. To determine the effect of classification revisions, NCHS calculates comparability ratios based on dual codings of the same data using previous and recent classifications. This provides a quantitative indication of the comparability between revisions.

Table I Data Matrix

	Variables						
Cases	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	0	1	1	1	1	1	1
2	1	0	1	1	1	1	1
3	1	1	0	1	1	1	1
4	1	1	1	0	1	1	1
5	1	1	1	1	0	1	1
6	1	1	1	1	1	0	1
7	1	1	1	1	1	1	0

Y = dependent variable; X = independent variable; 1=reported value; and 0=missing value. Taken from Riedel (2000, p. 114) with permission of Sage Publications.

Missing Data

Missing data consists of unit nonresponses and item nonresponses. For an attitude survey; unit nonresponses exist when no respondents are available to provide any information. Item nonresponses occur when there is information about the unit but no information about variables: for example, when information is absent about whether respondents are males or females.

The issue about how much missing data can exist in a data set without biasing the results is an unsettled one. Some experts suggest that 5% or less missing values for a variable will not seriously bias the results. However, large data sets can absorb the loss of information better than small ones. In addition, there are techniques available to determine the effect of missing information.

Table I is useful for understanding the effect of missing data; it is a hypothetical data matrix of seven cases with six independent variables (X_k) and one dependent variable (Y). A general strategy of researchers is to ignore missing data. To do so, the researcher employs either listwise deletion or pairwise deletion. Suppose the researcher decides to analyze this data using a multivariate statistical technique, that is, a statistic that uses all the independent variables (X_1 through X_6) to determine their relative importance in explaining the dependent variable (Y). A general requirement of these techniques is that values for each variable and each unit be completely reported. If one or more values are missing, the entire unit is dropped from the analysis. Listwise deletion occurs in the latter instance, and inspection of Table I indicates the researcher would have no cases to analyze; every case would have one value missing.

Suppose, on the other hand, the researcher decides to use a bivariate statistical technique. Bivariate techniques would compare Y to X_1, then compare Y to X_2, and so on. This would eliminate the problem associated with listwise deletion because only two cases would be lost in each comparison.

But pairwise deletion presents a different sort of problem. Examination of Table I indicates that a comparison of Y and X_1 would use a different subset of data from Y and X_2. In the first comparison, cases 1 and 2 would be

excluded, while in the second comparison, cases 1 and 3 would be excluded. If researchers assume the data are drawn from the same population and the results of bivariate tests can be compared to one another, they would be in error. In fact, comparisons are made from samples drawn from overlapping populations.

One solution to the problem is missing data imputation. Some secondary data sources routinely impute missing data while others do not, and it is not always clear from the documentation whether and what type of imputation has been used. The difficulty with imputed data sets is that the secondary data user does not have access to the original, or nonimputed data. Unless and until the secondary data user understands how the imputation or weighting technique was used to compensate for missing data, a serious bias may be introduced. While imputation is not an improvement over thorough and careful data collection procedures, it represents a important attempt to deal with the problem of missing data.

Research Issues

Designing Back to Front

Research with information gathered for another purpose is unlike research in which researchers gather their own observations. Using secondary data means the approach to research design has to be "back to front." Instead of designing a study and collecting data in accordance with it, the researcher needs to have a detailed knowledge of the characteristics of the data set, then consider that in relation to hypotheses that he or she wants to test.

For research with official records, the first step is to determine the nature and extent of data. One strategy for doing so is to arrange for an interview with the person who manages the records. Using blank documents, the strengths and limitations of agency records can be reviewed with that person. In addition to help with the research design, such a preliminary step allows the researcher to determine the quality of the data before going through procedures necessary to gain access.

Learn the Social and Legal Background

In assessing the quality of the data, it is important to learn the goals of the organizations that collect the data. This provides clues as to the kind of data available, because data collection supports the major activities and mission of agencies. Knowing something about the goals of organizations helps to understand what they define as a "case." For example, the U.S. Department of Education funds a National Public Education Financial Survey in which a state is the unit of analysis.

Organization of the Records

How are the records organized? Are they ordered alphabetically, by the type of event, or by initial action taken? How records are organized is important if the researcher wants to draw a representative sample. If the records are ordered alphabetically, then a simple random sample is sufficient; more complex samples are needed for other types of data organization.

Documentation

How adequate is the documentation? Archival data typically provide more adequate documentation than official statistics taken directly from agencies. The most problematic are official records in which abbreviations or shorthand is used in the records. In the latter instance, it is necessary to discuss with records managers what each of the symbols mean.

A major documentation problem is indications of missing data. Conventionally, missing values is a statement of ignorance: we do not know what the value is. Frequently, blanks are used to indicate missing data, but the meaning of the blank is unclear. For example, if information is missing about an offender's previous criminal history, does that mean he or she had no previous criminal history or does it mean the information was lost or never collected?

Changes in the Record-Keeping System

When a study covers several years, the data has to be evaluated for changes in record keeping. Periodic shifts include elimination or introduction of a new form or revisions in existing forms. In addition, when a new procedure or law has been implemented, what is the effective date that information began to be collected?

Changes in record keeping pose major issues of comparability. Other than NCHS, few organizations perform comparability studies to assess the effect of changes in record keeping. Coping with the changes may mean a simplified analysis. For example, suppose in earlier versions of the data, race is classified as "white" and "nonwhite," and a revision classifies race in several race and ethnic groups. Unless the researcher is satisfied with data only from the revision, he or she has to divide the variable into "white" and "nonwhite."

See Also the Following Articles

Aggregation • Cross-Cultural Data Applicability and Comparisons • Rare Events Research

Further Reading

Acock, A. C. (1997). Working with missing data. *Fam. Sci. Rev.* **10,** 76–102.

Glaser, B. G. (1963). The use of secondary analysis by the independent researcher. *Am. Behav. Sci.* **6,** 11–14.

Hakim, C. (1987). *Research Design: Strategies and Choices in the Design of Social Research.* Allen & Unwin, London.

Hyman, H. H. (1972). *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities.* John Wiley & Sons, New York.

Riedel, M. (2000). *Research Strategies for Secondary Data: A Perspective for Criminology and Criminal Justice.* Sage Publications, Thousand Oaks, CA.

Stewart, D. W. (1984). *Secondary Research: Information, Sources, and Methods.* Sage Publications, Beverly Hills, CA.

Selection Bias

James J. Heckman
The University of Chicago, Chicago, Illinois, USA

Glossary

generalized Roy model A model of sectoral choice with a more general sectoral selection criterion than that of the Roy model.
population distribution A mathematical description of the probability of the outcomes that random variables can assume.
Roy model A model of sectoral choice in which agents choose their sector of employment on the basis of where they get the highest income.
selection bias A bias arising, in general, from nonrandom sampling.
weighted distribution A distribution derived from a population probability distribution by weighting the sampling at some values of the population distribution differently than simple random sampling.

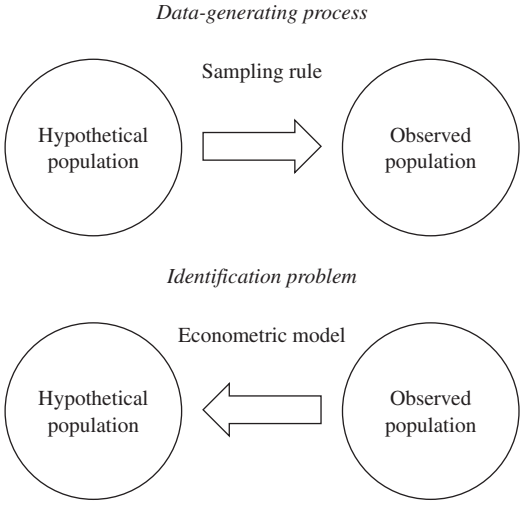


Figure 1 Relationship between hypothetical (counterfactual) populations and observed data.

A random sample of a population produces a description of the population distribution of characteristics. A sample selected by any rule not equivalent to random sampling produces an inaccurate description of the population distribution of characteristics, no matter how big the sample size. The problem of selection bias arises when a rule other than simple random sampling is used to sample the underlying population of interest. The resultant distorted representation of a true population in a sample as a consequence of a sampling rule is the essential source of the selection problem. The identification problem is the problem of recovering features of a hypothetical population from an observed sample. The hypothetical population can refer, for example, to the potential wages of all persons (whether or not they work and have observed wages) or to the potential outcomes of any choice problem when only outcomes from choices made. Selection rule distortions may arise from decisions of sample survey statisticians, or as a consequence of self-selection, so only a subset of possible outcomes is observed. There are two important characterizations of the selection problem. The first, which originates in statistics, involves characterizing the sampling rule depicted in Fig. 1 as applying a weighting to hypothetical population distributions to produce observed distributions. The second, which originates in econometrics, explicitly treats the selection problem as a missing-data problem and uses observables to impute the relevant unobservables.

Weighted Distributions

Any selection bias model can be described in terms of weighted distributions. Let Y be a vector of outcomes of interest and let X be a vector of "control" or "explanatory"

variables. The population distribution of (Y, X) is $F(y, x)$. To simplify the exposition, assume that the density is well defined and write it as $f(y, x)$. Any sampling rule is equivalent to a nonnegative weighting function $\omega(y, x)$ that alters the population density. People are selected into the sampled population by a rule that differs, in general, from random sampling. Let (Y^*, X^*) denote the random variables produced from sampling. The density of the sampled data $g(y^*, x^*)$ may be written as

$$g(y^*, x^*) = \frac{\omega(y^*, x^*) f(y^*, x^*)}{\int \omega(y^*, x^*) f(y^*, x^*)\, dy^*\, dx^*}, \qquad (1)$$

where the denominator of the expression is introduced to make the density $g(y^*, x^*)$ integrate to one as is required for proper densities. Simple random sampling corresponds to the case where $\omega(y, x) = 1$. Sampling schemes for which $\omega(y, x) = 0$ for some values of (Y, X) create special problems because not all values of (Y, X) are sampled. For samples in which $\omega(y, x) = 0$ for a nonnegligible proportion of the population, it is useful to consider two cases. A truncated sample is one for which the probability of observing the sample from the larger random sample is not known. For such a sample, Eq. (1) is the density of all the sampled Y and X values. A censored sample is one for which the probability is known or can be consistently estimated.

In many problems in economics, attention focuses on $f(y \mid x)$, the conditional density of Y given $X = x$. If samples are selected solely on the x variables (selection on the exogenous variables), $\omega(y, x) = \omega(x)$ and there is no problem about using selected samples to make valid inferences about the population conditional density. Sampling on both y and x is termed "general stratified sampling," and a variety of different sampling schemes can be characterized by the structure they place on the weights.

From a sample of data, it is not possible to recover the true density $f(y, x)$ without knowledge of the weighting rule. On the other hand, if the weight $\omega(y^*, x^*)$ is known and the support of (y, x) is known and $\omega(y, x)$ is nonzero, then $f(y, x)$ can always be recovered because

$$\frac{g(y^*, x^*)}{\omega(y^*, x^*)} = \frac{f(y^*, x^*)}{\int \omega(y^*, x^*) f(y^*, x^*)\, dy^*\, dx^*}, \qquad (2)$$

and by hypothesis both the numerator and denominator of the left-hand side are known, and $\int f(y^*, x^*)\, dy^*\, dx^* = 1$, so it is possible to determine $\int \omega(y^*, x^*) f(y^*, x^*)\, dy^*\, dx^*$. It is fundamentally easier to correct for sampling plans with known nonnegative weights or weights that can be estimated separately from the full model than it is to correct for selection when the weights are not known and must be estimated jointly with the model. Choice-based sampling, length-biased sampling, and size-biased sampling are examples of the former. Sampling arising from more general selection models cannot be put in this form because the weights require that the model be known in advance of any analysis of the data. Selection with known weights has been studied under the rubric of the Horvitz–Thompson estimates since the mid-1950s. Contributions to the choice-based sampling literature in economics were made by Manski and McFadden in 1981. Length-biased sampling is analytically equivalent to choice-based sampling and has been studied since the late 19th century by Danish actuaries. Heckman and Singer extended the classical analysis of length-biased sampling in duration analysis to consider models with unobservables dependent across spells and time-varying variables. In their more general case, simple weighting methods with weights determined independently from the model are not available.

The requirements that the support of (y, x) is known and $\omega(y, x)$ is nonzero are not innocuous. In many important problems in economics, the second requirement is not satisfied: the sampling rule excludes observations for certain values of (y, x) and hence it is impossible without invoking further assumptions to determine the population distribution of (Y, X) at those values. If neither the support nor the weight is known, it is impossible, without invoking strong assumptions, to determine whether the fact that data are missing at certain (y, x) values is due to the sampling plan or that the population density has no support at those values. Using this framework, a variety of sampling plans of interest in economics have been analyzed, showing what assumptions they make about the weights and the model to solve the inferential problem of going from the observed population to the hypothetical population.

A Regression Representation of the Selection Problem when there is Selection on Unobservables

A regression version of the selection problem when the weights $\omega(y, x)$ cannot be estimated independently of the model has been devised. Let there be two outcomes, Y_1 and Y_0, corresponding to outcomes in sector 1 and outcomes in sector 0. The outcomes are written as follows:

$$Y_1 = \mu_1(X) + U_1, \qquad (3a)$$

$$Y_0 = \mu_0(X) + U_0. \qquad (3b)$$

The decision rule that characterizes the sector of choice is based on I, a net utility, and

$$I = \mu_I(Z) + U_I, \qquad (3c)$$

$$D = \mathbf{1}[I \geq 0]. \qquad (3d)$$

The special case where $\mu_I(Z) = \mu_1(X) - \mu_0(X)$ and $U_I = U_1 - U_0$ so $I = Y_1 - Y_0$ is the Roy model. In this model, selection only occurs on outcomes; (Y_1, Y_0) are potential outcomes. For simplicity, assume that (U_1, U_0, U_I) are statistically independent of (X, Z) and that (U_1, U_0, U_I) have mean zero. Then $Y = DY_1 + (1-D)Y_0$ and Y_1 (or Y_0, but not both) are observed. In some applications, Y_0 (or Y_1) is never observed. In general,

$$E(Y_1 \mid X, D = 1) \neq \mu_1(X),$$
$$E(Y_0 \mid X, D = 0) \neq \mu_0(X).$$

The observed outcomes are a nonrandom sample of potential outcomes,

$$E(Y \mid X, Z, D = 1) = E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + E(U_1 \mid X, Z, D = 1) \quad (4a)$$

and

$$E(Y \mid X, Z, D = 0) = E(Y_0 \mid X, Z, D = 0) = \mu_0(X) + E(U_0 \mid X, Z, D = 0). \quad (4b)$$

In some cases, only Eq. (4a) or (4b) can be constructed because only Y_1 or Y_0 is observed. The conditional means of U_0 and U_1 are the "control functions," or bias functions. The mean observed outcomes (the left-hand variables) are generated by the means of the potential outcomes plus a bias term. The control function is the bias term.

Define $P(z) = \Pr(D = 1 \mid Z = z)$. As a consequence of the decision rule, Eq. (3d), it has been demonstrated that under general conditions these expressions may always be written as

$$E(Y \mid X, Z, D = 1) = \mu_1(X) + K_1[P(Z)] \quad (5a)$$

and

$$E(Y \mid X, Z, D = 0) = \mu_0(X) + K_0[P(Z)], \quad (5b)$$

where $K_1[P(Z)]$ and $K_0[P(Z)]$ are control functions and depend on Z only through P. The functional forms of the K depend on specific distributional assumptions. The value of P is related to the magnitude of the selection bias. As samples become more representative, $P(Z) \to 1$, $K_1(P) \to 0$. Figure 2 shows a plot of control function $K_1(P)$ versus P. As $P \to 1$, the sample becomes increasingly representative because the probability of any type of person being included in the sample is the same (and $P = 1$). The bias function declines with P. The population mean of Y_1 in samples can be computed with little selection (high P). In general, regressions on selected samples are biased for $\mu_1(X)$. The selection bias term is conflated with the function of interest. If there are variables in Z not in X, regressions on selected samples would indicate that they "belong" in the regression. Equations (5a) and (5b) are the basis for an entire econometric literature on selection bias in regression

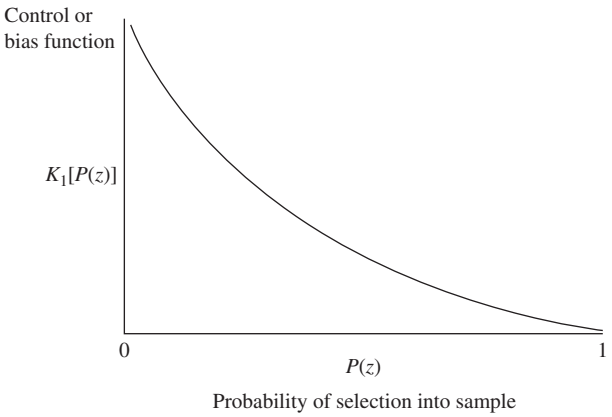


Figure 2 Control function or selection bias as a function of $P(Z)$.

functions. The key idea in all this literature is to control for the effect of P on fitted relationships.

The control functions relate the missing data (the U_0 and U_1) to observables. Under a variety of assumptions, it is possible to form these functions up to unknown parameters and identify the $\mu_0(X)$, $\mu_1(X)$ and the unknown parameters from regression analysis, and control for selection bias. In the early literature, specific functional forms for Eqs. (4) and (5) were derived assuming that the U were joint normally distributed:

Assumption 1 $\quad (U_0, U_1, U_2) \sim N(0, \Sigma)$.

Assumption 2 $\quad (U_0, U_1, U_2) \perp\!\!\!\perp (X, Z)$.

Assumption 1 coupled with Assumption 2 produces precise functional forms for K_1 and K_0. There is a two-step estimation procedure for censored samples: (1) Estimate $P(Z)$ from data on the decision to work and (2) using an estimated $P(Z)$, form $K_1[P(Z)]$ and $K_0[P(Z)]$ up to unknown parameters. Then Eqs. (5a) and (5b) can be estimated using regression. This produces a convenient expression linear in the parameters when $\mu_1(X) = X\beta_1$ and $\mu_0(X) = X\beta_0$. A direct one-step regression procedure has been developed for truncated samples. Equations (5a) and (5b) are the basis for an entire literature that generalizes and extends the early models, and remains active to this day.

Identification

Much of the econometric literature on the selection problem combines discussions of identification (going from populations generated by selection rules back to the source population) with discussions of estimation in solving the inferential problem of going from observed samples to hypothetical populations. It is analytically useful to distinguish the conditions required to identify the

selection model from ideal data from the numerous practical and important problems of estimating the model. Understanding the sources of identification of a model is essential to understanding how much of what is being obtained from an empirical model is a consequence of what is put into it.

Paul Holland used the law of iterated expectations to write the conditional distribution of an outcome, say Y_1 on X, in the following form:

$$F(Y_1 \mid X) = F(Y_1 \mid X, D = 1)\ \Pr(D = 1 \mid X) + F(Y_1 \mid X, D = 0)\ \Pr(D = 0 \mid X). \quad (6)$$

It is possible to observe Y_1 only if $D = 1$. In a censored sample, it is possible to identify $F(Y_1 \mid \mathrm{X},\ D = 1)$, $\Pr(D = 1 \mid X)$ and hence $\Pr(D = 0 \mid X)$; Y_1 is not observed when $D = 0$. Hence, $F(Y_1 \mid \mathrm{X})$ is not identified. Independent work of James Smith and Finis Welch made a similar decomposition of conditional means (replacing F with E).

Holland questioned how $F(Y_1 \mid X)$ could be identified and compared selection models with other approaches. Smith and Welch discussed how to bound $F(Y_1 \mid X)$ or $E(Y_1 \mid X)$ by placing bounds on the missing components, $F(Y_1 \mid X, D = 0)$ and $E(Y_1 \mid X, D = 0)$, respectively. (Smith and Welch use their analysis to bound the effects of dropping out on the black–white wage gap.) A clear precedent for this idea was the work of Peterson, who developed nonparametric bounds for the competing risk model of duration analysis, which is mathematically identical to the Roy model, which is the model of Eqs. (3a)–(3d) when $I = Y_1 - Y_0$. The competing risks model replaces $\max(Y_0, Y_1)$ with $\min(Y_0, Y_1)$ for selecting outcomes. In this model, $D = 1$ if $I = Y_0 - Y_1 > 0$.

The normality assumption widely made in the early literature has been called into question. Arabmazar and Schmidt presented Monte Carlo analyses of models showing substantial bias for models with continuous outcomes when normality was assumed but the true model was nonnormal. The empirical evidence is more mixed. Normality is not a bad assumption for analyzing models of self-selection for log wage outcomes once allowance is made for truncation and self selection. Normality of latent variables turns out to be an acceptable assumption for discrete-choice models except under extreme conditions.

Heckman and Honoré have considered identification of the Roy model ($I = Y_1 - Y_0$ in the notation of Eqs. (3a)–(3d)) under a variety of conditions. They establish that under normality, the model is identified even if there are no regressors, so there are no exclusion restrictions. They further establish that the model is identified (up to subscripts) even if only Y is observed, but analysts do not know if it is Y_1 or Y_0. The original normality assumption used in selection models was based on powerful functional form assumptions. They develop a nonparametric Roy model and establish conditions under which variation in regressors over time or across people can identify the model nonparametrically. The distributional assumptions can be replaced with different types of variation in the data to identify the Roy version of the selection model. Heckman and Smith and Carneiro, Hansen and Heckman extend this line of analysis to the generalized Roy model where the decision is based on a more general I. It turns out that the decision rule with $I = Y_1 - Y_0$ plays a crucial role in securing identification of the selection model. In a more general case, when I may depend on $Y_1 - Y_0$ but on other unobservables as well, even with substantial variation in regressors across persons or over time, only partial identification of the full selection model is possible. When the models are not identified, it is still possible to bound crucial parameters, and an entire literature has grown up elaborating this idea.

Bounding and Sensitivity Analysis

Starting from Eq. (6) or its version for conditional means, the work of Smith and Welch, Holland, and Glynn, Laird, and Rubin characterized the selection problem more generally without the structure of Eqs. (3a)–(3d), offering Bayesian and classical methods for performing sensitivity analyses for the effects of different identifying assumptions on inferences about population mean.

Selection on observables solves the problem of selection by assuming that Y_1 is independent of D given X, so $F(Y_1 \mid \mathrm{X},\ D = 1) = F(Y_1 \mid X)$. This is the assumption that drives matching models. It is inconsistent with the Roy model of self-selection. Various approaches to bounding this distribution, or moments of the distribution, have been proposed in the literature, all building on insights by Holland and by Peterson. To illustrate these ideas in the simplest possible setting, let $g(Y_1 \mid X,\ D = 1)$ be the density of outcomes (e.g., wages) for persons who work ($D = 1$ corresponds to work). Assume censored samples. Missing is $g(Y \mid X,\ D = 0)$ (e.g., the density of the wages of nonworkers).

In order to estimate $E(Y_1 \mid X)$, Smith and Welch used the law of iterated expectations to obtain

$$E(Y_1 \mid X) = E(Y_1 \mid X, D = 1)\ \Pr(D = 1 \mid X) + E(Y_1 \mid X, D = 0)\ \Pr(D = 0 \mid X).$$

To estimate the left-hand side of this expression, it is necessary to obtain information on the missing component $E(Y_1 \mid X,\ D = 0)$. Smith and Welch proposed and implemented bounds on $E(Y_1 \mid X,\ D = 0)$; for example,

$$Y_{\mathrm{L}} \leq E(Y_1 \mid X, D = 0, Z) \leq Y^{\mathrm{U}},$$

where Y^{U} is an upper bound and Y_{L} is a lower bound. (In their problem, there are plausible ranges of wages

that dropouts can earn.) Using this information, the following bounds were constructed:

$$E(Y_1 \mid X, D = 1)\ \Pr(D = 1 \mid X) + Y_{\mathrm{L}}\ \Pr(D = 0 \mid X) \leq E(Y_1 \mid X) \leq E(Y_1 \mid X, D = 1)\ \Pr(D = 1 \mid X) + Y^{\mathrm{U}}\ \Pr(D = 0 \mid X).$$

A sensitivity analysis produces a range of values for $E(Y \mid X)$ that are explicitly dependent on the range of values assumed for $E(Y \mid X, D = 0)$. Later work has developed this type of analysis more systematically for a variety of models.

Glynn, Laird, and Rubin present a sensitivity analysis for distributions using Bayesian methods under a variety of different assumptions about $F(Y_1 \mid X, D = 0)$ to determine a range of values of $F(Y \mid X)$. Holland proposes a more classical sensitivity analysis that varies the ranges of parameters of models. The objective of these analyses of bounds and the Bayesian and classical sensitivity analyses is to clearly separate what is known from what is conjectured about the data, and to explore the sensitivity of reported estimates to the assumptions used to secure them. Others have demonstrated the extra restrictions that come from using the structure of Eqs. (3a)–(3d) to produce bounds on outcomes.

Much of the theoretical analysis presented in the recent literature is nonparametric, although in practice, much practical experience in statistics and econometrics demonstrates that high-dimensional nonparametric estimation is not feasible for most sample sizes available in cross-sectional econometrics. Some form of structure must be imposed to get any reliable nonparametric estimates. However, feasible parametric versions of these methods run the risk of imposing false parametric structure. The methods used in the bounding literature depend critically on the choice of conditioning variables X. In principle, all possible choices of the conditioning variables should be explored, especially in computing bounds for all possible models, although, in practice, this is never done. If it were, the range of estimates produced by the bounds would be substantially larger than the wide bounds already reported.

Acknowledgment

This work has been supported by National Science Foundation Grant SES-0241858.

See Also the Following Articles

Bayesian Statistics • Population vs. Sample • Weighting

Further Reading

Arabmazar, A., and Schmidt, P. (1981). Further evidence on the robustness of the Tobit estimator to heteroskedasticity. *J. Econometr.* **17,** 253–258.

Ahn, H., and Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econometr.* **58,** 3–29.

Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* **59,** 307–345.

Arabmazar, A., and Schmidt, P. (1981). Further evidence on the robustness of the Tobit estimator to heteroskedasticity. *J. Econometr.* **17**(November 1981), 253–358.

Bera, A. K., Jarque, C. M., and Lee, L.-F. (1984). Testing the normality assumption in limited dependent variable models. *Int. Econ. Rev.* **25,** 563–578.

Carneiro, P., Hansen, K., and Heckman, J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *Int. Econ. Rev.* **44,** 361–422.

Glynn, R., Laird, N., and Rubin, D. (1986). Selection modeling vs. mixture modeling with nonignorable response. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.), pp. 119–146. Springer-Verlag, New York.

Gronau, R. (1974). Wage comparisons—A selectivity bias. *J. Pol. Econ.* **82,** 1119–1144.

Heckman, J. (1974). Shadow prices, market wages and labor supply. *Econometrica* **42,** 679–694.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Social Measure.* **5,** 475–492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47,** 153–161.

Heckman, J. (1980). Addendum to sample selection bias as a specification error. In *Evaluation Studies Review Annual* (E. Stromsdorfer and G. Farkas eds.), Vol. 5, pp. 69–74. Sage Publ., San Francisco.

Heckman, J. (1987). Selection bias and the economics of self selection. In *The New Palgrave: A Dictionary of Economics* (J. Eatwell, M. Milgate, and P. Newman, eds.), pp. 287–296. Stockton, New York.

Heckman, J. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *Q. J. Econ.* **115,** 45–97.

Heckman, J. (2001). Micro data, heterogeneity and the evaluation of public policy: Nobel lecture. *J. Pol. Econ.* **109,** 673–748.

Heckman, J., and Honoré, B. (1990). The empirical content of the Roy model. *Econometrica* **58,** 1121–1149.

Heckman, J., and MaCurdy, T. (1986). Labor econometrics. In *Handbook of Econometrics* (Z. Grilches and M. D. Intriligator eds.), Vol. 3, Chap. 3, pp. 1917–1977. Elsevier, New York.

Heckman, J., and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data* (J. Heckman and B. Singer, eds.), pp. 156–245. Cambridge University Press for Econometric Society Monograph Series, New York.

Heckman, J., and Robb, R. (2000). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In *Drawing Inferences From Self-Selected Samples* (H. Wainer, ed.), pp. 63–107. Lawrence Erlbaum, Mahwah, New Jersey.

Heckman, J., and Singer, B. (1985). Social science duration analysis. In *Longitudinal Analysis of Labor Market Data* (J. Heckman and B. Singer, eds.), pp. 39–110. Cambridge University Press, Cambridge, U.K.

Heckman, J., and Smith, J. (1998). Evaluating the welfare state. In *Econometrics and Economics in the 20th Century: The Ragnar Frisch Centenary* (S. Strom, ed.), pp. 241–318. Cambridge University Press for Econometric Society Monograph Series, New York.

Heckman, J., and Vytlacil, E. (2000). Local instrumental variables. In *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Theory and Econometrics: Essays in Honor of Takeshi Amemiya* (C. Hsiao, K. Morimune, and J. Powell, eds.), pp. 1–46. Cambridge University Press, Cambridge, U.K.

Heckman, J., and Vytlacil, E. (2005). Econometric evaluation of social programs. In *Handbook of Econometrics* (J. Heckman and E. Leamer, eds.), Vol. 6. North Holland, Amsterdam.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66,** 1017–1098.

Holland, P. (1986). A comment on remarks by Rubin and Hartigan. In *Drawing Inferences from Self Selected Samples* (H. Wainer, ed.), pp. 149–151. Springer-Verlag, New York.

Manski, C. F. (1994). The selection problem. In *Advances in Econometrics: Sixth World Congress* (C. Sims, ed.), pp. 143–170. Cambridge University Press, Cambridge, U.K.

Manski, C. F. (1995). *The Identification Problem in the Social Sciences.* Harvard University Press, Cambridge.

Manski, C. F., and McFadden, D. (1981). Alternative estimators and sample designs for discrete choice analysis. In *Structural Analysis of Discrete Data With Econometric Applications* (C. Manski and D. McFadden, eds.). MIT Press, Cambridge.

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics* (R. Engle and D. McFadden, eds.), pp. 2111–2245. Elsevier-North Holland, New York.

Peterson, A. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proc. Nat. Acad. Sci.* **73,** 11–13.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Distributions* (G. Patil, ed.), pp. 311–324. Pergamon Press, Calcutta.

Rao, C. R. (1985). Weighted distributions. In *A Celebration of Statistics* (A. Atkinson and S. Fienberg, eds.). Springer-Verlag, Berlin.

Robins, J. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS* (L. Sechrest, H. Freeman, and A. Mulley, eds.), pp. 113–159. Public Health Service, Washington, D.C.

Rosenbaum, P. (1995). *Observational Studies,* Chap. 4. Springer-Verlag, Leipzig, Germany.

Sheps, M., and Menken, J. (1973). *Mathematical Models of Conception and Birth.* University of Chicago Press, Chicago.

Smith, J., and Welch, F. (1986). *Closing The Gap: Forty Years of Economic Progress for Blacks.* Rand Corporation, Santa Monica.

Sellin–Wolfgang Scale of Severity

Paul E. Tracy
University of Texas, Dallas, Richardson, Texas, USA

Glossary

complex criminal event A crime can involve multiple and separate violations, including bodily injury, property theft, property damage, and intimidation by verbal threat or a weapon.
crime severity Quantitative aspects of the seriousness of a criminal event as measured through kinds and amount of actual harm caused by the crime.
hierarchy rule Federal Bureau of Investigation method of counting crimes by using only the most serious offense regardless of how many other crimes were committed during the same crime event.
ratio scale A measurement scale that not only uses equidistant points along the scale but also has a meaningful zero point. Ratio scales are the most sophisticated scales because they incorporate all the characteristics of nominal, ordinal, and interval scales. As a result, a large number of statistical calculations are applicable.
Sellin–Wolfgang scale A ratio scale of the severity of a criminal event that uses a weighting scheme to score all the components of a crime.

This article discusses the measurement of the gravity, severity, or seriousness of crime through the use of the Sellin–Wolfman scale. The characteristics of this ratio scale of crime severity are described, examples of the scoring system are presented, and potential uses of the scale for research and policy are discussed.

Classification and Measurement of Crime

Social scientists and criminal justice officials have long recognized the need for precise and accurate indicators of the amount of criminal behavior existent in a given place and time. Without such measures of crime, it would be difficult, if not impossible, to determine with any degree of certainty the level of criminal activity and the effectiveness of efforts to address crime. General agreement exists among scholars and practitioners that adequate measures of crime are necessary not only for testing scholarly research hypotheses but also for the evaluation of criminal justice policies and procedures and the rational allocation of an increasingly scarce pool of criminal justice system resources. As a consequence of this general requirement for high-quality social indicators surrounding crime, the escalation in criminal justice-related statistical information has been dramatic in the past 10–20 years. In fact, the growth in the amount and complexity of criminological research in no small way has resulted from the burgeoning of the statistical data made available to researchers by federal agencies. Furthermore, the audience for this crime-related information has increasingly included the general public as well as academics, legislators, and criminal justice officials.

Although numerous crime-related data are available, the most common and widely used database is the Uniform Crime Report (UCR) compiled by the Federal Bureau of Investigation (FBI). The UCR uses a scheme called the Standard Classification of Offenses (SCO) for classifying criminal behavior. The SCO system classifies criminal events in two categories: (i) index or part I offenses and (ii) nonindex or part II offenses. The seven offenses in the index offense category comprise the well-known "crime index" and consist of nonnegligent criminal homicide, rape, robbery, aggravated assault, burglary, larceny, and vehicle theft. All other offenses, from simple assault to parking violations, are contained in the part II category. With very few modifications, this system has been the basis of the UCR reporting system since its inception in 1930.

The basic rationale for the adoption of the part I offenses as the basis of a crime index was the assumption that these so-called "serious" offenses would be reported to the police most often and thus constitute the closest measure of the total amount of crime that is committed. Thus, the sum of the seven index offenses is treated as the volume of serious crime that is known to the police. Although this system seems reasonably capable of producing an index of crime, an index that details both the volume and the seriousness of criminal behavior, the method used to classify and count offenses renders the system misleading, if not erroneous, in several respects.

First, the index classification system does not provide for multiple offenses (i.e., a complex criminal event that comprises several distinct crimes). That is, according to the hierarchy principle of the SCO, only the crime that has the highest rank order in the list of ordered categories shall be counted. Thus, for example, an incident composed of a rape, an aggravated assault, and a robbery would be recorded for UCR index purposes as only one crime—rape—because rape has the highest rank order of the three crimes committed.

Second, the classification of offenses according to the broad legal label attached to them ignores the fact that each legal category consists of a variety of offenses that should not be equated and contribute the same amount to the crime rate. A robbery, for instance, may be the armed holdup of one or more persons, the infliction of serious harm, and the theft of large sums of money. On the other hand, a robbery may also be the taking of a child's lunch money by a schoolmate. Many criminal acts that lie between these two extremes would all be classified as robberies, regardless of the degree of injury or the amount of property loss. The other index offenses are similarly affected by the broad continuum of behavior that is subsumed under each category.

Third, the SCO does not differentiate between criminal events that are successful and those that are merely attempted. Equating these two categories clearly masks the amount of actual harm or loss incurred by the community and gives a distorted view of crime severity.

Last, there is no weighting system in the compilation of the index crime rate. Thus, two auto thefts are allowed to contribute as much to the crime rate as do two homicides.

The method of classifying and counting criminal offenses for index purposes just described has two overriding deficiencies. First, by counting only one offense when at least two are conjoined, and by using an arbitrary set of ordered categories, the UCR reporting system provides only a partial enumeration of the specific offense actually known to the police. This clearly provides misleading data concerning the actual volume of criminal behavior. Second, by equating all offenses that carry the same generic legal label, and by confounding completed and attempted acts, considerable differences in the degree of seriousness of various offenses are concealed. In other words, the UCR method provides no solution to the problem of how to deal statistically with a complex of offenses or with simple offenses that vary appreciably in seriousness but that carry the same legal title.

Wolfgang was among the first to express this skepticism regarding the characteristics of the UCR classification system and raise doubts about the usefulness of the crime index system. Wolfgang and his colleague and mentor, Thorsten Sellin, at the Center for Studies in Criminology and Criminal Law at the University of Pennsylvania, determined that because the FBI's measurement approach misrepresents and even masks the actual volume and seriousness of criminal behavior, an alternative measurement scheme needed to be developed. Their goal was to capture both the quantitative and the qualitative dimensions of criminal behavior and, thus, offer greater substantive utility to research and policy making on crime.

The First Sellin–Wolfgang Scale: 1964

At the time when Sellin and Wolfgang sought to develop a scale of the severity, gravity, or seriousness of crime, efforts to scale various social phenomena were not new. The work of Likert, Guttman, and others had been widely employed in creating categorical and ordered scales. However, Sellin and Wolfgang determined that such scales were inadequate for weighting the amount of harm in a criminal event because they have neither a zero point to anchor the scale nor can distances along the scale be reliably determined. For example, although it is relatively easy to determine that a homicide is a more serious crime than a rape or a robbery, the usual scaling methods do not adequately capture the severity differentials or distance among these crimes. Furthermore, even within a particular offense type, such as robbery, available scales do not allow a researcher to measure the crime severity differentials among a robbery with injury and theft, a robbery with just injury, and a robbery with just theft. Sellin and Wolfgang thus sought to develop a ratio scale of perceived severity to overcome these shortcomings by generating a continuous measure of crime weighted for the severity of the event.

Sellin and Wolfgang turned to the literature on psychological scaling and found a variety of procedures for developing a ratio scale. They noted that Thurstone's 1927 method of paired comparisons had been widely adopted in psychological and social psychological measurement, but a Thurstone technique was not feasible owing to the fact that Sellin and Wolfgang anticipated that a very large number of offense comparisons were needed to capture

the various components of offense severity. Fortunately, Stevens and Galanter had recently made significant contributions in the field of psychophysics by developing a less complex ratio scaling method based on magnitude estimation procedures. The magnitude estimation method is a procedure in which a subject makes direct numerical estimates of a series of subjective impressions. Usually, the subject is presented with a base stimulus, known as the modulus, which may have an arbitrary score of 10. Subjects then receive another stimulus and judge its intensity compared to the modulus using any range of numbers they choose. If a subject believes that the new stimulus is twice as intense as the modulus, then a score of 20 would be reported for this item, whereas if the new item is deemed to be half as intense as the modulus then a score of 5 would be assigned.

Scaling the Gravity of Crime

Armed with this new technique for ratio scale construction, Sellin and Wolfgang began their effort to develop a quantitative measure of crime severity by extracting the components or aspects of criminal conduct from the Philadelphia Crime Code. Although Sellin and Wolfgang questioned the value of the particular crime index developed by the FBI, they were in full agreement that a crime index system must be based on certain components of offensive conduct. However, instead of selecting these crime components on the basis of the title given them by the criminal code, they believed that the nature of the harm inflicted in the criminal event should govern the selection of an index. Thus, they concluded that a scale of offense gravity should be constructed utilizing events that involve violations of the criminal law that inflict bodily harm on one or more victims and/or cause property loss by theft, damage, or destruction. Sellin and Wolfgang maintained that these harmful effects were more crucial to the establishment of an index of crime severity than the specific legal labels attached to the events. The Sellin–Wolfgang criterion of selecting events for a crime index differs in two major respects from the one reflected in the UCR system. First, it does not allow the inclusion of offenses that produce none of the components of harm just described. Thus, the offenses utilized in their scale all share one very important feature—some degree of measurable social harm to the community. Second, their system includes many offenses that are not counted among the index crimes category in the UCR. Simply stated, Sellin and Wolfgang chose the criterion of discernable consequences over that of an ordered set of legal categories that may or may not appropriately reflect the seriousness of criminal behavior.

Sellin and Wolfgang also determined that the class of violations to be utilized in the scoring system should be subdivided into three categories in order to indicate the major effect associated with the offense. The first category includes those events that produce bodily harm to a victim or to victims even though property theft or damage may also be involved. The second class of events consists of those offenses that do not involve injury but have a property theft component even when accompanied by damage. The last category consists of those offenses that involved only damage to property. In addition, because they believed that an event should not be evaluated solely in terms of the injuries and losses that occur, the system takes account of certain other factors in the event that aggravate the crime. For example, a crime is aggravated if the offender engages in intimidating behavior (either verbal/physical or by weapon). Furthermore, a property crime may be aggravated if the offender damages the premises by forcible entry. Thus, the crime severity scale takes account of both the components (injury, theft, and damage) and the aggravating factors (intimidation and premises forcibly entered).

Developing the 1964 Scale

Sellin and Wolfgang used these components of offense severity to develop 141 single-sentence offense descriptions or scenarios that were shown to a pilot group of 17 raters (university students). The raters were instructed to rate the criminal event on the basis of their perception of its severity or seriousness on a 7-point scale of severity. The small pilot study was deemed a success because it demonstrated that the offense descriptions were adequate and subjects could render comparative judgments.

The second stage of the research involved a determination by Sellin and Wolfgang as to the type of scaling technique to be used to develop the actual scale of crime severity. They followed the suggestion of Stevens and Galanter and employed both an 11-point category scale and a magnitude estimation ratio scale for comparative validation purposes. The second stage also involved the selection of the raters who would make the judgments concerning crime severity. They considered a variety of subject groups and selected a total of 569 respondents: 286 police officers, 245 university students, and 38 juvenile court judges. The ratings of the 141 offense scenarios by these respondents constitute the "primary index scale." Sellin and Wolfgang then compared the results of the 11-point category scale with the magnitude estimation ratio scale and found that for each of the rating groups, there was remarkable consistency when the category scores were plotted against the magnitude values. In light of this consistency, Sellin and Wolfgang could confidently adopt the magnitude estimation scores as the basis for constructing the scale, and thus derive a set of ratio scores by which the relative gravity of crimes could be documented.

Table I Offense Scores and Ratio Values, l964

Offense scenario	*Mean magnitude scale values*	*Ratio score*[a]
Larceny $1	16.93	1
Larceny $5	22.09	1
Larceny $20	27.77	2
Larceny $50	32.31	2
Larceny $1000	52.99	3
Larceny $5000	69.13	4
Burglary $5	40.62	2
Robbery $5 (no weapon)	52.25	3
Robbery $5 (weapon)	86.33	5
Assault (death)	449.20	26
Assault (hospitalized)	115.60	7
Assault (treated and discharged)	69.32	4
Assault (minor harm)	22.50	1
Rape (forcible)	186.30	11
Vehicle theft (no damage)	27.19	2
Forcible entry	18.53	1
Intimidation (verbal)	30.15	2
Intimidation (weapon)	64.24	4

[a] Ratio scores are derived by dividing the geometric mean of an offense scenario by the geometric mean of larceny of $1 (i.e.,16.93).

Among the set of 141 offense scenarios was a set of 18 core items that Sellin and Wolfgang used to vary systematically the various aspects of crime severity so as to derive the weights for the various components that must be scored to compute a severity score for a criminal event. The actual scale values were computed as follows and displayed in Table I. The appropriate measure of central tendency for ratio judgments is the geometric mean that in practice is calculated by taking the antilog of the arithmetic mean of the logarithms of the responses, or the antilog of

$$\frac{\sum_{I=1}^{N} \log X_I}{N}.$$

Replications

The work of Sellin and Wolfgang has been replicated on several occasions both in the United States and in other countries, which confirms the validity of the principle that the severity of crime can be scaled using subjective judgment procedures. These replications were first done internationally; in Canada by Normandeau, in Puerto Rico by Velez-Diaz and Megargee, and in Taiwan by Hsu. In the United States, Rossi *et al.*, Figlio, and Wellford and Wiatrowski conducted similar studies of crime severity judgments.

The Second Sellin–Wolfgang Scale: A National Survey of Crime Severity

In the early 1980s, Wolfgang and colleagues at the Criminology Center at the University of Pennsylvania decided to update the original scale because approximately 5 years had passed since the first scale had been developed and they hypothesized that changes in the public's perception of crime might have changed. The research that ensued employed the same scaling techniques as its predecessor, but the researchers decided to expand the number of offense scenarios and use a nationally representative sample. The major purposes of the new survey were (i) to determine at the national level the public's perception of the relative severity of various kinds of crime; (ii) to investigate the perceived severities of criminal offenses according to regions, states, size of place, and a range of sociodemographic characteristics of the population; and (iii) to determine if the data generated by the survey would produce a structure resembling a quantitative scale similar to that previously reported in the literature.

The National Survey of Crime Severity (NSCS) was administered as a supplement to the National Crime Victimization Survey (NCVS), an ongoing national survey sponsored by the Bureau of Justice Statistics. The purpose of the NSCS was to measure household and individual victimizations by the major crimes of assault, burglary, robbery, larceny, and vehicle theft. The NCVS has utilized a rotating sample design with approximately 60,000 households interviewed over a 6-month period. The NCSC estimates were based on data collected during the period July through December 1977. The NCSC sample was a 50% sample of the NCVS full sample and was spread over 376 sample areas with coverage in each of the 50 states and the District of Columbia. The national-level magnitude estimation scores and ratio values that comprise the severity scale weights are shown in Table II.

The Sellin–Wolfgang Severity Scoring System

Offense Components

In order to score criminal events, the following items, insofar as they are applicable to a given event, must be collected and recorded so that the component of harm can be scored.

Number of Victims Injured

Each victim who receives some degree of bodily injury during a criminal event must be accounted for. Physical injuries usually occur as a direct result of assaultive events,

Table II Offense Scores and Ratio Values, 1985

Offense scenario	*Mean magnitude scale values*	*Ratio score*[a]
Larceny $1	21.9	1.0
Larceny $10	37.8	1.7
Larceny $50	63.0	2.9
Larceny $100	78.5	3.6
Larceny $1,000	150.2	6.9
Larceny $10,000	239.3	10.9
Burglary and theft $10	70.6	3.2
Robbery $10 (verbal threat)	144.8	6.6
Robbery $10 (weapon)	180.0	7.3
Assault (death)	778.4	35.6
Assault (hospitalized)	261.4	12.0
Assault (treated and discharged)	186.0	8.5
Assault (minor harm)	32.2	1.5
Rape (forcible)	565.6	25.8
Vehicle theft (recovered)	97.7	4.5
Vehicle theft (not recovered)	176.7	8.1
Forcible entry	43.6	1.5

[a] Ratio scores are derived by dividing the geometric mean of an offense scenario by the geometric mean of larceny of $1 (i.e., 21.9).

but they may be a by-product of other events as well. The following are the four levels of bodily injury:

Minor harm: An injury that requires or receives no professional medical attention. The victim may, for instance, be pushed, shoved, kicked, knocked down, and receive a minor wound (cut, bruise, etc.).

Treated and discharged: The victim receives professional medical treatment but is not detained for further medical care.

Hospitalization: The victim requires inpatient care in a medical facility, regardless of its duration, or outpatient care for three or more clinical visits.

Killed: The victim dies as a result of the injuries, regardless of the circumstances in which they were afflicted.

Number of Victims of Acts of Forcible Sexual Intercourse

This event occurs when a person is intimidated and forced against his or her will to engage in a sexual act (e.g., rape, incest, and sodomy). Such an event may have more than one victim, and the score depends on the number of such victims. A continuous relationship such as may occur in forcible incest is to be counted as one event. A forcible sex act is always accomplished by intimidation. Thus, the event must also be scored for the type of intimidation involved. Intimidation is scored for all victims in a forcible sex act (such is not the case for other events). The victim of one or more forcible sex acts during an event is always assumed to have suffered at least minor harm. Even when medical examination may not reveal any injuries, the event must be scored for minor harm. This level of injury should also be scored (rather than treated and discharged) when the victim is examined by a physician only in order to ascertain if venereal infection has occurred or to collect evidence that the sex act was completed.

Physical or Verbal Intimidation or Intimidation by a Dangerous Weapon

This is an element in all events in which one or more victims are threatened with bodily harm (or some other serious consequences) for the purpose of forcing the victim(s) to obey the request of the offender(s) to give up something of value, to assist in a criminal event that leads to someone's bodily injury and/or property theft or damage. In addition to rape, robbery is a classic example. Ordinary assault and battery, aggravated assault and battery, or homicide are not to be scored for intimidation merely because someone was assaulted or injured. The event must have also included the threat of force for intimidation to have been present. With the exception of forcible sex acts, criminal events involving intimidation are scored only once regardless of the number of victims who are intimidated. The types of intimidation are:

Physical or verbal: Physical intimidation means the use of strong-arm tactics such as threats with fists and menacing gestures. Verbal intimidation means spoken threats only, not supported by the overt display of a weapon.

Intimidation by weapon: Display of a weapon, such as a firearm, cutting or stabbing instrument, or blunt instrument capable of inflicting serious bodily injury.

Number of Premises Forcibly Entered

As used here, forcible entry means unlawful entry, even when not by "breaking" into premises of a private character to which the public does not have free access or the breaking and entering into premises to which the public ordinarily has free access. Such an entry is, in itself, an event to be scored if it causes some degree of damage to property (e.g., a broken lock, window, or door) even though it is not followed necessarily by an injury to a person or by a theft of and damage to property inside the premises.

Usually only one distinct premise will be entered, such as a family dwelling, an apartment, or a suite of offices, but some events may embrace several such entries. The scoring depends on the number of premises forcibly entered during the event and occupied by or belonging to different owners, tenants, or lessees. Contrary to the "hotel rule" used in the UCR, each hotel, motel, or lodging house room broken into and occupied by different tenants should be scored. If a building was forcibly entered

Table III Crime Severity Scoring Sheet (1985 Weights)

Effects of event: I (injury) T (theft) D (damage) (Mark all that apply)			
Crime severity component	*No. of victims*	× *Scale weight*	= *Total*
1. Injury			
a. Minor harm	_____	1.47	_____
b. Treated and discharged	_____	8.53	_____
c. Hospitalized	_____	11.98	_____
d. Fatality	_____	35.67	_____
2. Forcible sex acts	_____	25.92	_____
3. Intimidation			
a. Verbal or physical	_____	4.90	_____
b. Weapon	_____	5.60	_____
4. Premises forcibly entered	_____	1.50	_____
5. Motor vehicles stolen			
a. Recovered	_____	4.46	_____
b. Not recovered	_____	8.07	_____
6. Property theft and damage[a]	_____		_____
	Total severity score		

[a] The severity score for any value of theft or damage is produced as follows: $\log_{10}{}^{Y} = 0.26776656 \times \log_{10}{}^{X}$; where Y is the crime severity weight, and X is the total dollar value of theft or damage.

and further entries were made inside, the total number of entries scored should include the forcible entry of the building even when the building belongs to someone who is victimized by a further entry inside.

Number of Motor Vehicles Stolen and Recovery Status

As used here, motor vehicle means any self-propelled vehicle—automobile, truck, motorcycle, tractor, or airplane. Disregard self-propelled lawn motors and similar domestic instruments in this section; the value of such items is accounted for in the theft/damage section. Because motor vehicles may be either stolen and recovered or stolen and never returned to the legal owner, the number of vehicles in each category must be accounted for separately and will receive a different score value.

Total Dollar Amount of Property Loss through Theft or Damage

Regardless of the kind of event scored and the number of victims, the total value of all property stolen or damaged must be determined, whether it is wholly or partially recovered and whether or not the loss is covered by insurance.

Motor vehicle thefts require special handling. The score of the event does not depend on the value of the vehicle stolen. Thus, the dollar value of the vehicle is ignored in this element. However, if the vehicle is recovered damaged and/or property has been taken from it, the loss is the sum of the cost of the damage and the value of the stolen articles.

The Crime Severity Scale Illustrated

The offense components discussed previously constitute the scale items in the Sellin–Wolfgang scale of the gravity of the crime. The scoring system used to evaluate the seriousness of crime can best be presented by first describing the elements of the system and then illustrating the scoring procedure with hypothetical offenses. Table III depicts the elements of the system. The first item that must be collected is the identification number. This is the number given to a particular criminal event. It may be a central complaint number, a district number, or some similar designation. If the same event is represented by more than one such number, all numbers should be recorded so that the event can be scored as a whole. In most cases, an event will be described in complaint or investigation reports carrying but one identifying number. In some cases, however, one event may become the subject of reports with different numbers (two or more such reports describing the same event). For instance, in a rape event with two victims, each victim may file his or her own complaint and thus it would be necessary to coordinate the separate reports before the event could be scored.

In order to classify the event, the presence of I (injury), T (theft), and D (damage) components must be determined. Because the construction of subindices is often necessary, as many of the components as apply should be circled. From this procedure, it is possible to derive six classifications of an event: I, T, D, IT, TD, and ITD. It is possible, therefore, to use this classification scheme as a solution to the problem of dealing with the complex criminal event in which more than one offense type is present simultaneously.

Following the determination of the class to which the event belongs, the event is scored for seriousness. Column 1 lists the various offense components and the particular levels of each. Column 2 refers to the number of victims who experience each level of the offense components. The exceptions to the rule of accounting for the number of times each component occurs involve criminal events other than rape in which intimidation is present and in which this component is scored only once regardless of the number of victims and the value of property loss that is summed across all victims. Column 3 gives the scale weight assigned to each element of the offense. Column 4 is reserved for the total score for a given component; this is obtained by multiplying the number in Column 2 (where applicable) by the weight listed in Column 3. By adding all the numbers in Column 4, the total score for the event is computed.

Illustrations of how the proposed scoring system works are given next. For the purpose of showing how it differs from that of the UCR system, the problems have been adapted from the "Uniform Crime Reporting Handbook" issued by the FBI. The problems as originally listed there generally do not contain all the necessary information. Therefore, hypothetical data have been supplied in parentheses.

Problem 1

A holdup man forces a husband and his wife to get out of their automobile. He shoots the husband, gun whips and rapes the wife (hospitalized) and leaves in the automobile (recovered later) after taking money ($100) from the husband. The husband dies as a result of the shooting.

Solution: Effects of event: ***Injury*** ***Theft*** Damage

Severity component	*No. of victims*	*Severity weight*	*Score*
Injury: hospitalized	1	11.98	11.98
Injury: death	1	35.67	35.67
Forcible sex act	1	25.92	25.92
Intimidation: weapon	1	5.60	5.60
Vehicle: recovered	1	4.46	4.46
Property loss: $100	n.a.	3.43	3.43
		Total score	87.06

In this event, the husband was killed (35.67), and the wife was raped (25.92), threatened with a gun (5.60), and did sustain injuries requiring hospitalization (11.98). The car was stolen and recovered (4.46). The total value of the property loss was $100 (3.43). In comparison to the UCR solution of one nonnegligent criminal homicide, the Sellin–Wolfgang scale finds an injury–theft event with a total score of 87.96.

Problem 2

Two thieves break into a warehouse (damage $20) and have loaded considerable merchandise (worth $3500) on a truck. The night watchman is knocked unconscious with some blunt instrument (treated and discharged). The thieves drive away in the stolen truck (not recovered).

Solution: Effects of event: ***Injury*** ***Theft*** ***Damage***

Severity component	*No. of victims*	*Severity weight*	*Score*
Injury: treatment	1	8.53	8.53
Premises entered	1	1.50	1.50
Vehicle: not recovered	1	4.46	4.46
Property loss: $3520	1	8.91	9.91
		Total score	23.40

This offense involves the forcible entry of a building (1.50), injury to the night watchman requiring treatment (8.53), theft of an unrecovered motor vehicle (4.46), and property loss of $3520 (8.91). The UCR would classify this event as a one robbery, whereas the Sellin–Wolfgang system reveals that it is a complex event that involves the combination of the three primary effects of crime (injury, theft, and damage) and has a total seriousness score of 23.40.

Problem 3

Three men break into a public garage (damage $20) after closing hours. They steal cash from the garage office ($50) and two automobiles from the lot. One vehicle was recovered undamaged; the other was not found.

Solution: Effects of event: Injury ***Theft*** ***Damage***

Severity component	*No. of victims*	*Severity weight*	*Score*
Premises entered	1	1.50	1.50
Vehicle: recovered	1	4.46	4.46
Vehicle: not recovered	1	8.07	8.07
Property loss: $70	1	3.12	3.12
		Total score	17.15

The UCR solution to this problem would be to record one burglary. The Sellin–Wolfgang system classifies the event as a theft–damage crime that involved forcible entry (1.5), two motor vehicles stolen with one recovered

(4.46) and the other not found (8.07), and property loss totaling $70 (3.12). The total score for the event is 17.15.

Problem 4

An automobile containing clothing and luggage valued at $375 is stolen. The car is recovered (undamaged), but the clothing and luggage are missing.

Solution: Effects of event: Injury ***Theft*** Damage

Severity component	*No. of victims*	*Severity weight*	*Score*
Vehicle: recovered	1	4.46	4.46
Property loss: $375	1	4.89	4.89
		Total score	9.35

In this example, the two scoring systems are similar because the UCR would record one auto theft, whereas the Sellin–Wolfgang classification would record a vehicle theft and property theft. However, the Sellin–Wolfgang scale further signifies that the vehicle was recovered (4.46) and that there was a loss of property in the amount of $375 (4.89), which results in a final score of 9.35.

Problem 5

Answering a broadcast of an armed robbery in progress, police become engaged in a gun battle with three armed robbers; one of the bandits is killed and the other two are captured. (Presumably no one was injured except the offender, who was killed.)

Solution: Effects of event: ***Injury*** Theft Damage

Severity Component	*No. of victims*	*Severity weight*	*Score*
Injury: death	1	35.67	35.67
		Total score	35.67

If no one was injured except the offenders, this would be a theft event if theft had actually occurred before the police arrived. If so, the event would be scored for intimidation by weapon (5.60) plus the score for the value of property taken—for instance, $100 (3.43)—which totals 9.03 for the event. If the robbers had not carried out the offense because the police came before any property was taken, the event would be rated as an attempted robbery and not scored at all within the index of crime severity. In the final analysis, this event would be scored for a felony murder (35.67) to account for the death of one of the robbers. The UCR would score this event as one robbery.

Problem 6

Answering a riot call, police find that seven persons were in a fight. A variety of weapons are strewn about. None of the participants is particularly cooperative. Each one claims innocence but is vague regarding who is responsible for the assault. Three of the seven are severely wounded (all were hospitalized) while the remaining four receive only minor cuts and bruises (no medical treatment).

Solution: Effects of event: ***Injury*** Theft Damage

Severity Component	*No. of victims*	*Severity weight*	*Score*
Injury: minor harm	4	1.47	5.88
Injury: hospitalized	3	11.98	35.94
		Total score	41.82

The UCR procedure for the enumeration of this event calls for the designation of three aggravated assaults. The Sellin–Wolfgang scoring process accounts for these same effects (35.94) as well as the four minor injuries (5.88). Taken together, these consequences produce a combined score of 41.82 for this injury event.

Problem 7

Ten persons are present in a nightclub when it and the 10 persons are held up by armed bandits. Two of the victims resist the robbery and are seriously injured (hospitalization). (The combined property loss is $1800.)

Solution: Effects of event: ***Injury*** ***Theft*** Damage

Severity component	*No. of victims*	*Severity weight*	*Score*
Injury: hospitalized	2	11.98	23.96
Intimidation	n.a.	5.60	5.60
Property loss: $1800	n.a.	7.44	7.44
		Total score	37.00

The UCR classification of the event as one robbery clearly hides several important ingredients. The Sellin–Wolfgang scale produces a combined injury–theft event that involved two hospitalized victims (23.96), intimidation by a dangerous weapon (5.60), and dollar loss of $1800 (7.44). The overall score of 37.00 indicates that the recording of one robbery could be very misleading.

Problem 8

Six rooms in a hotel are broken into (damage $60) by two sneak thieves on one occasion. (The total value of property stolen from the rooms, occupied by different tenants, amounted to $1200).

Solution: Effects of event: Injury ***Theft*** ***Damage***

Severity component	*No. of victims*	*Severity weight*	*Score*
Premises entered	6	1.50	9.00
Property loss: $1260	n.a.	6.76	6.76
		Total score	15.76

The UCR classification system would designate this event as one burglary and would not account for either the property damage or the theft losses. Alternatively, the Sellin–Wolfgang system provides for six victims of burglary and incorporates the significant dollar loss from damage and theft.

These examples of the Sellin–Wolfgang severity scale show that it yields a more accurate measure of the severity of a criminal event than other methods currently in use. Although other systems measure the quantity and quality of crime, they do not produce the same degree of precision available with the Sellin–Wolfgang scale. In particular, the UCR system of counting index crimes determines the degree of seriousness of a crime by selecting the single element in the offense that has the legal label that is highest in the rank order of offenses. Furthermore, it treats all aggravated assaults, robberies, and burglaries as equally serious and each offense contributes one unit to crime rate measures.

The method for dealing with the relative gravity of criminal offenses discussed and illustrated previously has the same ultimate aim as the UCR scheme but pursues it in a different manner. Instead of focusing on an ordered set of crimes, the Sellin–Wolfgang scoring system utilizes a scale that assigns different weights to certain designated elements of an index event. When these score values are added together, they provide a score for the total event—a score that can be placed on a severity continuum reflecting the quantity and quality of criminal behavior.

Applying the Crime Severity Scale

It should be stressed that the seriousness scoring system described previously has great potential for improving the measurement of crime. It seems that this benefit applies to researchers and criminal justice practitioners alike. However, there appears to be some question whether in practice the acknowledged value of the scale warrants the extra effort required by the scoring system, an effort not necessary with the simple enumeration system of the UCR, for instance. That is, there are critics of the Sellin–Wolfgang scale who have concluded that the UCR classification and counting methods may be more than adequate for representing the volume and seriousness of criminal behavior and the additional costs and difficulties surrounding the implementation of the gravity scale overshadow the potential benefits. It needs to be recognized that these critics are essentially referring to the process of applying the scale to aggregate-level data by merely weighting the frequency of index offenses recorded in the UCR by the mean Sellin–Wolfgang severity score of similar offenses.

This is a somewhat artificial critique because the Sellin–Wolfgang scale was not created to accommodate post hoc calculations. The Sellin–Wolfgang scale begins with individual criminal events. Through the scoring procedure outlined previously, the criminal event is evaluated for the presence of several important severity components and seriousness weights are assigned. Although the system can and should be used to construct aggregate rates of crime weighted for severity, criminal events must be scored at the individual level before any aggregate weighted rates are calculated. Simply, the process of merely multiplying the frequency of a crime by an average severity score compounds the measurement problems associated with the classification of the event in accord with UCR rules in the first place. Any such procedure ignores the wealth of data represented by the criminal event (especially because of the hierarchy principle) and thus vitiates the potential of the scale to capture the quantitative components of offense severity.

Research Applications

Crime Severity Rates

One of the most frequent research issues that confronts criminology is the construction and analysis of crime rates. Crime rates form the basis of analyses designed to investigate changes in crime over time or variation across certain levels of aggregation (e.g., national, regional, state, county, and city). Usually, researchers use the data available in the UCR for measuring total, violent, and property index offenses. It was noted previously that the UCR system gives equal weight to each of the offenses in the crime index to represent the total amount of serious crime and, when subdivided into violence and property, reflects the amount of category-specific crime. The essential problem with the rates of crime derived from the UCR is that the impact of the more serious and less frequent offenses (e.g., homicide, rape, and robbery) is attenuated by the more frequent and less serious offenses. For

Table IV Aggregate Crime Severity Measures

Measure	*Meaning*
1. $\frac{\sum \text{Seriousness score across crimes}}{\text{Total juvenile population}} \times 100{,}000$	Juvenile harm: Severity crime rate per 100,000 juveniles
2. $\frac{\sum \text{Seriousness score across crimes}}{\text{Total adult population}} \times 100{,}000$	Adult harm: Weighted crime rate per 100,000 adults
3. $\frac{\sum \text{Seriousness score across crimes}}{\text{Total population}} \times 100{,}000$	Community harm: Weighted rate per 100,000 adults
4. $\frac{\sum \text{Seriousness score across crimes}}{\text{Total crimes}}$	Average crime severity
5. $\frac{\sum \text{Seriousness score across events}}{\text{Total offenders}}$	Average offense severity per offender
6. $\frac{\sum \text{Seriousness score across offenders}}{\text{Total offenders}}$	Average offender severity

example, because homicide comprises only approximately 2% of the violent crime rate, more than a 50% increase in homicide would be needed to affect a 1% increase in the violent crime rate. Clearly, this aspect of the rate structure of UCR crimes seriously jeopardizes the value of such rates for research purposes, particularly with respect to measuring significant shifts in the severity of crime over time or differences across comparison areas.

Alternatively, crime rates could be calculated that reflect the relative severity of each individual crime; consequently, such rates would better capture changes in the actual quantity of social harm associated with criminal behavior. In their pioneering work, Sellin and Wolfgang suggested several possible indices or rates that could be based on crimes weighted for severity. These rates are shown in Table IV. Although these rates were designed primarily for application to juvenile delinquency, they have direct application to adult data as well. These weighted rates have been utilized in several studies.

Formula 1 provides the main comparative statistic for a weighted index of delinquency or "juvenile harm" as it was called by Sellin and Wolfgang. The index uses the familiar offense rate calculation: (i) The severity of all acts of delinquency are scored and then summed, (ii) the result is divided by the juvenile population at risk for delinquency (i.e., usually age 7 to 17 in most jurisdictions), and (iii) this result is then multiplied by a constant (usually 100,000) to allow comparisons across jurisdictions. The resulting statistic addresses the crucial issue of the amount of juvenile harm, or weighted crime severity, inflicted on the community by 100,000 juveniles.

Formula 2 is the complement to formula 1 because it provides a weighted index of adult harm. Formula 3 provides a weighted rate of overall community harm by indicating the amount of crime severity per 100,000 population of community members. These rates are analogous to the UCR index crime rates but they appear to be far more valuable because the elasticity of the most serious crime components (although relatively infrequent) are built into the weighting scheme. Furthermore, because age-specific rates are computed, it is possible to attribute the relative share of social harm that pertains to the important distinction between juvenile delinquents and adult criminals.

Formulas 4–6 are also useful measures because they represent statistics pertaining to average severity scores. Formula 4 calculates the average crime severity across all index crimes. Formula 5 provides the average crime severity across crimes and apportions it across the offenders involved regardless of how many such offenders there were, and formula 6 calculates the average crime severity attributable to offenders regardless of the number of events in which they were involved.

These six formulas are a sample of the weighted rates that can be computed using the Sellin–Wolfgang scale. Naturally, it is possible to use other denominators to encompass different gender, race/ethnicity, or social class groups and analyze comparative severity rates. It is also possible to compute rates for the three main categories of Sellin–Wolfgang index crimes—injury, theft, and damage. In this way, subindices can be constructed in order to compare severity scores both across and within offense types. Regardless of which rates are utilized, the use of severity scores to weight the various components of a criminal event produces an index system that can accurately as possible measure the real or actual extent of harm associated with illegal behavior in a given area across time periods or across areas within a single time period.

Wolfgang *et al.*, Tracy *et al.*, and Nevares *et al.* have shown that weighted rates of delinquency allow researchers to uncover important relationships between sociodemographic factors and crime that are generally not discernible using only frequency measures.

Criminal Careers

It should be clear that another very useful application of the scale concerns what might be called offense-specific analysis. It has been conclusively demonstrated that criminal offenses can be evaluated and scaled for severity components, thus providing a basis for comparing the relative severity between crimes. This can be accomplished in two ways. First, a numerical score can be computed and assigned to the event overall or for various subcomponents such as injury. Second, the event can be classified into one of three major categories depending on which major severity component (i.e., injury, theft, or damage) of crime severity characterizes the event. The value of these two approaches can be best illustrated with respect to research on criminal careers.

Ordinarily, an offender's criminal career is typified by the number of offenses he or she has committed. The offenses may be grouped into various classes of severity, such as crimes against persons or property. Furthermore, a "rap sheet" may indicate a long series of crimes that stretch over several years or even decades, including robberies, thefts, burglaries, and assaults as the legal code defines such illegal behavior. However, these labels do not in and of themselves give any indication of the severity of the illegal conduct in terms of either absolute severity or whether such severity fluctuates during an offender's career or escalates and becomes more serious as the career progresses.

Alternatively, if the offense career were scored for severity using the Sellin–Wolfgang system, this would provide a valuable enhancement to the study of criminal career progression. It would be possible, for instance, to study whether at an early stage of the career the offenses increased in severity as the career progresses regardless of the particular legal labels attached to the behavior and whether such increased severity was associated with a longer criminal career thereafter. The severity scaling of the offenses may also reveal differences, otherwise undetected, among offenders who produce particular harmful effects such as through injury or theft components and the extent to which these different offender types are likely to continue their illegal conduct or desist from crime. It may also be possible through severity scoring to find differences among offenders concerning such correlates as age, gender, race/ethnicity, and social class that do not readily appear when only the frequency of illegal conduct is studied.

This strategy for evaluating or analyzing a criminal career has clear benefits for research designed to go beyond the mere description of criminal careers by studying the explanatory factors underpinning longer and more serious careers from their less serious or shorter counterparts. Clearly, the severity scoring system provides a means for comparing the occasional offender with his or her more recidivistic counterpart, a basis that does not merely count offenses or use broad legal labels but rather quantitatively rates their degrees of actual social harm. This improves the attempt to delineate patterns of criminal conduct that may be hidden by the broad legal labels that are usually referenced when rap sheets as opposed to offense reports are used as the data collection source. As a result, our understanding of the phenomenon of crime may be enhanced. Furthermore, our ability to control, if not prevent, prolonged criminal careers may also be improved.

Wolfgang *et al.*, Tracy *et al.*, and Nevares *et al.* have shown that when delinquency offenses are scored for seriousness, the patterns of delinquency and offense escalation across the career become meaningful aspects of the analysis of juvenile delinquency careers. Furthermore, using crime severity scores, Tracy and Kempf-Leonard have show that offense escalation early in a delinquent's career portends a greater than average likelihood that the offender will make the transition to adult criminality.

Survey Data Weighted for Crime Severity

Increasingly, surveys of victims, such as the NCVS administered by the Bureau of Justice Statistics, U.S. Department of Justice, have come to occupy a central place in the measurement of crime. By interviewing respondents in a national probability sample of households, researchers can generate highly valuable estimates of the incidence of crime, information that does not depend on whether the victim reported the crime to the police or the manner in which the police responded to crimes that were reported. From these surveys, criminologists have learned that a considerable amount of crime actually occurs but is not reported to the police and cannot be included in crime rate calculations. Clearly, therefore, victimization surveys are an important adjunct to police statistics. Similarly, self-report surveys are used to illicit data on the hidden dimension of crime—the crimes that people commit for which there may not have been a report or there was no arrest. These data not only address the incidence of crime but also provide information about the prevalence of criminality across sociodemographic correlates of official statistics versus hidden crime.

It is not only feasible but also highly desirable to apply the Sellin–Wolfgang severity scaling procedures to these surveys. In so doing, research could address topics other than just the incidence of unreported victimization or the prevalence of hidden crime. For example, comparisons could be made between the severity of offenses that victims report to the police and those that victims choose to let go unreported. Are there strong severity differences

that explain the reporting phenomenon? Furthermore, research could investigate whether different segments of the population experience different degrees of harm. In terms of self-report data, the application of severity scoring procedures provides the opportunity to compare the severity of offenses that are known to the police with those for which the offender was never caught. In any event, victimization and self-report data weighted for severity components would enhance the important function that these measurement approaches serve in augmenting official crime statistics.

Evaluation Research

The Sellin–Wolfgang severity scale also appears to have potential application in program evaluation research. Generally, offender recidivism, or the lack thereof, is used as the basic success or outcome measure. Another important outcome measure that should be evaluated is the severity of crime. By scaling the offenses committed by program participants, evaluators could examine the possible effects of treatment, if not in preventing recidivism altogether, then at least in the extent to which such treatment affected a reduction in the severity of the post-treatment criminal behavior committed. For example, one might investigate the relative effectiveness of intensive, moderate, and minimal probation or parole supervision models or a violent offender program. By using crime severity data, such program evaluations could be rendered more substantial and perhaps lead to more definitive conclusions concerning the effectiveness of certain treatment methods.

Practitioner Applications

Law Enforcement

The use of the crime severity scale is not limited to research applications; it can be implemented in various spheres of the criminal justice system. For example, Heller and McEwen tested the utility of the 1964 version of the Sellin–Wolfgang scale for law enforcement functions. The results indicated that the scale may be useful in several ways. First, it can be employed as the basis for work assignments for detectives: cases with higher than average severity scores could be allocated first instead of arbitrarily choosing cases for investigation. In this regard, the scale was also suggested as providing a means to estimate a severity-of-offense clearance rate that would reflect more accurately the effectiveness of police operations. Second, the scale could also be used in the allocation of patrol personnel to shifts (watches) with the higher severity scores. Last, the scale could also be applied in the determination of patrol beats so that patrols would cover the higher severity areas more effectively rather than the places where the volume of crimes was highest (but might have lower severity scores).

Prosecution

The Sellin–Wolfgang scale has value in the prosecution area. The scale has been successfully implemented in the Prosecutor's Management Information System in Washington, DC, to estimate the urgency of a case prosecution. The scale can also be used to assist in the selection of cases for special prosecution procedures as in the career criminal programs that have been adopted in many jurisdictions. These programs are designed to provide more effective handling of the career criminal. The special procedures may involve more extensive investigations before trial and uniform case processing from indictment through sentencing. Naturally, these career criminals must first be identified so that they can be designated for special procedures. The usual procedure is to count rap sheet offenses until some threshold is reached. However, some career criminal prosecution programs also attempt to account for the severity of the criminal's career, but evaluation in this regard usually consists only of the determination of whether the offenses are crimes against the person versus property or felonies versus misdemeanors.

The identification and prosecution of career criminals could be greatly enhanced by using the crime severity scale. Prosecutors would have a quantitative measure with which to compare offenders and the overall severity of their offense record. Consequently, prosecutors could more easily identify the most suitable candidates for special handling by the career criminal unit, and the selection could be justified with reference to not only the volume of crimes but also the quality of the offense career by accounting for the extent of social harm inflicted by career criminal designees.

Sentencing

Another stage of the criminal justice system for which the Sellin–Wolfgang scale would seem to have particular relevance is the sentencing stage. With respect to determining the appropriate sentence for convicted criminals, the severity of the offense would likely be one of the most relevant aspects. As early as the 17th century, legal theorists of the classical school called for a punishment system that bases the nature and extent of criminal sanctions on the degree of harm inflicted on the victim and the wider community by the offender. It can be seen that punishments graduated for offense severity would serve the goals of both retribution and deterrence while at the same time possibly reduce the disparity and capriciousness that characterize other sentencing methods. Hence, offenders convicted of equally serious crimes would receive the same penalty and the sanction should be more harsh than that applied to less serious violators.

The classical school doctrine of "let the punishment fit the crime" dominated the criminal law until the late 19th century when perspectives about rehabilitating offenders became more persuasive. The rehabilitation approach, with its reliance on indeterminate sentencing models, became the primary thrust of criminal sanctions and the dominant philosophy in corrections until the late 1970s. At this time, owing to growing concerns that treatment approaches were not effective at reducing offender recidivism, an alternative approach known as the "just or commensurate deserts" became more prevalent. Essentially, the just deserts approach is a revival and extension of the classical perspective that punishment should be commensurate with the blameworthiness as reflected in the current offense as well as the offender's prior criminal record.

Clearly, the Sellin–Wolfgang severity scale has substantial value in sentencing under the just deserts approach. The scale could be used to rank the severity of both the current offense and the offenses in the prior criminal career along a quantitative continuum and thus ensure that punishments were appropriately matched to the degree of crime severity—a scientific measure of just deserts. Such gradients could be constructed in terms of both classes of events (i.e., injury, theft, or damage) and the severity rating scale. Either way, the scale could provide a meaningful operational definition of the just deserts principle and simultaneously ensure that disparity in sentences was minimized.

Conclusion

The Sellin–Wolfgang severity scale was first developed in the early 1960s and updated in 1977. This article described its development and reviewed the scoring procedure that should be used to capture the qualitative and quantitative dimensions of criminal conduct. It has been shown that the severity scaling system has been used to enhance criminological research, particularly in the area of criminal careers. It was also demonstrated that the scale has widespread applicability to a variety of research and practitioner applications.

The Sellin–Wolfgang scale of crime severity was innovative when it was first developed in 1964, and it was improved when it was updated in 1985, but it has been approximately 20 years since the scale was last updated. During this time, criminology has witnessed significant changes in the awareness of and attention focused on a broad range of offense types that heretofore were given only passing concern. For example, there is now a much higher priority to the study of crimes against women and children as society has come to appreciate the significance of domestic violence and the nature of child sexual abuse. Among the crime issues currently receiving significant attention, one of the most controversial concerns the problem of "child sex offenders," also known as "child molesters," "violent sexual predators," and so on. The debate has been fueled by a series of highly publicized homicides throughout the United States involving children who were kidnapped, sexually assaulted, and murdered. These cases have stimulated a host of "sexual predator laws" or sex offender notification statutes such as the federal Megan's Law and even state laws such as Ashley's Law in Texas.

The 1980s also ushered in a keener appreciation for the drastic economic consequences associated with white-collar crime, such as the savings and loan scandals and the junk bond fraud cases. Notably, the so-called Whitewater federal prosecutions in the 1990s that involved a host of notable Arkansas politicians, bankers, and their associates points to the heightened public policy significance attributed to such large-scale white-collar crime. Unfortunately, the past 20 years has also seen numerous incidents of terrorist crimes, such as the sabotaging of airliners with significant loss of human lives; the bombing of the World Trade Center in New York; the bombing of the federal building in Oklahoma City, which resulted in the deaths of 168 people and injuries to countless others; and the ultimate act of terrorism—the attack on September 11, 2001, which caused the loss of thousands of lives, countless injuries, and hundreds of millions of dollars in property losses in New York, Washington, DC, and Pennsylvania.

These crime-related developments during the past 20 years surely must have had an effect on the public's perception of crime and the effectiveness of efforts to control crime and punish offenders accordingly. Thus, it is necessary, if not crucial, to update the 1977 National Survey of Crime Severity by sampling a nationally representative group of respondents. These respondents could be administered a set of questionnaires with an updated set of offense descriptions or scenarios to capture the developments reviewed previously. This research would ensure that the Sellin–Wolfgang scale of offense severity would continue to be an important measurement approach with contemporary relevance and significance to researchers and policymakers.

See Also the Following Articles

Criminal Justice Records • Criminology • Experiments, Criminology • Police Records and The Uniform Crime Reports

Further Reading

Blumstein, A. (1974). Seriousness weights in an index of crime. *Am. Sociol. Rev.* **39,** 854–864.

Figlio, R. (1975). The seriousness of offenses: An evaluation by offenders and nonoffenders. *J. Criminal Law Criminol. Police Sci.* **66,** 189–200.

Galanter, E. (1962). The direct measurement of utility and subjective probability. *Am. J. Psychol.* **75,** 208–220.

Galanter, E., and Messick, S. (1961). The relation between category and magnitude scales of loudness. *Psychol. Rev.* **38,** 363–372.

Heller, N. B., and McEwen, J. T. (1975). Applications of crime seriousness information in a police department. *J. Res. Crime Delinquency* **12,** 44–50.

Hindelang, M. S. (1974). The Uniform Crime Reports revisited. *J. Criminal Justice* **2,** 1–17.

Hsu, M. (1973). Cultural and sexual differences on the judgement of criminal offenses. *J. Criminal Law Criminol. Police Sci.* **64,** 348–353.

Jacoby, J. E. (1972). *A System for the Manual Evaluation of Case Processing in the Prosecutor's Office: First Annual Report*. National Center for Prosecution Management, Washington, D.C.

Nevares, D., Wolfgang, M. E., and Tracy, P. E. (1990). *Delinquency in Puerto Rico: The 1970 Birth Cohort Study.* Greenwood, Westport, CT.

Normandeau, A. (1970). De la Criminalité dans 8 pays. *J. Criminal Law Criminol. Police Sci.* **57,** 172–177.

Rossi, P. H., Waite, E., Boise, C. E., and Berk, R. E. (1974). The seriousness of crimes. *Am. Sociol. Rev.* **39,** 224–237.

Sellin, T., and Wolfgang, M. E. (1964). *The Measurement of Delinquency.* Wiley, New York.

Tracy, P. E., and Kempf-Leonard, K. (1996). *Continuity and Discontinuity in Criminal Careers: The Transition from Delinquency to Crime.* Plenum, New York.

Tracy, P. E., Wolfgang, M. E., and Figlio, R. M. (1990). *Delinquency Careers in Two Birth Cohorts.* Plenum, New York.

Velez-Diaz, A., and Megargee, E. L. (1971). An investigation of differences in value judgements between youthful offenders and nonoffenders in Puerto Rico. *J. Criminal Law Criminol. Police Sci.* **61,** 549–553.

von Hirsch, A. (1976). *Doing Justice: The Choice of Punishments.* Hill & Wang, New York.

Welford, C., and Wiatrowski, M. (1975). On the measurement of delinquency. *J. Criminal Law Criminol. Police Sci.* **66,** 175–188.

Wolfgang, M. E. (1963). Uniform Crime Reports: A critical appraisal. *Univ. Pennsylvania Law Rev.* **111,** 708–738.

Wolfgang, M. E., Figlio, R. M., and Sellin, T. (1972). *Delinquency in a Birth Cohort.* University of Chicago Press, Chicago.

Wolfgang, M. E., Figlio, R. M., Tracy, P. E., and Singer, S. I. (1985). *The National Survey of Crime Severity.* U.S. Government Printing Office, Washington, DC.

Shewhart, Walter

Mark Wilcox
Cranfield University, Cranfield, Bedford, United Kingdom

Glossary

common causes The description given to variation in a stable process. This can also be described as natural variation in a system—in other words, nothing unusual is happening (e.g., no special causes). Common cause variation is also present in processes that are out of control and are constant over time and throughout the system.

control Here we see the relationship between prediction and control: "For our present purpose *a phenomenon will be said to be controlled when, through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. Here it is understood that prediction within limits means that we can state, at least approximately, the probability that the observed phenomenon will fall within limits*" (Shewhart; emphasis in original).

prediction The premise is that if a process is in statistical control, then the future behavior of that process is predictable for the foreseeable future. The notion of prediction comes from theory. The construction of a chart allows the evidence to be interpreted, using theory to predict what may happen next. Data on its own does not provide a prediction. "Rational prediction requires theory and builds knowledge through systematic revision and extension of theory based on comparison of prediction with observation" (from Deming).

Shewhart control chart A chart with two axes used to plot data taken from processes and products. The plotted data points on the chart illustrate "variation" in the "thing" being measured. The chart will also show the mean and upper and lower control limits. Charts are an operational definition of the distinction between common and special causes of variation. Charts can be used for (1) judgment of the past, (2) stability of the present, (3) prediction of the future by looking back on a set of results from a process, and as an *Operation* to attain and maintain an ongoing process.

special causes Causes outside natural variation in the process. In manufacturing, it could be some faulty material, an untrained employee making a mistake, or a worn machine tool, for example. They are often sporadic and may have a local impact on the process as a whole. Process instability is shown on the control chart as evidence of special causes that the operator or engineer will have to find, remove, or correct to get the process back into control.

subgroups Sets of measurements collected at specific times or points in a process to monitor the process. Each subgroup is treated as a sample and will display a pattern of behavior that may be treated as part of a population. If the patterns of behavior are consistent then the process may be assumed to be in control.

system "A system is a network of interdependent components that work together to try to accomplish the aim of the system" (from Deming).

variation In a system, variation is based on the premise that no matter how hard we try to do something there will always be some variation in the process and the outcome. Variation is a result of common and special causes.

Dr. W. A. Shewhart (1891–1967) was the founder of the modern quality movement with his application of modern statistical method to mass-production in the 1920s. Shewhart invented the control chart in 1924 and introduced the concept of assignable and common causes of variation. His contribution has never had the recognition it deserves in the West, while his work was developed by others in the post-World War II reconstruction in Japan. Deming, Juran, Ishikawa, and many others in their wake, took Shewhart's ideas and applied them to the improvement of quality, that grew throughout the 20th century. Significantly, the notion of continual improvement can be traced back to his publications in 1931 and 1939. Deming's PDSA wheel was always referred to as the Shewhart Cycle in his writing and teaching. Deming, more than any other, continued to praise the contribution made by this remarkable man, as he developed his own ideas until his death in 1993.

Introduction

Dr. W. A. Shewhart was employed as an industrial researcher at the Hawthorne Plant of Bell Telephones in the United States in the 1920s and 30s. Shewhart was asked to find a more economic way of controlling quality. The cost of relying on inspection was becoming prohibitive for a company employing thousands of workers producing telephones and telephone systems that consumed 110,000 different piece parts. Shewhart's task was twofold. The first was to reduce the number of defectives, and the second was to find economic and reliable methods of sampling products that could only be tested by destruction, e.g., the blowing of fuse. "The attempt to solve these two problems gave rise to the introduction of the operation of statistical control involving the use of the quality control chart in 1924, and may therefore be seen as the starting point of the application of statistical technique in the control of quality of a manufactured product" (Shewhart).

By 1931, Shewhart had published his treatise in which he systematically developed the principles of his statistical method. Indeed, much of this book is about statistical method being applied to mass production. However, with more careful reading we find that Shewhart read widely and in great detail. The appendices of this volume contain a lengthy section outlining the different books and the contribution they all made to his work. These include statistics and probability, economics, physics, mathematics, logic, philosophy of science, and psychology, plus an array of other peripheral works.

Shortly after this book was published, it appears that Shewhart came across the work of pragmatist philosopher Professor C. I. Lewis, who had published his theory of knowledge, *Mind and The World Order* in 1929. Shewhart was reputed to have read this book 14 times, and was apparently indebted to its influence on the development of his ideas over the next few years. Lewis' influence becomes apparent in Shewhart's 1939 book and can be seen in Deming's publications in 1986 and 1993.

While many have applied statistical methods over the past 80 years, Wilcox suggests that to fully understand what Shewhart was trying to achieve, we must understand the influence of Lewis and others from the philosophy of science (e.g., Whitehead and Eddington quoted in Shewhart; it should be noted that Shewhart first produced a control chart in 1924 to put these points into historical perspective).

Shewhart's work is based on an ontology of flux and the control chart is an epistemological device (or operational definition) to show how processes are in flux, or showing variation. Elements of Lewis' work can be traced back to Heracleites, and the notion that the world is in motion. Lewis argued that "mind" interacts with and interprets the here and now as the world passes by in motion. Control charts are a device for illustrating this phenomenon. If this point is not understood, then the chances are the application and use of control charts will not be fully realized.

Furthermore, Shewhart's aims were to help managers and engineers predict the future behavior of the processes in their organizations. This point is also often lost in the interpretation and application of the statistical method.

This article will therefore show Shewhart's work as a methodology for interpreting an organization as a system. The system is made up of interdependent processes. Getting the processes into a state of statistical control will then achieve maximum economic benefits. If all the processes are in a state of statistical control, then "Mass production viewed in this way constitutes a continuing and self corrective method for making the most efficient use of raw and fabricated materials" (Shewhart) Shewhart's work is probably the original theory for continual improvement and this article will show how he arrived at this concept.

The Problem with Inspection

Shewhart's task as a research worker at Bell Telephone's Hawthorne Plant was to find a more economic way of controlling quality. A basic telephone in the 1920s consisted of over 200 parts. A telephone system had over 110,000 parts. The annual production involved billions of parts, which were resourced from around the globe. Clearly, the cost of inspection, at source and during production, was becoming prohibitive. Furthermore, the marketing people had designed a slogan "as alike as two telephones," which was a potential embarrassment as the reliability of the equipment was not very good. Indeed, Shewhart also observed that the harder the managers tried to improve the quality, the worse things got. So, in short, inspection was costly and not providing reliable products.

Shewhart's solution was to try to gain control over the whole process of production. If control can be attained, then the future quality of the products is predictable. Shewhart used some interesting data from The Food Research Institute of Stanford University. The researchers had studied the returns of bread to bakeries as a loss to the system. Ten bakeries were studied over 36 weeks. Some had far more returns than others. From this, the deduction was made that the bakery with the least number of returns (1.99%) also showed better control over the 36 weeks. This was demonstrated by showing the data—not in tabular form, but on a control chart. The returns of the 10 bakeries were compared. The bakery with 1.99% returns also appeared to have the most stable system. Therefore, Shewhart argued: greater control, equals fewer defects.

Inspection, does not increase the degree of control over the system of production. This is only achieved by bringing the processes and the system as a whole into a state of statistical control.

Defining Quality

The word quality has been around for thousands of years and can be found in Plato's republic, for example. It has been particularly problematic for writers on quality, because of the subjective nature of the concept. For instance, what may be a quality product or service for you, may not meet my needs, and therefore does not satisfy my requirements. Shewhart gave the definition of quality some thoughtful attention. His chapter on defining quality in his 1931 book is worthy of attention by scholars struggling with this point. Space does not allow a full exposition of all of the points, but his distinction between the objective and subjective sides of quality are worth noting. Objective quality exists "independent of man," and comprises the physical attributes that can be measured, such as shape, size, and weight.

The subjective side of quality is related to the notion of value. This is broken down into four features: (1) Use, (2) Cost, (3) Esteem, and (4) Exchange. Shewhart recognizes the economic importance of the subjective features, but realizes they are notoriously difficult to measure and control. However, he suggests that the engineer must be aware of the customer's needs and have the ability to translate them into the physical (objective) characteristics of the product. "In taking this step intuition and judgment play an important role as well as the broad knowledge of the human element involved in the wants of individuals" (Shewhart).

We see here probably one of the first pieces of management writing, which advocates satisfying customer needs, long before the theories on marketing and the later work in the 1990s on customer satisfaction. We can also see the link to Deming's mantra about the customer being the most important person on the production line. These connections show the importance of Shewhart's work and why it may be seen as an epoch in the history of management.

A Solution and An Epoch

Shewhart's aim was to gain control over the processes of production. He did this by developing his statistical method for mass production. The statistical method was originally developed in the natural sciences, not in engineering or mass production. Shewhart's thesis could be interpreted as a polemic against what he called the "exact sciences." Statistical theories were based on the so-called "laws of nature" that exhibit variation and probabilities: "It follows, therefore, since we are thus willing to accept as axiomatic that we cannot do what we want to do and cannot hope to understand why we cannot, that we must also accept as axiomatic that a controlled quality will not be a constant quality. Instead, a controlled quality must be a variable quality. This is the first characteristic" (Shewhart). So instead of trying to make products that are exactly alike, we must accept that there will always be some variability. The variability is the result of a constant cause system, the roots of which we can never know. Three postulates were developed to underpin his theory of control:

1. *All chance systems are not alike in the sense that they enable us to predict the future in terms of the past*. So, for example, the economic laws that control inflation are very different from the laws governing the tossing of a coin. We can predict that the tossing of coin in 50 years time will provide similar results, because the cause systems are simple to understand. Conversely, we cannot predict the rate of inflation in 50 years time because there are far too many variables in an economic system.
2. *Constant systems of chance causes do exist in nature*. This is probably the most contentious of Shewhart's claims, a point he rightly acknowledges. "To say that such systems exist in nature, however, is one thing; to say that such systems of causes exist in a production process is another" (Shewhart). Shewhart backs up this claim from his own research which he suggests shows chance cause systems at work. What this suggests is that we can measure mass-production systems following the rules of statistical theory, and the data will show variation similar to that found in the natural world.
3. *Assignable causes of variation may be found and eliminated*. Assignable causes of variation are a cause of defects in the system and a greater source of defects. Shewhart's theory claims to be able to identify assignable causes of variation, which are outside the system of chance cause variation. Once identified using Shewhart's method, assignable causes may be eliminated. Shewhart appreciated making the distinction between chance and assignable causes, and in effect drew a line between those he could economically remove and those that should be left to chance. This is how he arrived at the 3 sigma limits: "We assumed, therefore, upon the basis of the this test, that it was not feasible for research to go much further in eliminating causes of variability." Thus, he established the 3 sigma limits for process control, not by statistical theory alone, but ultimately by judgment.

By such a procedure, processes can be brought into a state of statistical control. When systems are in control, they are predictable—all things being equal. If they are predictable, then the need for inspection is reduced. The result is a reduction in the cost of inspection and rejection, and a more uniform quality of output.

These three postulates form the basis of Shewhart's ideas. A controlled or constant system of chance causes is the ultimate aim—with the qualification that one is producing what the customer wants!

Measuring

Shewhart had to devise a reliable method for measuring the objective features of quality as the starting point in the construction of a control chart. He took into account the precise and accurate methods of measurement from Goodwin and the theory of errors, to account for variation in the measuring process. However, if there is variation (in everything), then the process of measurement becomes problematic.

Measurement(s) provide the data points that will eventually appear on the control charts. We can see how Shewhart perceived the engineer, reacting to and upon, an ever-changing environment.

"The operation of control is in this sense a *dynamic process* involving a chain of actions, whereas the criterion of control is simply a tool used in this process. The successful quality engineer, like the successful research worker, is *not a pure reason machine* but instead is a *biological unit* reacting to and acting upon an ever changing environment" (Shewhart; emphasis added).

According to Lewis, all knowledge of objects comes from our conceptual interpretation of the given. There is no knowledge without interpretation—and the first step in interpreting data stems from the observer, who takes the measurement. No measurement—and therefore no knowledge of the quality levels—can take place without the presence and perception of the observer. The act of measurement is a process, and the values are sequences of numbers plotted on a control chart.

The control chart was a unique invention by Shewhart in 1924. The chart represents the variation in the observations over time, including the measurement variation. The individual measures, capture a moment in time, which when seen collectively, manifest variation. The central line is the mean or median. The two outer lines are the upper and lower control limits calculated from the individual measures to represent 3 sigma limits. The control chart becomes the voice of process (Burr quoted in Deming)—in effect it tells a story. Collectively, the control charts have the ability to become the voice of the organization.

Prediction

Shewhart was trying to construct a theory of control and prediction premised on the notion that everything is in flux: "A phenomenon will be said to be in control when, through the use of past experience we can predict at least within limits how a phenomenon may be expected to vary in the future. Here it is understood that the prediction within limits means that we can state, at least approximately, the probability that the observed phenomenon will fall within given limits."

The data points on a control chart represent the evidence and the first step in the process of prediction. Prediction depends upon the relationships among (1) the data provided as evidence, (2) the prediction made on the basis of this evidence, and (3) the degree of belief in the prediction which is related to data described in (1). Therefore, the validity of the prediction depends upon the integrity of the (data) evidence collected and the stability of the process.Similarly, Shewhart refers to:

> *Nonstatic character of knowledge . . . we are forced to consider knowledge as something that changes as new evidence is approved by more data, or as soon as new predictions are made from the same data by new theories. Knowledge in this sense is somewhat of a continuing process, or method, and differs fundamentally in this respect from what it would be if it were possible to attain certainty in the making of predictions.*

This indicates how Shewhart thought that everything was in flux, or movement, showing the connection of past, present, and future. He was clearly influenced by a quotation from Lewis (1934)" . . . knowing begins and ends in experience; but it does not end in the experience in which it begins" (Quoted in Shewhart).

Shewhart linked the past, present, and future together to illustrate the notion of flux or flow. Then he took the three concepts of specification, production, and inspection from the exact methods of mass production and turned them into a circular spiral, into what became the origins of continual improvement.

"*The three steps constitute a dynamic scientific process of acquiring knowledge*" (Shewhart; emphasis in original). To show how this works, Shewhart explains how scientists and statisticians join forces. Scientists decide on the specification (step 1), and join forces with the statisticians (step 2) to eliminate assignable causes of variation to a point where predictions can be made. Statisticians need the scientists to eliminate the causes, because of their knowledge of the process (physical laws). When the state of statistical control has been attained the statistician can continue without the scientist, (step 3) and "*set up rules that lead to the most efficient prediction*" (Shewhart).

Shewhart described how this would work in practice:

> *In fact an economic standard of quality is not a written finality, but is a dynamic process. It is not merely the imprisonment of the past in the form of specification (step I), but rather the unfolding of the future as revealed in the process of production (step II) and inspection*

(step III), and made available in the running quality report (Shewhart; emphasis in original).

Thus we see the roots of continual improvement, and the basis of what Deming developed into the Shewhart cycle or PDSA Deming Wheel. Shewhart's discourse encapsulates the notion of flux in the passage above, and made the distinction between scientific and emotive language. He was well aware of the problems of communicating his theories to a somewhat skeptical engineering audience who were steeped in the language of the "exact" sciences. It was a long time before his work received the attention it deserved, and even in the 1980s Deming was predicting it would be another 50 years before the full impact of Shewhart's work would be realized.

See Also the Following Article

Deming, William Edwards

Further Reading

Blankenship, B., and Petersen, P. B. (1999). W. Edwards Deming's mentor and others who made an impact on his views during the 1920s and 1930s. *J. Management Hist.* **5,** 454–467.

Deming, W. E. (1986). *Out of the Crisis.* MIT Press, Cambridge, MA.

Deming, W. E. (1991). A tribute to Walter A. Shewhart on the occasion of the 100th anniversary of his birth. *SPC INK Newsletter* Winter.

Deming, W. E. (1994). (ition) *The New Economics,* 2nd Ed. MIT Press, Cambridge, MA.

Goodwin, H. M. (1908/1920). *Elements of the Precision of Measurements and Graphical Methods.* McGraw-Hill, New York.

Lewis, C. I. (1929/1956). *Mind and The World Order: Outline of a Theory of Knowledge*. Dover, New York.

Mauléon, C., and Bergman, B. (2002). On the theory of knowledge in the quality movement—C.I. Lewis' contribution to quality pioneers. In *Proceedings of the 8th Annual Research Seminar*, Fordham University, New York, February, pp. 156–164.

Plato (1934). *Plato's Theory of Knowledge: The Theaetetus and the Sophist Translated with a Running Commentary* (F. M. Cornford, ed.). Routledge & Keegan Paul, London.

Shewhart, W. A. (1931/1980). *Economic Control of Quality of Manufactured Product.* ASQC, Milwaukee.

Shewhart, W. A., and Deming, W. E. (1939/1986). *Statistical Method from The Viewpoint of Quality Control.* Dover, New York.

Whitehead, A. N. (1929/1960). *Process and Reality.* MacMillan, New York.

Wilcox, M. (2002). Whither the pragmatism? In *Proceedings of the 8th Annual Research Seminar* (pp. 258–269). Fordham University, New York.

Wilcox, M. (2004). Prediction and pragmatism in Shewhart's theory of statistical control management decision. *Focus on Management History*, **42,** 152–165.

Small Area Estimation

Ferry Butar Butar
Sam Houston State University, Huntsville, Texas, USA

Glossary

composite estimator The weighted average of a direct and a synthetic estimator.
direct estimator The estimator obtained from values of the variable of interest only from units in the area or domain of interest.
empirical Bayes estimation A Bayes estimate of the unknown parameter of interest that is obtained by using a prior distribution. The unknown parameters of the prior distribution in the estimator are then estimated by some classical methods.
empirical best linear unbiased prediction (EBLUP) A mixed linear model is used to produce the best linear unbiased predictor (BLUP). The variance components involved in the BLUP are estimated by a standard method (i.e., ANOVA). When the variance components of a BLUP are replaced by their estimates then it is called an EBLUP.
hierarchical Bayes estimation This method models prior parameters in stages. The parameter of interest is estimated by its posterior mean, and its posterior variance is a measure of precision.
indirect estimator The estimator obtained from values of the variable of interest not only from units in the area or domain of interest but also from other areas or domains of interest.
small area Small area (domain) generally refers to a subgroup of a population from which samples are drawn. The subgroup may refer to a small geographical region (e.g., state, county, and municipality) or a particular group obtained by a cross-classification of various demographic factors such as age, gender, and race.
synthetic estimator The estimator from the larger area that is used to estimate the smaller area with the assumption that the characteristics of the larger area are similar to those of the smaller areas.

Small area statistics are needed in regional planning and fund allocations. The direct survey method is an unreliable estimate for a subnational region due to small samples available from the region. Demographers have long used a variety of indirect methods for estimating small area populations in postcensal years. In estimating small areas, it is often necessary to borrow strength by using values from related areas to increase the effective sample size. The model-based approach to small area estimation, empirical best linear unbiased prediction, empirical Bayes, and hierarchical Bayes methods are based on explicit small area models. The model-based approach is very effective and offers several advantages.

Introduction

The history of small area statistics can be traced back to the 11th century in England. Records of births, baptisms, marriages, deaths, etc. were used to produce various small area statistics. At that time, sources of small area statistics were limited to various administrative records available from local governments.

The sampling design and the sample size of most large-scale national surveys are usually determined so as to produce reliable estimates of various characteristics of interest at the national level. Often, there is a need to produce similar estimates at the subnational levels (e.g., states and counties). The direct survey method is an unreliable estimate for a subnational region due to small samples available from the region. A similar situation occurs when estimates are needed for domains obtained by classifying the population according to various demographic characteristics (e.g., age, race, and sex). For example, Fay and Herriot considered the estimation of per capita incomes of small places (population less than 1000). In order to obtain their estimates, they combined information from two sources. The direct information came from the Current Population Survey, and the

second source of data derived from tax returns for the year 1969 and housing information from the 1970 census. They provided the empirical Bayes [which is the same as best linear unbiased prediction (BLUP)] estimator of per capita income, which is a weighted average of the Current Population Survey estimator of the per capita income and a regression estimator that utilizes tax return data and the data on housing. Another example is the drug prevalence estimation given by Meza *et al*. Since drug use is a relatively rare event, many counties in Nebraska have few or zero cases; therefore, the estimate based on individual counties is unreliable due to the small sample size. Such problems in survey sampling literature are known as small area estimation problems. Small area (domain) generally refers to a subgroup of a population from which samples are drawn. The subgroup may refer to a small geographical region (e.g., state, county, and municipality) or a particular group obtained by a cross-classification of various demographic factors, such as age, gender, and race. Other terms used to describe small area estimation include local area, small domain, subdomain, small group, subprovincial, indirect, and model dependent.

Reliable small area statistics are needed in regional planning and in allocations of government resources. Due to budgetary constraints, it is not possible to collect adequate sample sizes from the small areas. When information on one or more relevant covariates is available from various administrative records, synthetic estimators (i.e., regression estimators) have been proposed in the small area literature. Although the synthetic estimators have small variances compared to the direct survey estimators, they tend to be biased because they do not use the information on the characteristic of interest directly obtainable from the sample surveys. A compromise between the direct survey and the synthetic estimation is the method of the composite estimation.

Morrison described small area estimation methods used prior to 1970. Purcell and Kish reviewed demographic methods as well as statistical methods of estimation for small domains. In 1980, the National Research Council gave detailed information as well as an evaluation of the Census Bureau's procedure for making postcensal estimates of the population and per capita income for local areas. Schaible provided estimates on small area used in U.S. federal programs.

Demographic Methods of Small Area Estimation

Demographers use a variety of methods for estimating small area population in postcensal years. These methods, called symptomatic accounting techniques (SATs), utilize current "symptomatic" (e.g., number of births and deaths) data from administrative registers as well as related data from the latest census. The SAT methods consist of the vital rates, Census Component Methods II, the administrative records method, the housing unit method, and regression symptomatic method. Except for the regression symptomatic method, these methods do not use sampling.

The components method uses birth, death, and migration to estimate population. The net migration is the sum of the immigration and net interarea migration minus emigration, or $m_{0t} = i_{0t} + n_{0t} - e_{0t}$, where i_{0t} is the immigration during the time period between 0 and t, n_{0t} is interarea migration, and e_{0t} is the emigration for the period between time 0 and t. Then the population for small area is estimated by

$$P_t = P_0 + b_{0t} - d_{0t} + m_{0t}, \tag{1}$$

where P_0 is the population of the small area in the census year 0, and b_{0t} is the number of births and d_{0t} is the number of deaths in the small area during period 0 and t. Unlike net migration, registration of births and deaths is usually complete in the United States and Canada. In practice, net migration is often difficult to estimate. In the United States, military migration is obtained from administrative records, whereas civilian migration is obtained from school enrollment records (Component Method II) and from income tax returns (administrative records).

Direct and Synthetic Estimation

Direct Estimation

Direct survey estimation is the most understood and widely used technique. These estimators are motivated from the randomization principle of survey sampling and typically use information only from the small area and the time period of interest. The direct estimators are unbiased estimators. A simple example of direct estimation is the sample mean, which is an unbiased estimator. The slope estimated from regression using only data from small area is also unbiased. The direct estimate is an unstable estimate for small area due to the small sample size from the region; therefore, the variance is very large. Certain considerations at the design stage (e.g., less clustering and more stratification, and sample allocation that provides more samples for the small areas) and estimation stage (e.g., calibration) can improve direct estimation.

Synthetic Estimation

Suppose the characteristics of the small areas are the same as those of a large area in which the small area is located. For the large area, an unbiased direct estimator is

obtained from a sample survey. When this unbiased direct estimate is used to derive an indirect estimator for the small area, then this estimate is called a synthetic estimate.

Suppose the auxiliary variables are not available. Let $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, where $i = 1, \ldots, m$, small area means are the parameter of interest, and the sizes, N_i, of small areas are known ($i = 1, \ldots, m$). The usual survey estimate of $\bar{Y}_i$ is $\bar{y}_{is} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, $i = 1, \ldots, m$. The variance of $\bar{y}_{is}$ for given n_i is of order n_i^{-1}; hence, the estimator, $\bar{y}_{is}$, is likely to yield large standard errors unless n_i is large. A simple synthetic estimator of $\bar{Y}_i$ is the weighted mean direct estimates,

$$\bar{y}_s = \frac{\sum_{i=1}^{m} n_i \bar{y}_{is}}{n}, \tag{2}$$

where $n = \sum_{i=1}^{m} n_i$. The mean square error of $\bar{y}_s$ is of order n^{-1}. If the mean of small area is approximately equal to the overall mean (i.e., $\bar{Y}_i = \bar{Y}$), then the mean square error (MSE) will be small and the estimator will be very efficient. If this assumption is violated, then the synthetic error has a greater concern for small areas rather than large. The estimation could be an underestimation in some areas and an overestimation in others. As a result, there is not a qualified improvement in smallest areas. The synthetic estimation is a bias estimator due to applying the same correction to areas with different coverage.

Now suppose the auxiliary information X, with known population mean $\bar{X}_i$ is available. The synthetic ratio estimator of $\bar{Y}_i$ is

$$\bar{y}_{is}^{r} = \frac{\bar{y}_s}{\bar{x}_s} \bar{X}_i. \tag{3}$$

In the usual direct estimate, one uses the ratio $\bar{y}_{is}/\bar{x}_{is}$ instead of $\bar{y}_s/\bar{x}_s$. Again, the assumption that $\bar{y}_s/\bar{x}_s = \bar{y}_{is}/\bar{x}_{is}$ is very strong and will usually not be true in practice. The rationale for the synthetic estimator is that the distribution of a characteristic of interest is highly related to the demographic composition of population. If the assumption is not valid, the synthetic estimators for some of the small areas can be highly design biased.

The variance of a synthetic estimator will be small compared to the variance of a direct estimator because it depends only on the precision of the direct estimator in a large area. To find the $\operatorname{mse}(\bar{Y}_{is})$, first find the $\operatorname{mse}(\hat{Y}_{is})$, since $\bar{Y}_{is} = \hat{Y}_{is}/N_i$, then $\operatorname{mse}(\bar{Y}_{is}) = \operatorname{mse}(\hat{Y}_{is})/N_i^2$. An approximate design unbiased estimator of MSE of $\hat{y}_{is}$ will be obtained using the unbiased direct estimator of $\hat{Y}_i$. The mean square error is

$$\operatorname{MSE}(\hat{Y}_{is}) = E(\hat{Y}_{is} - \hat{Y}_i)^2 - E(\hat{Y}_i - Y_i)^2 + 2\operatorname{Cov}(\hat{Y}_{is}, \hat{Y}_i). \tag{4}$$

If $\operatorname{Cov}(\hat{Y}_{is}, \hat{Y}_i)$ is approximately zero, where $\hat{Y}_i$ is a direct unbiased estimator of Y_i, then an approximate unbiased estimator of MSE of $\hat{Y}_{is}$ is given by

$$\operatorname{mse}(\hat{Y}_{is}) = (\hat{Y}_{is} - \hat{Y}_i)^2 - v(\hat{Y}_i), \tag{5}$$

where $v(\hat{Y}_i)$ is a design unbiased estimator of the variance of $\hat{Y}_i$. In practice, the assumption of $\operatorname{Cov}(\hat{Y}_{is}, \hat{Y}_i) \approx 0$ is realistic because $\hat{Y}_{is}$ is less variable than $\hat{Y}_i$.

The synthetic estimate also can be found by regression. Suppose X_i is a vector of covariates which are related to $\bar{Y}_i (i = 1, \ldots, m)$. Then $X_i' \hat{\beta}$, where $\hat{\beta}$ is an estimator of the vector of regression coefficients obtained by fitting $\bar{Y}_i'$s on the X_i's is called a synthetic estimator of $\tilde{\bar{Y}}_i (i = 1, \ldots, m)$. In case $X_i = 1$ (i.e., when there is no relevant covariate available), the synthetic estimator of $\tilde{\bar{Y}}_i$ reduces to $\bar{Y}$, the overall mean.

Composite Estimation

Composite estimates are indirect estimates that borrow strengths from other areas of interest. Although the synthetic estimators have small variances compared to the direct survey estimators, they tend to be biased because they do not use the information on the characteristic of interest directly obtainable from the sample surveys. A compromise between the instability of a direct estimator, $\hat{Y}_{i1} = \hat{Y}_i$, and the potential bias of a synthetic estimation, $\hat{Y}_{i2} = \hat{Y}_{is}$, is the method of the composite estimation. Broadly defined, a composite estimator is the weighted average of a direct survey estimator and a synthetic estimator. The synthetic and the composite estimators are usually obtained by implicit or explicit models that borrow strengths from related sources. The general form of a composite estimator of θ_i is

$$\hat{Y}_{iC} = W_i \hat{Y}_{i1} + (1 - W_i) \hat{Y}_{i2}, \tag{6}$$

where $0 \le W_i \le 1$ is determined from the data by some optimal method. Composite estimators generally perform better than both the survey estimators and the synthetic estimators in the average MSE sense. The weight W_i is determined by minimizing the classical MSE of $\hat{Y}_{iC}$ with respect to W_i. The design MSE of the composite estimator is

$$\begin{aligned} \operatorname{MSE}(\hat{Y}_{iC}) = {} & W_i^2 \operatorname{MSE}(\hat{Y}_i) + (1 - W_i)^2 \operatorname{MSE}(\hat{Y}_{is}) \\ & + 2Wi(1 - W_i) E(\hat{Y}_i - Y_i)(\hat{Y}_{is} - Y_i). \end{aligned} \tag{7}$$

By minimizing Eq. (7) with respect to W_i, the optimal weight W_i is

$$W_i(\text{opt}) \approx \frac{\operatorname{MSE}(\hat{Y}_{is})}{\operatorname{MSE}(\hat{Y}_i) + \operatorname{MSE}(\hat{Y}_{is})}, \tag{8}$$

assuming that the covariance term $\operatorname{Cov}(\hat{Y}_i, \hat{Y}_{is}) \approx 0$. These weights can be very unstable. Purcell and Kish

use common weight W and then minimize the average MSE with respect to W, and the weight is given by

$$W(\text{opt}) = 1 - \frac{\sum_{i=1}^{m} v(\hat{Y}_i)}{\sum_{i=1}^{m} (\hat{Y}_{is} - \hat{Y}_i)^2}. \quad (9)$$

If the variances of the $\hat{Y}_i$ are equal, then Eq. (9) becomes

$$W(\text{opt}) = 1 - \frac{m\bar{v}}{\sum_{i=1}^{m} (\hat{Y}_{is} - \hat{Y}_i)^2}, \quad (10)$$

where $\bar{v} = m^{-1} \sum_{i=1}^{m} v(\hat{Y}_i)$.

Empirical Best Prediction

Small area models can be classified into two types. The focus here is on model-based indirect estimators that combine data from administrative and census data with the sample survey data. The first type, area specific auxiliary information, x_i, is available for areas $i = 1, \ldots, m$. The population area mean is assumed to be related to x_i through a linear model

$$\theta_i = x_i' \beta + v_i, \quad i = 1, \ldots, m, \quad (11)$$

where β is a vector of the regression parameters, and v_i's are the random small area effects to be independent identically distributed (i.i.d) normal with a mean of 0 and variance of τ^2. Assume that the direct survey estimator Y_i of θ_i is known as

$$Y_i = \theta_i + e_i, \quad i = 1, \ldots, m, \quad (12)$$

where e_i's are the sampling errors with i.i.d normal with mean of 0 and known variance σ_i^2. Thus, combining Eqs. (11) and (12), the basic area level model is

$$Y_i = x_i' \beta + v_i + e_i, \quad i = 1, \ldots, m, \quad (13)$$

which is a special case of the generalized mixed linear model. This model involves design-based random variables, e_i, and model-based random variables, v_i. Assume that v_i and e_i are uncorrelated. In practice, the sampling variance of σ_i^2 may not be known, but it can be smoothed by some methods to stabilize the estimator $\hat{\sigma}_i^2$.

The second type of model, unit value y values, y_{ij}, are assumed to be related to auxiliary information $x_{ij} = (x_{ij1}, \ldots, x_{ijp})'$. A nested error linear regression model is

$$y_{ij} = x_{ij}' \beta + v_i + e_{ij}, \quad i = 1, \ldots, m; \; j = 1, \ldots, N_i, \quad (14)$$

where v_i's are i.i.d. normal random variables with mean 0 and variance of τ^2 and are independent of element errors $e_{ij} \overset{iid}{\sim} N(0, \sigma_i^2)$, and N_i is the number of population elements in the ith area.

For example, consider the Fay and Herriot model [i.e., Model (13)]. Model (13) is a special case of a generalized linear mixed model. We are interested in finding the best prediction (BP) [which is the same as empirical Bayes (EB) or BLUP in this case] estimator that minimizes the MSE in the class of linear unbiased estimator of $\hat{\theta}_i$. The BP estimator is a linear combination of fixed and random effects and is given as

$$\begin{aligned} \tilde{\theta}_i(\tau^2) &= x_i' \tilde{\beta}(\tau^2) + (1 - B_i)(y_i - x_i' \tilde{\beta}(\tau^2)) \\ &= (1 - B_i) y_i + B_i x_i' \tilde{\beta}(\tau^2), \end{aligned} \quad (15)$$

where $B_i = \sigma_i^2 / (\sigma_i^2 + \tau^2) (i = 1, \ldots, m)$ $(i = 1, \ldots, m)$ and $\tilde{\beta}(\tau^2)$ is the weighted least squares estimator of β with weight $(\sigma_i^2 + \tau^2)^{-1}$. The BP estimator in Eq. (15) is the weighted average of the direct estimator y_i and the regression synthetic estimator $x_i' \beta$. If the survey is reliable (i.e., σ_i^2 is small), then B_i is close to zero and the direct survey estimate is the BP estimate. If σ_i^2 is large relative to τ^2, then B_i is close to 1, and the BP estimate is close to synthetic estimator $x_i' \beta$. When both τ^2 and β are unknown, they can be estimated by the classical methods, such as the standard analysis of variance estimator, maximum likelihood estimator, and restricted maximum likelihood estimator. Inserting $\tilde{\beta}(\tau^2)$ for β and $(\hat{\tau}^2)$ for τ^2 in $\tilde{\theta}_i(\tau^2)$, the following empirical best prediction (EBP) (which is EBLUP in this case) estimator of θ_i is given by

$$\hat{\theta}_i(y_i; \hat{\beta}(\hat{\tau}^2, \hat{\tau}^2)) = (1 - \hat{B}_i) y_i + \hat{B}_i x_i' \hat{\beta}(\hat{\tau}^2). \quad (16)$$

In estimating the variance component τ^2 by MLE, there is a possibility that the value of τ^2 could be negative, especially for a small or moderate m, the number of small area. If τ^2 is negative, one simply assigns a value of zero. In the Bayesian method, the value of τ^2 is always positive; the problem is how one chooses a prior distribution of τ^2.

A measure of variability of $\hat{\theta}_i$ is given by $\text{MSE}(\hat{\theta}_i) = E(\hat{\theta}_i - \theta_i)^2$, where the expectation is taken over the mixed model. The estimator of MSE can be obtained by estimating the MSE of BP. This naive estimator is an underestimate of the true measure of uncertainty since it does not take into account the variability of the variance components. Several attempts that have been made to estimate the MSE of EBP include the delta method, bootstrap, and jackknife resampling. Extensive research on the measure of uncertainty of EBP has been conducted by Butar, Chen, Datta, Lahiri, Ghosh, Fuller, Jiang, Pfeffermann, Prasad, Rao, and other researchers during the past two decades.

Extension of the basic area level model (Eq. 13) and unit level model (Eq. 14) includes multivariate models, models with correlated sampling errors, time series and cross-sectional models, spatial models, random error variance linear models, logistic linear mixed models, models for mortality and disease rates, exponential family models, semi- and nonparametric models.

Hierarchical Bayes Estimation

Hierarchical Bayes (HB) estimation is also used in small area estimation. The primary objective of this approach is to account for small area variation that is generally ignored by other small area approaches. The model is built in stages, hence the name hierarchical. Consider the Fay–Herriot model example discussed previously: (i) Conditional on θ_i, y_i's are independent with $y_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, \sigma_i^2)$; and (ii) prior distribution of $\theta_i \overset{ind}{\sim} N(x_i'\beta, \tau^2)$, $i = 1, \ldots, m$, where y_i is the income per capita from survey estimator for the ith area, σ_i^2 is the sampling variance of y_i, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})'$ is a known vector from tax return data for the year 1969 and the data on housing from the 1970 census, and θ_i is the true income per capita. Here, the goal is to find the posterior distribution or the posterior mean (which is the same as the Bayes estimator under squared error loss function). The hyperparameters of the Bayesian model are estimated from the marginal distribution of y_i, and the estimators are substituted into the Bayes estimator to yield the EB estimator. The EB estimator here is the same as the EBP or the EBLUP.

In HB estimation, prior distributions on the hyperparameters are specified, the parameters of interest are estimated by the posterior mean, and their variability is measured by the posterior variance. For example, consider the unit level model mentioned in Eq. (14) and rewrite it as the HB model: (i) $y_{ij} \mid \theta_i \overset{ind}{\sim} \text{Ber}(e^{\theta_i}/(1 + e^{\theta_i}))$, (ii) $\theta_i = x_i'\delta + e_i$, (iii) $e_i \overset{ind}{\sim} N(0, r^{-1})$, (iv) $\delta \sim$ Uni Uniform (R^p) independent of $r \sim \text{Gamma}(a/2, b/2)$, where a and b are known; $i = 1, \ldots, m; j = 1, \ldots, N_i$, y_{ij} is alcohol abuse for the individual j and county i, $\pi_i = e^{\theta_i}/(1 + e^{\theta_i})$ is the true proportion of alcohol abuse in county i, and $x_i' = (x_{i1}, \ldots, x_{ip}) \in R^p$ are the covariates from the 1995 survey, such as adult liquor law arrest rate, adult drug arrest rate fraction, and any drug or alcohol diagnosis rate.

The posterior means or proportions and the variances are not closed forms; therefore, numerical integrations may be used. An alternative to the numerical integrations is Gibbs sampling. Gibbs sampling is a Markov chain Monte Carlo (MCMC) sampling method that requires the knowledge of the full conditional distribution. The problem here is to choose the priors so that the posterior is proper and the samples can be generated. Arbitrary starting values will be assumed. The first few iterations of the simulated Markov chain will be discarded in order to reduce the effect of the starting values. The posterior mean generated vector and covariance matrix of θ_i can be approximated using a large number of θ_i values generated using the Monte Carlo method. Convergence of the method will be investigated. A computer program called Bayesian inference using Gibbs sampling (WinBUGS) is widely used to implement MCMC and to calculate posterior quantities from the MCMC output. WinBUGS runs are monitored using a menu-driven set of S-Plus functions, called Convergence Diagnostic and Output Analysis. The WinBUGS software package is freely available. The Gibbs sampling method is a popular method of Bayesian data analysis.

Design Consistency and Data Consistency

The concepts of design consistency and data consistency have received considerable importance in the small area estimation literature. Design consistency is a large sample property that ensures the convergence (with respect to the sampling design) of an estimator to the parameter of interest. This is certainly a desirable property since we frequently encounter a few areas with relatively large samples. Data consistency is another desirable property that provides direct estimates of larger areas simply by an appropriate aggregation of small area estimates.

See Also the Following Articles

Bayesian Statistics • Maximum Likelihood Estimation • Population vs. Sample • Sample Size

Further Reading

Butar, F. B., and Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small area estimators. *J. Stat. Planning Inf.* **112,** 63–76.

Fay, R. E., and Herriot, R. A. (1979). Estimates of income for small places: An application of James–Stein procedure to census data. *J. Am. Stat. Assoc.* **74,** 269–277.

Ghosh, M., and Rao, J. N. K. (1994). Small area estimation: An appraisal. *Stat. Sci.* **9**(1), 55–93.

Lahiri, P. (2003). A review of empirical best linear prediction for the Fay–Herriot small-area model. *Philippine Statistician* **52,** 1–15.

Meza, J., Chen, S., and Lahiri, P. (2003). Estimation of lifetime alcohol abuse for Nebraska counties. Unpublished manuscript.

Morrison, P. (1971). *Demographic information for cities: A manual for estimating and projecting local population characteristics, RAND Report R-618-HUD.* RAND, Santa Monica, CA.

Platek, R., Rao, J. N. K., Särndal, C. E., and Singh, M. P. (eds.) (1987). *Small Area Statistics*. Wiley, New York.

Prasad, N. G. N., and Rao, J. N. K. (1990). The estimation of mean squared errors of small area estimators. *J. Am. Stat. Assoc.* **85,** 163–171.

Purcell, N. J., and Kish, L. (1979). Estimation for small domain. *Biometrics* **35,** 365–384.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley, New York.

Schaible, W. L. (1992). Use of small area statistics in U.S. federal programs. In *Small Area Statistics and Survey Designs* (G. Kalton and J. KordosR. Platek, eds.), Vol. 1, pp. 95–114. Central Statistical Office, Warsaw.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling 0.5.* MRC Biostatistics Unit, Cambridge, UK.

Snowball Sampling

Richard Wright
University of Missouri, St. Louis, St. Louis, Missouri, USA

Michael Stein
Lindenwood University, St. Charles, Missouri, USA

Glossary

ethnographic An approach to social research that focuses on the definitions, values, and meanings that actors bring to the situation(s) being studied. Ethnographic research invariably seeks to explain behavior in terms of native nomenclature and categories and involves at least some direct fieldwork.
fieldwork The gathering of social data *in situ*, that is, in the setting in which it is naturally embedded.
representative sample A sample that proportionally reflects relevant characteristics of the total population from which it is drawn.

Snowball sampling is a chain referral method whereby a sample is constructed from a base of initial contacts, who are asked to provide introductions to their associates, who, in turn, are asked to refer others. This process continues until a sample has been built. Snowball sampling is designed for the explicit purpose of obtaining systematic information in situations in which convenience sampling is inappropriate and probability sampling is unrealistic.

Appropriate Uses of Snowball Sampling

Snowball sampling is useful primarily in the study of hidden populations such as intravenous drug users or active street criminals outside of clinical or institutional settings. These populations have strong incentives to conceal their identities and/or activities from outsiders. As a result, they are difficult to reach through conventional channels and instead require introductions provided by trusted insiders. As Glassner and Carpenter (1985) pointed out, the parameters of these "wild" populations typically are unknown and often unknowable, making random sampling impossible.

Although snowball samples may be criticized as unrepresentative, it is important to keep in mind that, in certain circumstances, they likely are far more representative than those drawn from a clinical or institutional setting. For example, no one would seriously claim that individuals attending a drug treatment program are representative of drug users or that prisoners are representative of criminals. In addition, people are animals and, like other animals, they do not behave naturally when they are institutionalized. Imagine going to a zoo to study the hunting strategies of lions. If the goal is to study real-world behavior rather than adaptation to institutional life, a snowball sample may yield better data.

Procedural Considerations

Getting Started

Perhaps the most challenging phase of the snowball sampling enterprise involves making initial contact with a member of the hidden population of interest to the researcher. By definition, such individuals keep a low profile and are unlikely to step forward of their own volition. Various ways of making this initial contact have been suggested. McCall (1978), for instance, recommended using a chain of referrals (essentially snowball sampling

within snowball sampling) to locate someone who can introduce the researcher to a member of the proposed study population.

> *If a researcher wants to make contact with, say, a bootlegger, he thinks of the person he knows who is closest in the social structure to bootlegging. Perhaps this person will be a police officer, a judge, a liquor store owner, a crime reporter, or a recently arrived Southern migrant. If he does not personally know a judge or a crime reporter, he surely knows someone (his own lawyer or a circulation clerk) who does and who would be willing to introduce him. By means of a very short chain of such referrals, the researcher can obtain an introduction to virtually any type of criminal.* (p. 31).

This strategy can be effective and efficient, but it has pitfalls. From a practical standpoint, relying on referrals provided by outsiders—especially members of agencies charged with punishing, treating, or otherwise controlling the population of interest to the researcher—may backfire, by raising suspicion that the study is part of an effort to identify deviants so they can be dealt with by authorities. Hidden populations are skittish and susceptible to such rumors at the best of times; they are likely to interpret the researcher's link to officialdom, no matter how benign, as a signal to stay away. Another pitfall associated with using a chain of referrals initiated through official channels is that the resulting sample may be highly unrepresentative. It is likely that such a sample will include a disproportionate number of individuals who are known to care and/or control authorities. By the same token, such a sample almost certainly will exclude the most secretive and sophisticated individuals who avoid associating with colleagues known to these authorities.

A commonly suggested alternative to using a chain of referrals to make initial contact with members of the proposed study population is to frequent locales favored by these individuals, and thereby establish rapport with them. This is a standard ethnographic technique that has the advantage of mimicking the normal interactional processes through which most relationships of trust are established. But this strategy, too, has several potential drawbacks. First, members of the group the researcher wishes to study may not congregate in particular locales. Second, even if group members do favor certain locales, the researcher may not know about them, especially beforehand when such information could be most helpful. Third, this strategy requires an extraordinary investment of time because the researcher must devote many hours to establishing a reputation for trustworthiness before attempting to initiate the snowball sampling process.

Another popular way to establish initial contact with members of the proposed study population is to enlist the services of a field-based recruiter, an insider who already has the trust and respect of that population. For example, a number of criminologists have used offenders who have retired from crime or who remain only peripherally involved in lawbreaking to gain introductions to active criminals such as residential burglars, drug dealers, and armed robbers. One advantage in using a trusted and respected insider to initiate the snowball sampling procedure is that the individual's reputation for integrity increases the likelihood that the people contacted will cooperate in the research. Another advantage is that an insider, almost by definition, knows things about the proposed study population that remain hidden from outsiders, and thus can help with everything from verifying the eligibility of potential participants to brokering the disputes and breakdowns in communication that often arise during encounters between researchers and the researched.

Despite these advantages, using an insider to make initial contacts can present difficulties of its own because the better that individual is connected to the proposed study population, the more likely it is that he or she will fail to appreciate the esoteric aims and methods of the researchers. This is perhaps best illustrated by the experiences of Wright and Decker (1994), who employed a well-connected ex-offender as a field recruiter to introduce them to active residential burglars.

> *Not all of the difficulties we experienced . . . were created by the offenders; the project fieldworker—who, after all, bridged legitimate society and the criminal underworld—sometimes failed to follow stipulated procedures and had to be reminded of the importance of adhering to legal and ethical standards. On one occasion, for instance, we were riding in the back seat of the fieldworker's car when we heard him mention to the offender sitting in front that he did not have any auto insurance. We immediately terminated our research for the day, gave the fieldworker some money, and told him to get insurance. He assured us that he would do so right away, adding, "but first I have to get a driver's license"* (pp. 29–30).

The strategies outlined in this section do not exhaust the list of those used by researchers to make initial contacts for a snowball referral chain, but they are by far the most common approaches to doing so. Depending on the population wanted to reach, other possibilities include advertising in local newspapers or answering ads in personal columns. Both have been used with some success by researchers in the past, though the latter strategy has been challenged on ethical grounds because people who place ads in the personals are not seeking individuals to study their behavior. None of these techniques is foolproof, and it often is advisable to try to initiate snowball referral chains through several sources at once, both as a way of enhancing the chances of success and as a way of reducing the risk of tapping into just one network of like-minded members who are atypical of the population as a whole.

Maintaining Momentum

Once the researcher has located an initial contact, the next challenge is to maintain momentum by persuading this and subsequent contacts to provide introductions to similarly situated others. This can be difficult because, by their nature, hidden and/or deviant populations are permeated with mistrust. Virtually any researcher who has used snowball sampling can testify to the fact that initially promising contacts sometimes end up being unproductive and have to be dropped. Even productive contact chains are vulnerable to disruption and have a tendency to break down over time. There is no foolproof way to prevent such things from happening, but researchers can take simple steps to promote the snowball referral process.

Perhaps the best way to encourage contacts to provide introductions to others in their social circle is to offer them a "finder's fee" for each successful referral. Such payments, however small, are absolutely critical to securing the cooperation of individuals enmeshed in certain sorts of deviant or criminal subcultures. For example, it is a cardinal rule among street criminals that you must never do anything for nothing. Moreover, compensating individuals for the time and effort it takes to arrange introductions on the researcher's behalf provides tangible recognition that they are performing a valued and valuable service.

On its own, the offer of a finder's fee may not be sufficient to convince contacts to provide introductions to their associates, especially if the amount of money involved is modest. But not all rewards are monetary in nature, and there are other things that researchers can do to facilitate the snowball referral process. For starters, they can inform contacts that the research likely will be published and that their efforts are critical to its success. It would be difficult to overstate the potential effectiveness of this simple strategy, which encourages contacts to refer others by making them feel that they are participating in something important enough to be published. Beyond this, researchers can help to maintain the momentum of the snowball referral process by doing all they can to fit in with their contacts. This does not mean trying to be one of them, but rather learning enough of the study population's argot and/or other distinctive mannerisms so as to be able to interact comfortably with them.

Another big part of fitting in with contacts requires researchers to adhere strictly to their promises of confidentiality, which is easier said than done in the case of snowball sampling. The chain referral nature of snowball sampling means that some contacts inevitably know one another, often quite well. It is not uncommon for a contact to ask what the person who referred them (or, alternatively, the person they referred) said to the researchers about an incident they both witnessed and/or participated in. In most cases, such inquiries probably reflect nothing more than the contact's curiosity, though even then it is likely that word of any betrayal of confidence, however inadvertent, will quickly work its way up and down the referral chain, alienating previous contacts and deterring potential future recruits from participation.

Deciding When to Stop

Because snowball sampling is typically employed to reach populations whose parameters are unknown and often unknowable, it can be difficult to determine the optimal point at which to terminate the recruitment process. When can researchers be confident that additional referrals will not yield important new insights into the hidden world they are studying? The standard answer to this question is that researchers should continue to recruit participants until "sample saturation" is reached, which means that no new information is being provided by additional referrals. Although this is a good rule of thumb for researchers to follow, it is a far from perfect strategy that almost invariably leaves a nagging sense that, had the recruitment process been extended further, new information might well have surfaced somewhere down the line. Indeed, Heckathorn (2002) suggested that researchers should aim for more than sample saturation, demonstrating that, under certain strict conditions, it may be possible to obtain valid population estimates using snowball sampling. While this may be true, once the data provided by new recruits become highly repetitive, the value of any additional information likely will be overridden by the financial and other costs associated with generating it.

Be that as it may, researchers often have little say in the decision about when to terminate the recruitment process. It is in the nature of snowball sampling that, for a variety of reasons, referral chains frequently stall of their own accord. Sometimes researchers are able to overcome this problem by initiating a new referral chain, but success in doing so is far from guaranteed, especially when the breakdown results from a heightened perception on the part of members of the study population that the risks of participation have begun to outweigh any potential benefits. For example, a number of researchers have had their referral chains stall when the chance arrest of a study participant, or some other untoward event, became linked in the minds of would-be recruits with their project.

Snowball Sampling in Action

A study of armed robbers in St. Louis, Missouri by Wright and Decker (1997) provides a textbook example of the promises and pitfalls of snowball sampling in action. Concerned that the prison environment might distort the responses of incarcerated armed robbers, Wright

and Decker decided to try to locate and interview individuals actively involved in committing such offenses. To do this, they employed a snowball sampling technique similar to one they had used to recruit 105 active residential burglars for an earlier study of housebreaking. In their burglary study, Wright and Decker had initiated snowball referral chains by using a field-based recruiter who was well known and respected by several groups of street criminals operating in and around St. Louis. This fieldworker, an ex-offender who had retired from crime after being shot and paralyzed in a gangland-style execution attempt, had previously supported himself for many years as a highly skilled thief. He had been arrested just a few times and was never convicted. As a thief, he had acquired a solid reputation among his fellow criminals for toughness and integrity.

Wright and Decker turned to this individual once again for help in making contact with active armed robbers. He began by approaching some of his former criminal associates. All were still active offenders, and he found three who currently were doing armed robberies. He explained the research to them, stressing that it was confidential and that the police were not involved. He also informed them that those who agreed to be interviewed would be paid $50. He then asked the contacts to put him in touch with offenders actively involved in committing armed robberies, saying that they would receive $10 for each successful referral.

In adopting this strategy, Wright and Decker hoped to set in motion a self-perpetuating chain of referrals that would smoothly lead from one active armed robber to the next by word of mouth. Figure 1 outlines the networks through which the offenders were located. It also illustrates the uneven pace at which these networks were expanded.

Perhaps the best way to clarify the recruitment process is to select one of Wright and Decker's respondents, say, no. 56, who is situated about halfway down the figure just to the right of center, and identify the chain of referrals that led them to this individual. In this case, the fieldworker contacted a female acquaintance who made her living exclusively through non-violent street crimes. She introduced him to three active armed robbers—nos. 15, 16, and 21—but, more importantly, she also put him in touch with one of her male friends, another petty criminal, who helped the fieldworker find more than two dozen active armed robbers. Among these armed robbers was no. 24; he referred three additional offenders, including no. 36. Armed robber no. 36, in turn, provided the fieldworker with two additional contacts, one of whom, no. 50, introduced him to five further armed robbers; the last of these robbers was no. 56. As can be seen, the majority of the armed robbers were not referred directly by the original fieldworker, but rather through the efforts of various actors in the street scene, such as heroin addicts, gang members, and petty criminals. Wright and Decker almost certainly would not have been able to locate many of these armed robbers on their own, much less gain their cooperation.

Buried within Fig. 1 are various indicators of the difficulties that Wright and Decker encountered in generating their sample. Note, for instance, that armed robber no. 04, the first person that the fieldworker contacted, referred three other robbers before agreeing to be interviewed himself. When initially approached about the project, he denied any personal involvement in robbery, but added that, as a junkie and small-time heroin dealer, he came across a lot of people who did commit such crimes. Only after being named as an accomplice by one of his referrals did he admit to having taken part in the occasional stick-up. He had not admitted his involvement earlier because he was worried about the possibility of being set up for an arrest.

Armed robber no. 04 went on to provide Wright and Decker with many additional referrals. In fact, he acted as a backup fieldworker for them when, early in their research, the original fieldworker decided to slack off and went for over a month without recruiting a single armed robber. Desperate to keep the snowball growing, Wright and Decker turned to armed robber no. 04, agreeing to pay him $50 for each armed robber he located for them. This worked well in expanding the sample in the short term, but it led to considerable resentment on the part of the original fieldworker, who quickly regained his enthusiasm for the recruitment process. From then on, the two fieldworkers had to be kept apart; interviews had to be staggered accordingly and this further complicated Wright and Decker's efforts to construct a suitable sample.

Given the tensions between the two fieldworkers, Wright and Decker had to be careful to avoid the appearance of playing favorites. Each fieldworker brought his own unique mix of street connections to their project, and they did not want to alienate either one for fear of closing off their only viable conduit to potentially important subgroups of active robbers. As a close reading of Fig. 1 demonstrates, the armed robber samples generated by the two fieldworkers had different demographic characteristics. Almost all of the individuals recruited by armed robber no. 04 were older, African American males, whereas the original fieldworker was able to locate a more diverse range of armed robbers, including a sizable number of juveniles, over a dozen females, and several whites.

In an instructive aside, Wright and Decker note that none of the individuals referred by armed robber no. 04 seemed willing to put them in touch with fellow armed robbers directly. Instead, they insisted on using armed robber no. 04 as an intermediary. This suggests that he used his influence to maintain a firm grasp on the referral

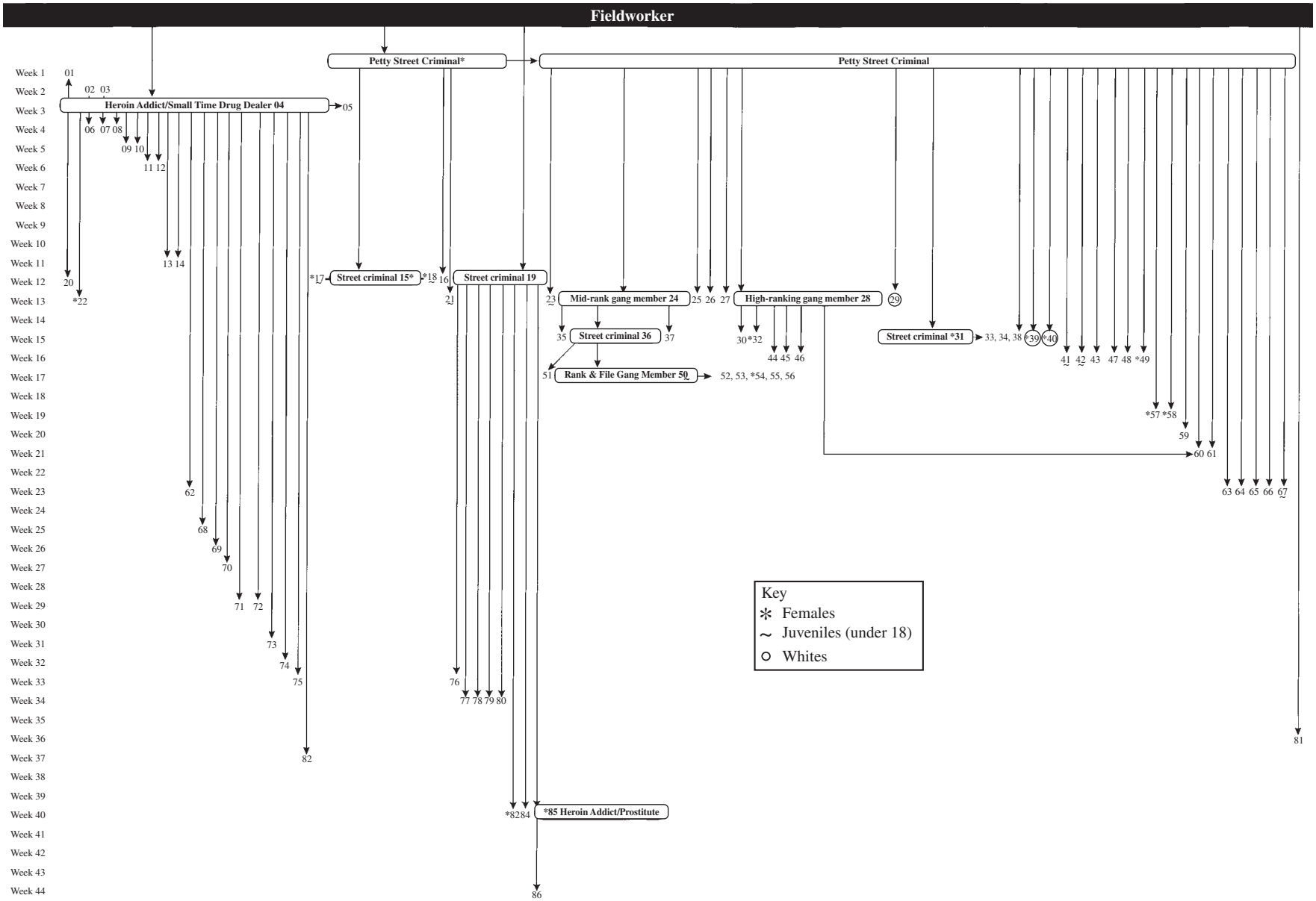


Figure 1 A snowball sample of active armed robbers in St. Louis, Missouri. (Reproduced with permission of the Northeastern University Press.)

process, presumably because he wanted the $50 finder's fee. Also instructive is the fact that, as Fig. 1 reveals, armed robber no. 04 had no success whatsoever in penetrating networks beyond the one in which he himself was a member. Wright and Decker speculated that this likely was because he inspired little trust outside his own circle of criminal acquaintances and had strong reasons to limit his recruiting efforts to his own neighborhood:

> *We know, for example, that armed robber No. 04 has a history of robbing illicit crap games and street corner crack dealers; in doing so, he has made deadly enemies and needs to "watch his back" wherever he goes, especially when traveling outside the boundaries of his own neighborhood.* (p. 22)

The most serious recruitment problem that Wright and Decker faced concerned armed robber no. 81, who is located on the extreme right-hand side of Fig. 1. He agreed to talk to them only after being repeatedly assured by the project fieldworker that they were not working for the police. During his interview, Wright and Decker discovered that he was a very well-connected armed robber and could serve as a valuable source of referrals. When they asked him if he might be willing to introduce them to his associates, he was initially reluctant but, after considerable reassurance that it was safe to do so, finally consented. As Fig. 1 indicates, however, this never happened; he was arrested and charged with armed robbery only hours after speaking to Wright and Decker. While the researchers had no hand in his arrest, the coincidence did not go unnoticed by other armed robbers in the area and, as a result, it became increasingly difficult for them to generate additional referrals. With the help of their fieldworkers, Wright and Decker did manage to induce a few more armed robbers to participate in their study, but the going got harder and harder and they terminated their research shortly thereafter.

The Enduring Relevance of Snowball Sampling

Snowball sampling is a powerful means of accessing hidden and/or deviant populations by using members of such populations as a source of recruitment for additional participants. Despite this fact, snowball sampling almost never proceeds in a straightforward manner for the simple reason that the sorts of populations it is used to access are often suspicious, unreliable, and unaccustomed to the arcane demands of systematic social research. In this sense, the term "snowball sampling" is a misnomer. Snowballs, after all, grow steadily bigger as they roll down an incline, whereas snowball samples grow in fits and starts, if they grow at all, and have to be restarted at irregular intervals owing to circumstances over which researchers often have little or no control.

The bottom line is that, for all its difficulties, snowball sampling frequently is the most effective strategy realistically available to researchers for contacting and studying hidden and/or deviant populations beyond the bounds of clinical or institutional control. For example, it has been used with substantial success in studies of populations at risk for AIDS (especially intravenous drug users and men who have sex with men), the homeless, and various sorts of active street criminals (including not only active armed robbers, but also active residential burglars, active drug dealers, and active car thieves). How else could such populations be accessed? As long as social scientists remain interested in the study of these and other hidden populations, snowball sampling will continue to be an important tool for gaining access to the real world of deviants and deviance.

See Also the Following Articles

Ethnography • Field Experimentation

Further Reading

Biernacki, P., and Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociol. Meth. Res.* **10,** 141–163.

Glassner, B., and Carpenter, C. (1985). *The Feasibility of an Ethnographic Study of Property Offenders.* U.S. National Institute of Justice, Washington, DC.

Heckathorn, D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Probl.* **44,** 174–179.

Heckathorn, D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Probl.* **49,** 11–34.

McCall, G. (1978). *Observing the Law.* Free Press, New York.

Watters, J., and Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Probl.* **36,** 416–430.

Wright, R., and Decker, S. (1994). *Burglars on the Job: Streetlife and Residential Break-ins.* Northeastern University Press, Boston, MA.

Wright, R., and Decker, S. (1997). *Armed Robbers in Action: Stickups and Street Culture.* Northeastern University Press, Boston, MA.

Wright, R., Decker, S., Rooney, A., and Smith, D. (1992). A snowball's chance in hell: Doing fieldwork with active residential burglars. *J. Res. Crime. Delinq.* **29,** 148–161.

Social Economics

Christine Rider
St. John's University, Queens, New York, USA

Glossary

institutionalism Accepting that institutions (organized patterns of behavior) influence individual economic activity.
normative economics Adding value judgments to economic analysis.
personalism Humans are a fusion of individuality and sociality, of body and spirit, of reason and emotion.
positive economics Economics as an objective, value-free, quantifiable social science.
rational economic man The autonomous economic agent who rationally calculates.
subsidiarity Any policy action should be taken by the private organization closest to the problem, and should be replaced by a government agency only when the private organization cannot function effectively even with government assistance.
utility Satisfaction realized from consuming a good or service.

Social economics is economics practiced as a social science involving human beings, who, in acting as economic agents, are often called on to address moral issues. Economics is the study of the material world, the study of the production, distribution, exchange, investment, and consumption of goods and services that are necessary for human survival (at the minimum) and which today make human life pleasant and comfortable for many. These activities take place in societies, human organizations that develop in order to organize and regulate human behavior. Social economics is the branch of economic analysis that explicitly recognizes the social dimension of economic activity. It recognizes that in economic affairs, humans act not only as individual beings as assumed in mainstream economics, but also as social beings, and, more importantly, recognizes that economics is a *moral* science. That is, it investigates the formation and impact of the values that influence what people and institutions in a society do; it adds an element of accountability to human behavior that is lacking in an "objective" approach, which is why many believe it is more relevant to modern reality. There are several strands of social economics, but all share a concern for understanding the philosophical roots of the discipline, a preference for identifying and solving social problems, and a holistic approach to analysis.

To summarize, social economists take account of the philosophical foundations of the discipline; attempt as accurate a description of an economy as possible; and use these insights to inform economic policy, in order to better provision human needs and wants.

Origins

Catholic Thought

Although it is possible to identify many earlier "social" economists who were outside and critical of the mainstream orthodox economics, social economics as a diverse and heterogenous strand of economic thinking began to coalesce in the period beginning 1941. At this time, the Catholic Economic Association, influenced by Thomas F. Divine and Bernard W. Dempsey, set out to replace the logical positivism of mainstream economics with a more specific, policy-oriented approach, which would necessarily involve incorporating ethical issues explicitly. Such a concern with normative values definitely contrasts with the distinction made by orthodox economics between normative and positive thinking, with the understanding that the former has no place in an objective science.

The emphasis on personalism is the main contribution to social economics made by Catholic social teaching,

according to Edward J. O'Boyle. What this means is that human beings are unique in the universe in that they alone are created in the image of God, and therefore cannot be reduced to the status of objects or commodities. While recognizing that humans obviously have physical/material needs, this view also recognizes the many other needs that make people human beings: for example, the need for work, for self-expression, for social relationships with others. Also, if basic economic needs are not met by the private sector, then, following the principle of subsidiarity, the state has a responsibility to help fill that need, legitimatizing the role of the state in the economy, a role that is often downplayed by mainstream economists.

Humanism

The humanist perspective, which can be traced back to Simonde de Sismondi, centers on the human person and human welfare. In economic affairs, the human person acts freely and independently, but, unlike the Rational Economic Man (REM) of standard economics, this perspective holds that the individual is not solely motivated by self-interest: interdependence encourages activity that enhances the common good. While the focus of REM might be appropriate if each economic agent had only one end to pursue at any one time, when there are plural goals, the standard approach fails. Choosing between different ends cannot be modeled mathematically—how to maximize utility given constraints—and instead requires other influences on choice making, especially when the choices are qualitatively different. It is the process of making these choices that involves individual freedom of judgment, not mechanical computation, and which is influenced by social relationships. This social setting calls for defining the economic agent more broadly than the autonomous, self-centered, utility-maximizing individual of conventional economics.

Ultimately, the end result is an economics that encompasses not only the needs of humans that are specifically individual and which, when met through individual action is called the good of the individual, but also the needs that derive from humans living together as one community, which when met through community action is called the good of the community, or simply the common good. These policy prescriptions designed to serve the common good clearly overlap with those derived from other influences on social economics: the right to material necessities, the right of economic democracy, and the right of future generations to economic sufficiency. Although the humanist strand derives from a nonreligious perspective, it essentially ends up in the same position as a religious one: that economic agents are neither objects nor instruments, but human beings with the ability to reason morally and act in accordance.

Institutionalism

Institutionalism is a uniquely American school of economic thinking in which seven key ideas can be identified. First, the economy should be studied as one unified system, in contrast to the orthodox focus on the many individual components that are then aggregated. Second, institutions—such as organized patterns of behavior, customs, beliefs, and laws—have an importance in influencing economic life. Third, society and the economy are seen in evolutionary terms as constantly changing into something new and unpredictable. Fourth, this in turn implies that the mainstream view that whenever shocked from an original equilibrium position, society and the economy always and predictably return in cyclic fashion to the same equilibrium state, is misplaced. Fifth, institutionalists believe that economic and social analysis is best understood in terms of conflict of interests between different groups in society, rather than as being ultimately harmonious. Sixth, a concern with social justice also colors this approach, with a more activist approach to social reform, the ameliorative aspect of social economics. Finally, institutionalists downplay the utility maximization goal of orthodox economic analysis, looking instead for a more nuanced and realistic approach to economic behavior. As can be seen, these concerns overlap with many of the other strands of economic thought making up social economics.

The founder of this school is considered to be Thorstein Veblen (1857–1929), who was writing at the time when American big business was flourishing, following post-Civil War industrialization. This experience seemed to leave many behind in the race to prosperity, and most likely influenced the emphasis on the "careful study of the economy as it really is," which is a hallmark of social economics. This emphasis was aided by the founding, in 1920, of the National Bureau of Economic Research (NBER) by several of Veblen's students, including Wesley Clair Mitchell. The NBER was established to collect and analyze economic data, especially for the study of business cycles, the periodic fluctuations in the level of economic activity.

Institutionalists, like other social economists, are critical of orthodox economic analysis, feeling that it is too protective of the status quo, even when reform is needed. While the Association for Evolutionary Economics is the main association for institutionalists, the overlap with social economics is clear. After 1970, when the Catholic Economics Association changed its name to the Association for Social Economics, many institutionalists joined forces, further blurring the distinction.

To summarize, all these different strands share a belief that economics should be practiced as a moral science where values do matter because they influence economic affairs through human behavior.

Similarities and Differences

Mainstream Economics

Most standard introductory economics textbooks will define economics as the study of the allocation of scarce resources among unlimited competing ends. The aim of economic activity is to maximize utility, the pleasure or satisfaction that comes from meeting material needs, given the constraint of limited resources. Certain key issues are as follows. First is the issue of resource *scarcity*; resources include both naturally occurring ones such as human labor power, agricultural and mineral resources, and those produced resources such as buildings, tools, and equipment. Second, it is assumed that there is potentially no limit to the uses to which these resources can be put, whether to meet basic human needs for food, clothing, and shelter, or to meet a variety of "wants"—for entertainment, enlightenment, luxury, dissipation, etc. In the more advanced industrial and postindustrial societies, these "nonbasic" wants have definitely multiplied, and entire industries, such as advertising and vacation travel, have developed to support and encourage the development of these new "needs."

Given scarcity and unlimited wants, economics can alternatively be described as the science of choice. For economists in the mainstream tradition, the task is to construct models that illustrate how resources are allocated to different uses, and, the parallel concern, to illustrate how income is distributed. (In a market capitalist economy, the main frame of reference, these two ends are two sides of the same coin: ownership of resources permits income to be earned which is then spent on the output that resource use gives rise to.)

Central to this tradition is the construct of the Rational Economic Man, the economic agent at the center of the model. Because economics aims to be as positivist a discipline as any of the natural sciences, but because of the impossibility of running laboratory experiments to test its hypotheses, REM is a convenient, objective simplification. REM has all the complete information necessary to make decisions, is not swayed by any noneconomic (say emotional, religious or altruistic) influences, and knows how to accomplish his goals. That goal is to maximize something, given the constraint of limitations. So if REM is a consumer, the task is to maximize the utility (pleasure or satisfaction) derived from the different choices of consumer goods and services that can be bought with a given income. If REM is a producer, the task is to maximize output produceable from the resources available. If REM is a profit-maximizing entrepreneur, the point is to maximize profit (net revenue) from the production process.

The ultimate goal for society is to expand the output of goods and services as much as possible. Also, the more efficient markets are, the better, because efficient market operations minimize waste and therefore make achieving this goal easier. With no market imperfections (defined as being anything which cannot be captured in the price, such as the cost of pollution, or as any noneconomic influence on prices, any of which would make this idealized state more difficult to achieve) the economy tends to a general equilibrium where goods and services are being produced in exactly the right amounts that will maximize society's utility. No policy intervention to redistribute income is necessary, because in this general equilibrium, no one person can be made better off without someone else being made worse off.

This position is called Pareto optimum, and describes an equilibrium, a point of rest, to which an economy will tend if it is composed of Rational Economic Agents operating in perfectly free markets which are not subject to any outside interference. The policy implication of the mainstream tradition is then clear. If prices in these markets do contain all the information necessary to achieve socially and individually desirable results, then it is important to keep them working this way. This legitimatizes state intervention only to remove imperfections; intervention is reactive, piecemeal, and justified only when needed. The survival of the (economically) fittest is the outcome of the operation of a competitive market system peopled by agents motivated by self-interest, and which will automatically produce a harmonious outcome.

Differences

In order to understand how and why social economics differs from orthodox mainstream (or neoclassical) economics, it is useful to start with the work of Adam Smith, the 18th century moral philosopher generally regarded as the founder of modern economics. Most strands of economic analysis can be traced back to Smith. Smith wrote two major books. In *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776), he tried to understand how a market economy held together. In this, his 18th century enlightenment views—in which an order based on natural law replaced a belief in the divine origin of things—are clear. As in the natural world which obeyed certain natural, and therefore rationally explicable, laws, so too could economic activity be understood once the laws which governed it were revealed. Central to this was his concept of the "invisible hand": the seemingly chaotic activities of markets where individuals pursued their own self-interest without any thought for others actually resulted in the common good *as though directed by an invisible hand*. That is, the free movement of prices and the price system effectively regulated activity. This strand is central to the subsequent development of orthodox economic thought.

His other book, *The Theory of Moral Sentiments,* was published in 1759, 17 years earlier than *The Wealth of Nations,* and is the work that shows how individual actions can produce social harmony rather than the anarchy that might be thought to result from the pursuit of self interest. Harmony results in societies because there are moral forces that restrain individual selfishness, such as, for example, sympathy, generosity, compassion, esteem, and mutual feelings for others. This strand of thinking can be seen as the origin of the social dimension of human economic behavior that is underscored in social economics.

Ethics

While much of social economics and the work of social economists themselves is critical of the mainstream, neoclassical tradition, there are certain identifiable themes that also make it a cohesive intellectual tradition in its own right. It is critical because of the observations that, in real life, market outcomes are not necessarily harmonious, do not result in an equitable income distribution, and leave many important human needs, both economic and noneconomic, unserved or underserved.

The first theme is the explicit recognition of the importance of the discipline's philosophical foundations: the ideas and values that influence the construction of the economic systems which coordinate economic behavior, and which also influence the approach of the economic analyst. How values are formed in a society and transmitted to society's members and groups is a valuable acknowledgement that no science in which human behavior is the subject can be value free. Social economics is concerned with "normative, moral—philosophical, and even theological perspectives" (from Nitsch) because it is concerned with the creation and perpetuation of a just economy and society. Orthodox economics sees as its ultimate goal the maximum expansion of material things, and while social economics recognizes the importance of efficiency and abundance, it also prefers that the resulting wealth be distributed equitably, and that minimal human needs are addressed adequately.

If the focus is simply on analyzing and promoting the operations of free markets in an objective way, this reduces people (labor), land (natural resources), and even money to commodities, or things, that can be produced for sale, or discarded if there is no demand for them. To social economists, this commoditization harms the creation of a just, humane society. Human beings are not things, hence the historical reality of the development of a protective safety net, in the form of the welfare state and government intervention, is not an imperfection, but rather a necessary element in the development of a just society. Similarly, social economists would not view money as neutral, as a commodity, but rather see it as a socially created mechanism that assists the integration and functioning of economic activity. The former view simply observes that a self-regulating market system automatically restores stability when beset by financial crises. The latter view sees these crises as harmful to human lives and the way production is organized, hence making it imperative that action be taken to remedy the problems. This difference highlights one way in which economics is seen and practiced as a moral science.

To be more specific, there are four ways in which values matter. First, they matter because social economists are value-directed: their values help inform the questions or issues to study, and the possible remedies to suggest. Second, because the values people hold influence how they act, it is important to see how these values are formed if we are to understand human action. Third, values make social economics problem-oriented, in that the search for the causes of problems becomes a holistic one, ready to move outside the strict boundaries of economics whenever the causes are noneconomic. Finally, values matter because social economics is ameliorative, especially when it comes to trying to make economic processes work better for those left behind.

Social

As noted, orthodox economic analysis centers on the construct of the Rational Economic Man, who springs, *ab ovo,* with a complete set of tastes, preferences, abilities, information, and so on. The actions of this atomistic individual are independent of the institutional context in which choices are made; little concern is given to where these attributes come from. In contrast, social economics focuses on the interdependency of human actions that take place in an institutional structure that both creates and is created by human actions, values, and history. Rather than focus on individual maximization of utility, a social economist will also be concerned with the social groupings that influence behavior, and on the achievement of the common good. Institutionalists in particular analyze this impact, but social economists deriving from the Catholic social thought tradition also analyze the impact on the wider community.

A good example of this is the periodic analysis of papal encyclicals, which provide guidance on the incorporation of ethical values into everyday life. This dovetails with social economics' concern to encourage the development of a just society, which is also economically effective by encouraging standards of conduct that do more than just serve one individual's interests. For example, the 1991 papal encyclical, *Centesimus Annus,* does not criticize private property and free markets, but does denounce consumerism as a moral vice. The policy implication here is that, especially in wealthy societies, "superfluous" income should be spent and capital formation attended to in ways that will help transform the world into a more equitable place. This policy orientation applies,

for example, to pension fund managers, who, following these prescriptions, would attempt to use these funds to further the development of human potential rather than simply seeking the highest quarterly return.

Historical

Mainstream economics is ahistorical; its purpose is to uncover those timeless laws that govern economic behavior at any time and in any place; it is essentially naturalistic, assuming that the natural world can be explained using mechanical concepts. This justifies a concentration on equilibrium analysis. Social economics embeds economies into their historical context and incorporates the reality of continuing institutional change and evolution. This limits the significance of equilibrium analysis, and of finding those tendencies that will produce a harmonious outcome. The reality of historical change is especially noticeable in those economists who favor an evolutionary approach to the discipline. It also makes it more appropriate for the study of nonmarket economies, for example, low income, developing economies or economies in transition from central planning.

Policy Oriented

Social economics emphasizes normative values, which is necessary for any policy application because value judgments have to be made to devise policy. That is, if policy is intended to change an undesirable reality and achieve a desirable end, the determination of "desirable" and "undesirable" is necessarily normative. Also, to state that social economics is especially concerned with social justice and welfare is also stating a normative value set. "Social justice is a set of normative values which define and specify ethically correct relationships among persons. Human welfare is a set of normative values which provide a standard for the measurement of the well being of a people" (from Hill).

One way in which this could be encouraged is by developing systems of social accountability in the private sector. This involves creating electoral systems of governance, and problem-resolving judicial systems in order to represent and further the common good. Social economists have suggested a variety of possibilities, some borrowed from actual practices in many countries. These include (but are not limited to) mutual stock ownership schemes, codetermination, workers' cooperatives, as well as community corporations (land trusts, financial intermediaries) which have a responsibility to the community. Behind all these different suggestions is the (old) idea of a self-governing, civil economy responsive to the public interest.

Historical Contributions

While there are many writers on economic topics who share many of the concerns of social economists, some can be more easily identified as "social economists." Although it may seem artificial to do so, this permits a distinction to be made between "social economists," Marxists, socialists, and other social critics and dissenters from accepted orthodox economic theory.

What follows is a suggestive listing; it by no means includes all those economic writers whose insights overlap with those of social economics, especially with reference to the present day.

Simonde de Sismondi (1773–1842)

Simonde de Sismondi was a Swiss economists who was among the first to criticize the then-economic orthodoxy, classical economics, for its reliance on the self-regulating nature of free markets, which were supposed to produce equilibrium tendencies. While in England during the early part of the 19th century, he was horrified by the appalling conditions in which a large part of the population of this wealthy country lived, which coexisted with enormous wealth of a privileged few. His argument was that because the maximization of output was unlikely to coincide with the greatest happiness of the population, state intervention to ensure a living wage for workers would be preferable. He favored greater state intervention to promote a more equitable distribution of income.

Twentieth Century Social Economists

Thorstein Veblen, the founder of institutionalism, never accepted the taxonomic, natural law-based approach to economic analysis that was common at the turn of the 20th century. Instead, he incorporated insights from other disciplines, including philosophy, history, and the natural sciences, to redefine economic activity as essentially evolutionary, and which is influenced by economic institutions, the customary and habitual methods by which societies come into contact with nature to meet their material needs.

Veblen's evolutionary approach saw contemporary American business culture as essentially wasteful, where work (and workmanship) were looked down on while wealth provided status. Affluence must be made obvious in conspicuous consumption, conspicuous leisure, and conspicuous waste (his expressions). His most popular book was *The Theory of the Leisure Class* (1899), in which he outlined his ideas of how the wealthy lived, a contrast with the neoclassical view of consumer sovereignty based on rational decision making.

Other early 20th century economists who contributed to the analysis of issues central to social economics include (alphabetically): Clarence E. Ayres, who emphasized institutional obstacles to technological change; John Maurice Clark; the historian and labor economist, John R. Commons; and Joseph A. Schumpeter.

Schumpeter's most famous contribution was his theory of creative destruction as an important explanation of both economic development and business cycles. This explains the uneven nature of economic change over time, caused by periodic clusters of innovations that require new materials, new equipment, and new methods of organizing production that make older technologies obsolete. Hence periodically, old equipment gets discarded and workers with outdated skills become unemployed, while the burst of new business investment associated with the introduction of new technologies leads to a business expansion and prosperity for those associated with the new industries and new skills. The introduction of these innovations, therefore, change not only material conditions of life, but also the distribution of political power and the institutional underpinnings of the economy.

Although not an institutionalist, Kenneth Boulding shared with them a dislike of conventional economics, and a willingness to use ideas from other social sciences to illuminate and explain economic reality as it unfolds through evolutionary change.

Similarly, John Kenneth Galbraith was, like Veblen, critical of the neoclassical conventional wisdom, and he adopted a more evolutionary approach to understanding modern capitalism. This, he wrote, was dominated by large enterprises rather than atomistic consumers, and continued prosperity for these enterprises depends on creating new wants. In this world, there is a social imbalance as the public sector is starved for funds while the output of commodities for private consumption expands, creating what he calls private affluence amid public squalor. Some key phrases of Galbraith's work include "countervailing power," which describes the way large enterprises deal with each other; the "affluent society"; and "technostructure," which refers to the new technologies that are vital to the survival of large firms and the technocrats who have more influence over operations than do owners and managers.

Some of Galbraith's policy implications are at odds with conventional thought. For example, in the United States, antitrust laws were developed to try to control if not limit the growth and behavior of large firms because of the perception that the resulting control of markets by these firms would result in high prices. Galbraith disagreed, saying that these firm's pricing policies are not in fact excessive; the real problems that are caused are those associated with an inequitable income distribution.

Contemporary Practitioners

There are many economists and related social scientists currently working in the social economics tradition. For example, working in the Catholic social thought tradition is Edward O'Boyle; representing a radical institutionalist perspective is William Dugger; and Mark Lutz is a proponent of the humanist tradition. Warren Samuels has applied the social economics viewpoint to a wide range of topics. Some other well-known social economists include David Ellerman, John Elliot, David George, Hans Jensen, Ronald Stanfield, and Charles Wilber. In addition, surveying past and recent issues of *The Review of Social Economy* gives a useful survey of both the topics covered by social economics, and of at least some of the current writers in this tradition, especially the younger ones whose names are not yet so widely known as these.

Publications

There are two main publications specifically associated with social economics, and many similar ones published by associations with overlapping memberships and interests. *The Review of Social Economy* is a quarterly publication of the Association for Social Economics. It "... investigates the relationships between social values and economics and the relation of economics to ethics, and focuses upon the social economy that encompasses the market economy. The journal is sponsored by the Association for Social Economics, by charter a pluralistic organization that accommodates different approaches to economics. Among the themes pursued are justice, need, poverty, cooperation, income distribution, solidarism, equality, freedom, dignity, community, pragmatism, gender, environment, economic institutions, humanism, economic methodology, and the work of past social economists" (from the *Review*'s Aims and Scope). The ASE also publishes the semi-annual *The Forum for Social Economics*.

The monthly *International Journal of Social Economics*'s mission statement reads as follows: "Increasing economic interaction, allied to the social and political changes evident in many parts of the world, has created a need for more sophisticated understanding of the social, political and cultural influences which govern our societies. The *International Journal of Social Economics* provides its readers with a unique forum for the exchange and sharing of information in this complex area. Philosophical discussions of research findings combine with commentary on international developments in social economics to make a genuinely valuable contribution to current understanding of the subject and the growth of new ideas. Coverage includes: economics and ethics; nuclear arms and warfare; economics of health care; the disintegration of the Soviet Union; religion and socioeconomic problems;

socioeconomic problems of developing countries; environmental sciences and social economics.

Institutions

The Association for Social Economics is the main association for those interested in social economics. It was established in December 1941 as the Catholic Economic Association, and when in 1970 it decided to broaden its focus, it renamed itself The Association for Social Economics. (The ASE was also a charter member of the Allied Social Sciences Association, the umbrella organization for various economics associations, which organizes annual meetings of these groups.) The ASE was formed "to advance scholarly research and writing about the great questions of economics, human dignity, ethics, and philosophy. Members seek to explore the ethical foundations and implications of economic analysis, along with the individual and social dimensions of economic problems, and to help shape economic policy that is consistent with the integral values of the person and a humane community" (from the ASE website).

During the 1980s, a new organization, the Society for the Advancement of Socio-Economics, came into existence, due to a large extent to the efforts of Amitai Etzioni. Both SASE and ASE share a concern that the market-efficiency oriented, individualistic, and mechanistic approach to economics of the mainstream tradition is inadequate, and both emphasize the importance of the social setting of economic behavior. However, socioeconomics sees itself as a better tool of analysis for a positivist explanation of reality. Key features include a broader sense of the individual, an "embedded market" (the economy seen as part of the surrounding society), and the central role of power relationships. Social economics adopts a normative, value-driven approach, while socioeconomics is intended to be more scientific and positive. Furthermore, the latter is intended to appeal to a broader group of social scientists, including psychologists.

Areas of Special Concern

Social Justice

An emphasis on social justice is an integral part of social economics, informing such areas of interest as income distribution, globalization, trade, poverty, and inequality in general. For example, while mainstream international economics theorizes that the opening up of world trade tends to lead to convergence—of prices, income levels and living standards—empirical evidence reveals a widening gap between rich and poor nations, and between rich and poor within any one society.

Many of those who combine social economics with Catholic social teaching emphasize studies in this area. If human material needs are not met, justice requires that action be taken, hence a concern with policy, which requires knowing where and how to intervene. Some economists in this tradition believe that strengthening the family and the community are important here, as both are where individuals spend most of their time, and are (or should be) the means by which unmet needs can be addressed. This particular focus avoids what could otherwise be a heavy-handed state intervention: heavy-handed because the state is too distant from the individual, and too often does not permit participation in the decision-making process, hence making the policy action less democratic.

An emphasis on social justice also provides an opportunity for the selective use of planning, because attempting to influence events or create institutions that will protect individuals makes democracy consistent with meeting human need. The mainstream approach is biased toward the operation of unfettered free markets, individualism, and free enterprise. However, given the realities of power relationships, and of concentrated, imperfect markets, one individual's decisions can have an adverse effect on others, which cannot be resolved by the operations of autonomous markets. But planning—giving an element of social control over society's resources—can do this, which implies the development of institutions to represent all the stakeholders in economic decisions, and which can add an element of accountability to actions.

Adherence to the economic orthodoxy encourages a policy prescription involving the extension of markets to activities that are not traditionally market based. Thus, for example, in Western Europe in the 1990s, and in the former centrally planned socialist economies in transition, the trend to privatization of state-owned enterprises is seen as the way to improve, for example, the provision of electricity, communications, and transportation services. In the United States, problems in health care especially, but also in education, have also been viewed as being candidates for transfer to market forces. While social economists obviously favor improvements in these and other areas that have an impact on human well being, they question whether treating education or health care like the production of automobiles or other commodities is justified. In all cases, the policy prescriptions compatible with this view look not just at maximizing output and income, but also at the impact on individuals' economic and noneconomic needs, on society as a whole, and on the implications for the future.

Such a consideration can also be adapted to the issues of international inequality, and question whether globalization and encouraging the development of markets

is always and everywhere the solution to problems of economic development.

Environment

Concern for ecological sustainability and the natural environment becomes of interest to social economists because it involves respect for the well being of current and future generations. If the goal of economic activity is merely to maximize something now, then future conditions—whether of scarcity or plenty—do not matter. (The implicit assumption is that, given conditions in which free competitive markets can operate, solutions will always appear through the invisible hand.) However, social economists would prefer that current needs should be met in a way that does not compromise the ability of future generations to meet *their* needs.

Standard economics approaches environmental issues in a market-oriented way. If there are externalities, in the form of pollution, resource degradation, species extinction, for example, then the rational solution is to internalize the problem, and increase prices so as to "pay" for the damage. If the future is considered at all, it is at a discounted rate, which generates the possibility that resources can become rationally extinct, and that future human lives are worth less than present lives. In contrast, social economics rejects the instrumental value of human beings and asserts that all persons living today and in the future are of inestimable worth. Once more, the aim is to transform society into a place where social justice does not come second to economic efficiency, and where human dignity, now and in the future, can be paramount.

See Also the Following Articles

Behavioral Economics: The Carnegie School • Economics, Strategies in Social Science • Ethical Issues, Overview • Taxation • Utility

Further Reading

Etzioni, A. (1988). *The Moral Dimension: Towards a New Economy.* The Free Press, New York.

Hill, L. E. (1996). Economic history as a source of socio-economic normative value. In *Social Economics: Premises, Findings and Policies* (E. J. O'Boyle, ed.). Routledge, London.

Lutz, M. (ed.) (1989). *Social Economics: Retrospect and Prospect*. Kluwer Academic, Boston.

Nitsch, T. O. (1989). Social economics: The first 200 years. In *Social Economics: Retrospect and Prospect* (M. Lutz, ed.). Kluwer Academic, Boston.

O'Boyle, E. J. (ed.) (1996). *Social Economics: Premises, Findings and Policies*. Routledge, London.

Review of Social Economy (1993). The challenges facing social economists in the twenty-first century. *Rev. Soc. Econ.* **51**.

Stanfield, J. R. (1979). *Economic Thought and Social Change.* Southern Illinois University Press, Carbondale, IL.

Waters, W. R. (1993). A review of the troops: Social economics in the twentieth century. *Rev. Soc. Econ.* **51**(3), Winter.

Social Experiments, History of

Trudy Dehue
University of Groningen, Groningen, The Netherlands

Glossary

chance Lack of knowledge; whimsicality (until the late 19th century); characteristic of reality; something people can "take" (20th century).
determinism The view that the laws of human nature and society are pregiven.
liberal welfare state The version of the 20th century welfare state that emphasizes economic liberalism and is comparatively reluctant with regard to social services.
randomized controlled trial Ideal experiment in social science and medicine conducted with randomly composed experimental and control groups.
social experiment Event disturbing normal social order (19th century); social science research design (20th century).
statistical mean Measure of normalcy (19th century); also a measure of mediocrity (from the late 19th century).

Nineteenth century social researchers amply discussed the idea of social experiments. Moreover, the 19th century is known as the age of passionate measurement of social phenomena. Nevertheless, 19th century experts agreed that scientific experimentation with human beings is not feasible. It was not before the 1910s that the present-day definition emerged of a scientific experiment as comparative measurement of experimental and control groups. Also, it was not until the 1950s that the randomized controlled trial (RCT) became the ideal experiment in the social sciences (and medicine). This article compares the views of 19th century authors who denied the feasibility of active experimentation with those of 20th century social scientists to whom conducting RCTs became the ideal research strategy. Definitions of scientific social research, historians have demonstrated, form part of general belief patterns on society and politics. Understanding the 19th century pattern that excluded scientific social experiments helps to recognize the 20th century convictions legitimizing them.

Introduction: Social Experiments as Randomized Controlled Trials

The Basic Scheme

In science, the term experiment has always referred to widely varying research procedures. Historians of science have demonstrated that only loose definitions such as controlled observation cover all research designs that go for experiments. Further specifications of how precisely such general aims should be accomplished have led to a wide range of recommendations and practices.

However, in contemporary social science and psychology (as well as in medical research), a particular definition of the scientific experiment has won the day. In these disciplines, a truly scientific experiment entails comparing experimental groups that received a treatment with control groups that did not receive the treatment and, if a difference is found, calculating its statistical significance. Moreover, for the sake of statistical soundness and comparability, the groups must be composed on the basis of chance. To eliminate the possible influence of expectations concerning the outcomes, preferably both the participants and the conductors of an experiment are kept unaware of the group to which each participant has been assigned. Briefly, in these disciplines a truly scientific experiment is a randomized controlled trial (RCT; in medicine, a randomized clinical trial), and ideally it is a double-blind RCT.

Since the 1950s, social science handbooks and publication manuals have presented the RCT as the methodological standard for investigating causal relations. Many

social scientists self-evidently use the word experiment as shorthand for an RCT and consider research projects that diverge from its scheme as quasi-experiments at best. Although quasi-experiments are also conducted according to strict rules of design and statistics, social scientists regard their results as less valuable.

The Practice of RCTs

The scheme of the RCT only represents the basic design of an actual experiment. Each individual experiment has to add tailor-made supplements to the rough construction. For instance, often researchers compare several treatments and hence work with more than one experimental group. The architects of an experiment must also devise special rewards and incentives for attracting participants and assigning them randomly to the groups. Performing the RCT demands a high level of control over both the experimenters' and the participants' behavior. The designers of experiments must develop means of controlling the participants for a certain period of time and ensuring that neither the experimenters nor the participants diverge from the experimental script. The experimenters often need special training before an experiment begins. They have to learn how to work with sophisticated instruments for recording the participants' responses to the treatment. Cooperating statisticians have to manage the resulting data according to established prescriptions. Finally, research reports must be composed according to elaborate rules of publication. In addition to much methodological and professional expertise, social science experiments require large amounts of money. Nevertheless, RCTs are done on a large scale with numbers of participants that can exceed many thousands.

Social Experiments in the 19th Century

Early Definitions

The RCT is a fairly recent development. Before the 1910s, no expert on social research advanced the idea of comparing artificially composed experimental and control groups, and before the 1920s none of them proposed to compose groups on the basis of chance. However, the expression of experimentation appeared in much earlier texts on social research. The historian David Carrithers pointed out that in the 18th century eminent scholars discussed the question of experimentation as a suitable method for investigating humans and society. David Hume's "Treatise of Human Nature," first published in 1739, is subtitled "Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects." Hume and his Enlightment contemporaries, however, did not employ the terminology of experimentation in relation to deliberate actions by social researchers. They merely used it as a metaphor borrowed from natural science for indicating disturbing events occurring in human life. Hume, for instance, discussed "wars, intrigues, factions, and revolutions" as exemplary "social experiments." Observing such exceptional situations, 18th century authors argued, is the human science substitute of the artificial laboratory experiment in natural science labs.

Nineteenth century views on social experimentation were largely, but not completely, similar to those of the 18th century. Although the 19th century is known as the age of "the rise of statistical thinking," "the avalanche of printed numbers," and "the politics of large numbers," no social researcher at the time associated social experimentation with measurement, statistics, or particular research designs. Like their 18th century predecessors, 19th century experts used the terminology of experimentation in relation to human calamities, such as revolutions, or natural disasters, such as avalanches and floods. However, they added slight innovations to the traditional views, which in the course of time would become crucial elements of social experiments in the present-day sense.

The first novelty was that apart from the acts of God or Nature, the acts of governments also became main examples. Authors such as Auguste Comte in France, Adolphe Quetelet in Belgium, and John Stuart Mill and George Cornewall Lewis in Britain extensively discussed social legislation as social experimentation. In this respect, their definitions were akin to the present-day one. The second novelty was that these 19th century scholars explicitly preserved the label of scientific experimentation for artificial experiments with active manipulation by researchers. In this respect too, their convictions were the same as those of 20th century scholars. However, 19th century spokesmen drew the reverse conclusion, namely that if scientific experiments require active manipulation by researchers, then social experiments cannot be scientific. They categorically excluded experimentation from the useable methods of research with humans. There were two main reasons why experimentation could not be a proper research strategy.

Nineteenth Century Objections against Scientific Experimentation

The first reason was of an ontological and epistemological nature. Nineteenth century scholars regarded people and societies as organic systems in which each part is closely related to all others, and in which every event or characteristic of human life is the result of numerous interrelated causes. Because scientific experimenters have to vary one precisely delineated cause, the experimental method is not appropriate. It was mainly for this

reason that Auguste Comte, in his 1842 "Cours de la Philosophie Positive," rejected the use of scientific experimentation and why John Stuart Mill did the same in his "System of Logic" published in 1843.

According to Mill, there was "a demonstrated impossibility of obtaining, in the investigations of the social science, the conditions required for the most conclusive form of inquiry by specific experience." He illustrated his view using an important issue of his time: "the operation of restrictive and prohibitory commercial legislation upon national wealth." Such matters cannot be decided, he argued, by comparative studies of two countries that only differ as to their commercial freedom:

> *If the two nations differ in this portion of their institutions, it is from some difference in their position, and thence in their apparent interests, or in some position or other of their opinions, habits, and tendencies; which opens a view of further differences without any assignable limit, capable of operating on their industrial prosperity, as well as on every other feature of their condition, in more ways than can be enumerated or imagined.*

The second reason why the experimental method was unsuitable was of a moral nature. George Cornewell Lewis extensively discussed the latter type of objection in his 1852 two-volume "Treatise on the Methods of Observation and Reasoning in Politics." Whereas Hume more than a century ago explicitly included the "experimental method" in the title of his book, Lewis no less explicitly excluded it from his title. Regarding experimentation, Lewis maintained, is "inapplicable to man as a sentient, and also as an intellectual and moral being." This is not, he added, "because man lies beyond the reach of our powers" but because experiments "could not be applied to him without destroying his life, or wounding his sensibility, or at least subjecting him to annoyance and restraint."

Explaining the Difference

How could the notions of organic structures and multiple causes be compelling to John Stuart Mill, whereas since the 1950s "independent variables" have commonly been isolated and tested in artificial experiments? How could George Cornewell Lewis demand prudence even to the level of not "annoying" people, whereas since the second half of the 20th century massive group experiments have scrupulously been done? In 1997, the historian of science, Robert Brown, published two articles on the difference between the 19th- and 20th century views on social experimentation. Brown considers it an "incoherence" that in the 19th century there was extensive talk of social actions actually being experiments, whereas conducting them in a scientific way was deemed technically impossible and morally unwarrantable. He ascribes this incoherence to methodological ignorance due to lack of motivation. Only after social research became involved with industrial management, Brown argues, could the proper methods for social experimentation develop.

Brown offers an impressive overview of 19th century writings on social experimentation, and he raises the important question of how to explain the difference between the earlier and the later views. His answer, however, violates two interrelated principles of the historiography of science. The first one is the rule that "finalism" should be avoided—that is, that older views should not be judged from later standards of science. This rule implies that allegations such as lacking methodological expertise and zeal can be acceptable explanations only if they are justifiable in terms of contemporary norms. It is unlikely that a distinguished methodologist such as Mill, who repeatedly acknowledged the condition of comparison in science, rejected experimental comparison in human science out of ignorance. In addition, the punctilious 19th century deliberations on the feasibility of scientific social experimentation amply testify to intense motivation. Therefore, the question is open as to why notions of organic structures and fears of annoying people were once compelling reasons against artificial experiments, and why this is no longer the case. The second principle of the historiography of science is that explaining incompatible past and present ideas demands studying them as responses to other relevant beliefs of their times. The following sections apply both rules to the question at hand.

Nineteenth Century Views on Society, Politics, and Social Research

Keeping Order and Collecting Numbers

Until the end of the 19th century, determinism governed social and political thinking. The authoritative writers ascribed the facts of life to given laws of nature and society rather than to human intent and design. An essential part of the determinist philosophy was that the State could only have a very restricted role. The task of governments was mainly to safeguard the natural and normal order and not to generate permanent social change. Social legislation was directed largely at preventing the breakdown of social harmony and restoring it in cases of actual disruption. Even John Stuart Mill, who devoted much of his writings to the subject of governmental responsibility, held that central interference should be restricted to a very small range of affairs. Direct social action could at best create temporary relief and reestablishment of the natural order (Fig. 1).

Figure 1 Poverty and private charity in the 19th century. (Reprinted with permission from Elsevier Inc.)

Limited, however, as this goal may have been, in the course of the 19th century it became an increasingly demanding one. The industrialization and urbanization of Western societies resulted in the creation of large working classes that formed serious threats to social harmony. These people became the central concern of the authorities. In relation to them, governments were interested in numbers and hence employed statisticians. Theodore Porter in "The Rise of Statistical Thinking, 1820–1900," Ian Hacking in "The Taming of Chance," and Michael Cullen in "The Statistical Movement in Early Victorian Britain" vivaciously describe the strong urge of 19th century statisticians to quantify every aspect of the working classes. Statisticians not only abundantly quantified crime, misery, and suicide but also collected endless data on the number of people sleeping in one bed, the number of prints slum dwellers had on the wall, the number of wives who could knit, the number of husbands who mended furniture, and the number of parents who knew how to sing a cheerful song.

Cullen analyzes the 19th century British debates on public health reform and education as ways to counter the danger of criminality and revolutions and as means to instill "the unalterable nature of certain social relationships and hierarchies." In a history of statistics in 19th century America, Steven Kelman adds that the large-scale entrance of poor and uneducated immigrants strongly fueled the fear of social instability. Worries about instability, enhanced in the United States by the immigration problem, accounted for a spectacular extension of statistical data about crime, illiteracy, poverty, prostitution, bad housing, and ill health.

The Social Experiment Reconsidered

Experimental research in the later sense of assessing attempts at social change would have been an incoherence in the 19th century determinist belief pattern on social regularity, normalcy, and the largely peacekeeping role of the State. In times of laissez-faire, the word experiment could impossibly be more than a metaphor for indicating that careful observation of disturbances offers knowledge on the right and balanced state of affairs. Hence, the same Mill who is famous for working out the scientific method of difference argued that experimentation cannot be used in medicine and that "still less is this method applicable to a class of phenomena more complicated than even those of physiology, the phenomena of politics and history." The same Lewis who firmly objected to subjecting people to "annoyance and restraint" could simultaneously maintain that "a famine or a commercial crisis . . . has an elective affinity with the rotten parts of the social fabric, and dissolves them by the combination" and aloofly add that "the study of monstrosities, or malformations, in the animal or vegetable kingdoms, has likewise been recommended as a means of tracing the laws of organic structure."

In the 19th century pattern of beliefs, indeterminism or chance had a negative connotation of lack of knowledge and whimsicality. Using chance as a scientific instrument for drawing population samples, composing equal groups, or calculating statistical significance was therefore inconceivable. Even less imaginable was a notion of chance as something people must take in order to improve their lives. The pertinent issue, therefore, is how social experiments could change from disturbing events into instruments for assessing social progress and how indeterminism could change into something people must make good use of in both science and general life.

Turn-of-the-Century Changes

Changes in Statistics

If social misery remained the same, at least changes occurred in statistics. In the course of the century, statisticians began to distinguish lawlike regularities in their figures, which in the long term would instigate a different view of chance. Adolphe Quetelet in particular is known as the man who formulated the normal curve with the mean as a new kind of true measure of things that replaced the former notion of absolute laws. This was not, however, an indication that social phenomena were malleable. The mean gave expression to *l'homme moyen* who represented normalcy, whereas dispersion from the mean expressed aberration. Quetelet also distinguished natural forces that produce steady movement in the right direction from man-made "perturbational" counterforces. Nineteenth century statistical experts regarded the steady statistical

laws of suicides, crimes, and misery as further evidence that the State is largely powerless.

Historians of statistics generally point to the famous British statistician and biometrician Francis Galton as a crucial figure in the transition from determinism to probabilism. Building on Quetelet, Galton took an important step in the eventual conversion of "chance" into something that opens possibilities of progress instead of an indication of error. In relation to his eugenic ideals of improvement of the human race, Galton was primarily interested in the dispersions from the mean rather than the mean itself. To him, *l'homme moyen* did not represent the exemplar of the human race but the ordinary man who needs correction: "Some thorough-going democrats may look with complacency on a mob of mediocrities, but to most other persons they are the reverse of attractive" (quoted in Porter, 1986: p. 29).

Changing Views on Social Experiments

Galton was also interested in other means of establishing progress through science. In the *Fortnightly Review* of 1872, he published an article under the telling title of "Statistical Inquiries into the Efficacy of Prayer." Assuming that piety should be profitable, Galton maintained that the use of prayer can be assessed on its earthly revenues. The article is intriguing, however, for another reason. It is one of the earliest to recommend a comparative research strategy for evaluating actions. Galton stated,

> *The principles are broad and simple. We must gather cases for statistical comparison, in which the same object is keenly pursued by two classes similar in their physical but opposite in their spiritual state; the one class being spiritual, the other materialistic. Prudent pious people must be compared with prudent materialistic people and not with the imprudent nor the vicious.*

In summary, in 1872 Galton proposed to isolate a single causal variable and to establish its effects by comparing groups equal in all other relevant aspects. As he confidently added, "We simply look for the final result—whether those who pray attain their objects more frequently than those who do not pray, but who live in all other respects under similar conditions." Eight years later, Galton's countryman William Stanley Jevons, a reputed economist and statistician, suggested comparative measurement for assessing social laws. Jevons published an article in the 1880 *Contemporary Review*, titled "Experimental Legislation and the Drink Traffic," that discussed the example of the free trade of beer in shops. According to the author, the decision by the British government to legalize the selling of beer was "a salient example of bad legislation." The effects of the law should have been tested scientifically instead of passed "by the almost unanimous wisdom of Parliament." If the government would have commanded "a social experiment" first, Jevons maintained, it would have known beforehand that the beer law would only create "a beastly state of drunkenness among the working classes."

He maintained that it was possible to "experiment . . . provided we can find two objects which vary similarly; we then operate upon the one, and observe how it subsequently differs from the other." His meaning of a social experiment clearly differed from the traditional one. As far as is known, Jevons was the first author on social issues to argue that deliberate experiments should be conducted for the sake of administrative knowledge making.

Most important, apart from an advocate of active social experimentation, Jevons was also an early exponent of the late 19th century upper middle-class movement for limits on economic liberalism. In 1882, he published a book titled "The State in Relation to Labour." Although Jevons still ascribed people's success and misfortune to their own doing, he pleaded for some State intervention with regard to the supposedly fixed laws of economy. Such appeals for restrained government interference in the free-market economy provided the context in which the notion of scientific social experimentation emerged, and the rules of truly scientific experimentation gradually expanded.

Twentieth century Views on Social Politics, Statistics, and Social Research

Three Principles of Restricted Liberalism

Three mutually related principles of 20th century welfare capitalism were of pivotal importance in the emergence of social experiments as RCTs. The first one was that of individual responsibility. Social functioning and malfunctioning remained an individual affair. This implied that the degree of social care should be limited, but it also implied that ameliorative attempts were to be directed first and foremost at problematic individuals rather than at further structural change. Helping people largely meant treating, educating, punishing, or prizing individuals in order to turn them into self-supporting citizens.

The second principle was that of efficiency. Many contemporary commentators feared adverse consequences of state charity, and there was widespread distrust that administrations would squander public funds. The more hesitant a society was about State charity, the larger the fear of squandering public funds and the stronger the urge to search for singular causal factors of misery and backwardness. Ameliorative actions financed with public money had to produce instant results with economical one-dimensional means.

Figure 2 Cartoon expressing distrust in politicians.

The third principle was that of impersonal procedures. The stronger the fear of abuse of social services, the less the belief in people's own stories of needs and the more pressing the call for impersonal techniques establishing the truth behind that story. In addition, not only was the self-assessment of the interested recipients of help to be distrusted but also that of administrators providing help. Measurement also had to control administrators' claims of efficiency. As Theodore Porter expresses it in his 1995 book, "Trust in Numbers," the more resistance to centralized government in a society, the more officials have to warrant their decisions in terms of standardized or "mechanical," rather than interpretative or "disciplinary," objectivity. When Jevons ridiculed the "unanimous wisdom of Parliament" and argued that it should make way for scientific proof in the 1880s, he announced an era in which government officials increasingly had to base their authority on impersonal knowledge (Fig. 2).

It was no coincidence that these early developments took place particularly in the United States and Britain. Employing a classification introduced in 1990 by Gøsta Esping-Andersen, economists and sociologists label these countries as the prototypes of the "liberal" version of the 20th century welfare regime, which they distinguish from the "social democrat" and the "corporatist" regimes that provide more organized protection of the vulnerable. In his 1997 book titled "The Reluctant Welfare State: A History of American Social Welfare Policies," Bruce Jansson concludes that of all capitalist welfare societies the United States still spends the least amount of money on social welfare but the most on research of its effects.

From the early 20th century, administrative officials turned for help to academic experts who, on their turn, adapted their questions and approaches to the new demands. Experts on psychological, sociological, political, and economic matters also began to organize their work according to the three principles of welfare capitalism: they concentrated on help directed at individuals, the efficiency of such attempts, and impersonal procedures for assessing the attempts.

Developing the RCT

Early 20th century social scientists established a closer alliance with statisticians, who also adjusted to the social changes and increasingly adopted the teachings of Francis Galton. Statisticians now generally regarded deviations from the mean as an indication of real population differences. In "The Politics of Large Numbers," Alain Desrosières explains how the change of focus from what binds people to what separates them induced the development of random sampling for drawing conclusions on entire populations. Desrosières defies the finalist view that 19th century researchers were mistaken not to select their subjects on the basis of chance. The aim of 19th century social surveys was not to collect data for the sake of administrative actions but to describe typical situations, such as destitution in workers' communities. Only in relation to attempts at social steering could 20th century statisticians turn Galton's earlier notion of the statistical identity of random population samples into the technique of random sampling for drawing conclusions on entire populations. Desrosières, Hacking, Porter, and other historians of statistics argue that probabilistic reasoning was worked out for calculating and controlling chance rather than regarding statistical laws as mere destiny.

Initially, 20th century social scientists focused on techniques for measuring social phenomena. Subsequently, some of them proposed to use the new measurement instruments for assessing differences before and after administrative actions. These social scientists also argued that investigating only one group leaves open too much space for personal discretion with regard to the real causes of effects.

In the 1910s, the sociologist Frederick Stuart Chapin discussed the option of comparing experimental and control groups. However, in a sense, Chapin was still a 19th century aristocrat; that is, he was a "patrician–technician," as the historian Robert Bannister suitably described him. Arguing that one should not withhold treatments from needy people just for the sake of research, Chapin added a 20th century version to the former moral objections against experimentation. Furthermore, he maintained that comparing groups would be impracticable because in real life there are no identical groups.

Educational psychologists introduced the solution of deliberately composing groups for the sake of experimentation. Psychologists had a long tradition of psychophysiological and psychical experiments using small artificial groups in laboratory settings. In his book, "Constructing the Subject," the historian Kurt Danziger describes how during the administrative turn of both government and

Figure 3 "Order is heaven's first law": an early 20th century classroom.

human science, many of these laboratory psychologists offered their services to educational administrations. In the school setting, it was both morally and practically possible to create experimental and control groups. Like volunteers in laboratories, the pupils and their teachers could easily be persuaded to cooperate (Fig. 3).

In 1997, Dehue analyzed how the new educational professionals adapted their former laboratory methods to the new demands. Educational psychologists in the 1910s worked out methods for ensuring that the groups were comparable. The earliest strategy was to solve the problem by "matching." Each child was subjected to preliminary tests on factors suspected of creating bias, and then each child's ratings were used to form groups with equal results. Matching, however, collided with two of the principles that the new social science shared with the U.S. welfare state. The technique was quite elaborate and hence inefficient, and, worse, it required personal imagination regarding the variables on which the groups should be equalized.

In the early 1920s, educational psychologists at Columbia University, renowned at the time for their rigorous turn to administrative research and quantification, introduced an alternative: if chance determines the assignment of the children to the groups, each systematic difference will be automatically cancelled out. Then no personal discretion is needed with regard to the factors to be controlled, and, most important, random assignment is much more economical. Thus, these psychologists introduced the RCT in its basic form.

The RCT perfectly epitomized the values of 20th century liberalism that interventions should be directed at individuals in need of integration, instantaneous effectiveness of ameliorative interventions should be unambiguously demonstrated, and that this should be done via impersonal procedures. The assumptions of individual responsibility, efficiency, and impersonal procedures also ensured the validity of the RCT. The maxim that people are basically self-ruling rather than the product of extraindividual social processes guaranteed that it was not an epistemological problem to take them out of their natural groups and randomly rearrange them in artificial groups without social cohesion. The importance of efficiency justified focusing on a single isolated "independent variable" rather than on social and historical patterns. The belief that the discretion of responsible politicians, the subjects, or their physicians, psychologists, families, and friends should be discarded rather than taken into account justified assessment on the basis of preestablished standardized procedures only.

Ronald Fisher's 1935 book, "The Design of Experiments," prescribed random assignment as a condition to valid application of analysis of variance. This work became highly influential because it provided extra ammunition to advocates of impersonal judgment in the social sciences (and, as Harry Marks describes, for the same reason in medical research). As, over the course of time, 19th century laissez-faire capitalism was replaced by 20th century welfare capitalism, it became morally acceptable to experiment with adults too. The vast literature on the human sciences since the 1940s demonstrates that, particularly in the United States, numerous RCTs were done with students, soldiers, patients, slum dwellers, criminal offenders, drug abusers, incompetent parents, spouse beaters, people on welfare, and various other groups who were selected for an intervention and were comparatively compliant to the experimenters' regime.

See Also the Following Articles

Basic vs. Applied Social Science Research • Chapin, Francis Stuart • Fisher, Sir Ronald • Galton, Sir Francis • Jevons, William Stanley • Randomization

Further Reading

Bannister, R. C. (1987). *Sociology and Scientism: The American Quest for Objectivity, 1880–1940.* University of North Carolina Press, Chapel Hill.

Boruch, R. (1997). *Randomized Experiments for Planning and Evaluation.* Sage, London.

Brown, R. (1997a). Artificial experiments on society: Comte, G. C. Lewis and Mill. *J. Historical Sociol.* **10,** 74–97.

Brown, R. (1997b). The delayed birth of social experiments. *History Hum. Sci.* **10,** 1–23.

Carrithers, D. (1995). The enlightment science of society. In *Inventing Human Science. Eighteenth century Domains* (C. Fox, R. Porter, and R. Wokler, eds.), pp. 232–270. University of California Press, Berkeley.

Cullen, M. J. (1975). *The Statistical Movement in Early Victorian Britain; The Foundations of Empirical Social Research.* Harvester, New York.

Danziger, K. (1990). *Constructing the Subject.* Cambridge University Press, New York.

Dehue, T. (1997). Deception, efficiency, and random groups: Psychology and the gradual origination of the random group design. *Isis* **88,** 653–673.

Dehue, T. (2001a). Establishing the experimenting society. *Am. J. Psychol.* **114,** 283–302.

Dehue, T. (2001b). Comparing random groups. The history of experimentation in psychology. In *International Encyclopedia of the Social and Behavioral Sciences* (N. J. Smelser and P. B. Baltes, eds.), Vol. 8, pp. 5115–5120. Pergamon, Oxford, UK.

Desrosières, A. (1998). *The Politics of Large Numbers. A History of Statistical Reasoning.* Harvard University Press, Cambridge, MA.

Gooding, D., Pinch, T., and Schaffer, S. (eds.) (1989). *The Uses of Experiment: Studies in the Natural Sciences.* Cambridge University Press, Cambridge, UK.

Hacking, I. (1990). *The Taming of Chance.* Cambridge University Press, New York.

Kelman, S. (1987). The political foundations of American statistical policy. In *The Politics of Numbers* (W. Alonso and P. Starr, eds.), pp. 275–302. Russell Sage Foundation, New York.

Marks, H. M. (1997). *The Progress of Experiment. Science and Therapeutic Reform in the United States, 1900–1990.* Cambridge University Press, New York.

Porter, T. M. (1986). *The Rise of Statistical Thinking, 1820–1900.* Princeton University Press, Princeton, NJ.

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton University Press, Princeton, NJ.

Social Interaction Models

Steven Durlauf
University of Wisconsin, Madison, Wisconsin, USA

Ethan Cohen-Cole
University of Wisconsin, Madison, Wisconsin, USA

Glossary

complementarity A property of individual preferences or payoffs in which higher levels of some choice variable by others imply that higher levels of the choice variable are relatively more attractive to the individual.
contextual effects Effects determined by characteristics of an agent's neighborhood.
endogenous effects Effects determined by neighborhood members' contemporaneous behavior.
neighborhood Defined abstractly to be any group of individuals that could be considered to have a definable impact on an individual.
peer group effects An influence on an agent's behavior due to the connection to or perception of a peer group.
phase transition A model exhibits phase transition if its properties qualitatively change for a small change in a parameter value. Phase transitions are thus a way of describing when threshold effects occur in an environment.
poverty traps A situation in which the incentives for increased income or wealth are offset by other effects such that a group of agents remain poor over long time periods.
role model effect An influence on an agent's behavior due to the connection to or perception of a role model.
social multiplier The notion that connections and interactions between individuals can amplify or reinforce direct influences on agents.

Social interactions models comprise a body of recent work by economists and other social scientists that attempts to analyze formally the interplay between individual decisions and social processes. Substantively, these models attempt to answer two broad classes of questions. First, how do the characteristics and choices of others affect an individual's decision making? Second, how are these social influences reflected in equilibrium behaviors observed in a group as a whole? Substantively, social interactions models extend the domain of economic reasoning by evaluating direct interdependences between individuals. As such, these models complement the traditional economic focus on individual interdependences that are mediated via prices. Social interactions refer to these direct interdependences. During the past 15 years, there has been a renaissance of interest among economists in the social determinants of individual behavior and aggregate outcomes. A key reason for this is the potential for social interactions to help explain outstanding social questions such as the prevalence of inner-city poverty. In this regard, there is now a large body of empirical studies that attempt to measure social interaction effects on individuals in the context of residential neighborhoods. Social interactions models place such studies in a firm theoretical context. Also, methodological advances have allowed the incorporation of such effects into traditional microeconomic models. The basic structure and implications of these models suggest that these tools may have general application across the social sciences. Drawing from this basis, this article presents an overview of social interactions models and their applications, including a discussion of econometric techniques and outstanding research questions.

Social Interactions: Theory

By describing how agents' choices depend on the actions and characteristics of others in a common neighborhood,

this article provides a basis for understanding how social factors combine with the market-based and individual-specific factors that are the basis of neoclassical economic reasoning.

Following Manski, who adopts this terminology from the sociology literature, one can think of an agent's interactions with his or her neighborhood as being composed of two factors: contextual and endogenous. The first refers to those factors that are group specific and based on characteristics of the group members. The second refers to how agents are affected by the contemporaneous behavioral choices of group members. These alternative factors are illustrated in the context of residential neighborhoods, which represent an important leading case in the social interactions literature. How do residential neighborhoods affect individual educational outcomes? One source is local public finance of schools. Such a mechanism links individual educational quality to the distribution of socioeconomic status among neighborhood families. Another mechanism is role model effects. In this case, a student's school effort may be influenced by the level of economic success he or she observes among adults in the neighborhood. Feedback from the socioeconomic status of adults in a community to student behavior is an example of a contextual effect. In general, contextual effects are not reflexive: Although a student is affected by the behaviors of adult role models, he or she does not influence those role models per se. This is most obviously the case when the student is affected by a past behavior of an adult.

In contrast, endogenous effects refer to direct interdependences in contemporary choices among members of a neighborhood. For example, one might argue that the educational effort of one student is influenced by the effort of his or her friends; this type of endogenous effect is also known as a peer group effect. Unlike contextual effects, endogenous effects such as peer effects are reflexive; one student's effort influences his or her friends just as he or she is influenced. This is what is meant by an endogenous effect. Notice that both contextual and endogenous effects are influences that are not directly adjudicated by prices—that is, there is no market to compensate an adult for providing a good role model or a student for providing desirable peer effects.

Policymakers have a particular interest in social interactions research in general, and endogenous effects in particular, in that they provide a mechanism for understanding two prominent issues in social science: poverty traps and social multipliers. To understand what is meant by poverty traps, suppose that the college attendance decision is strongly related to the percentage of graduates in the community. These connections in behaviors can lead to two communities with different levels of college graduates in the long term. The mechanism for this should be clear: High (low) attendance rates of one generation lead to high (low) rates for the next generation. Communities initially composed of poor (via lack of education) members will remain poor across time. This result can be explained with intertemporal social interactions (i.e., social interactions in which choices made at one time affect others in the future).

Another conception of poverty traps derives from peer group effects. When such effects are strong, the characteristics of individuals in the group are not unique determinants of the group's action; instead, dependence on history, reactions to common influences, etc. may determine which sort of average behavior actually transpires. The emphasis here is that strong contemporaneous dependences in behavior can generate multiple different self-reinforcing behaviors in groups. Within a given configuration of behaviors, each individual is acting "rationally" in the usual sense. Note that this does not suggest that each self-consistent configuration is equally desirable from the perspective of the members of the group. One can also interpret poverty traps as a socially undesirable collection of behaviors that are mutually reinforcing and consequently individually rational.

Social multipliers arise because social interactions can amplify the effects of individual incentives. For a policymaker, this means that alterations of private incentives across a group may have far larger per capita effects than that associated with one individual in isolation. Consider the impact of providing tertiary education scholarships to randomly chosen students across various high schools versus concentrating the funds among students within a given school. If the goal is to alter high school graduation rates, then the presence of social interactions can, other things equal, mean that the concentration of the scholarships will be more efficacious. With the assumption that the direct incentive effect of the scholarships is equal for all students, the advantage of concentrating the scholarships in one school is that they will induce neighborhood effects for all students in the school, including those who have not been offered scholarships. Spreading the scholarships would have essentially no impact on any of the "neighborhoods" and would consequently impact only the students who received the funds. More generally, neighborhood effects can amplify the effects of altering private incentives; this amplification is what is meant by a social multiplier. To date, the implications for policy design of such multipliers have been little explored.

The basic implications of these effects suggest that the notion of interactions within neighborhoods may have general application in varied social science contexts. Various research agendas focus on populations of agents organized into groups in which some type of non-price-related interactions occur. Each of these utilizes, at least abstractly, the notion of neighborhood-specific social interactions. Applications range from economic growth and development to crime and land use patterns. This work

does not require that neighborhoods be defined geographically, but it does rely on some notion of proximity versus distance in "social space," a notion originally given content by Akerlof.

Formal Theory

This section summarizes the previously discussed concepts into a formal model. First, consider the abstract problem of how social interactions influence individual choices and thereby produce interesting neighborhood behaviors in the aggregate.

This model has I individuals part of a common neighborhood denoted n. Each individual i chooses ω_i from a set of possible behaviors $\mathbf{\Omega}_i$. This individual-level decision will produce a probabilistic description of the choice given certain features of the individual and his or her neighborhood. This model constructs a probability measure $\mu(\cdot)$ for the vector of choices of all members of the group, ω, that is consistent with these individual-level probability measures and relates how neighborhood effects determine its properties. To capture how others influence each agent, define ω_{-i} as the vector of choices made by individuals other than agent i.

Continuing from the previous theoretical discussion, one may distinguish between the various types of influences on individual behavior. In addition to the contextual and endogenous factors defined previously, there are also two types of individual-specific characteristics: deterministic and random. These four types of influences have implications on how to model the choice problem. For simplicity, the model labels the four as follows:

X_i, a vector of deterministic (to the modeler) individual-specific characteristics associated with individual i

ε_i, a vector of random individual-specific characteristics associated with i

Y_n, a vector of predetermined neighborhood-specific characteristics (these measure the contextual effects)

$\mu_i^e(\omega_{-i})$, the subjective beliefs individual i possesses about behaviors of others in his or her neighborhood, described as a probability measure over those behaviors (this term captures potential endogenous effects).

Each of these components will be treated as a distinct argument in the payoff function that determines individual choices. As discussed previously, the social interactions terms are the final two. Even though these may be "nonstandard" in the context of traditional economic decision problems in that they are not price driven, individual choices are still defined via the maximization of some individual payoff function $V(\cdot)$; given the notation introduced, individual choices are thus assumed to follow

$$\omega_i = \arg\max_{\omega \in \Omega_i} V(\omega_i, X_i, \varepsilon_i, Y_n, \mu_i^e(\omega_{-i})). \tag{1}$$

Next, to close this model, one must choose a method of resolving a standard problem in economics: how individuals form beliefs about the behaviors of others. The benchmark assumption is that beliefs are rational in the sense that

$$\mu_i^e(\omega_{-i}) = \mu(\omega_{-i} \mid \varepsilon_i, Y_n, X_j, \mu_j^e(\omega_{-j})\ \forall j), \tag{2}$$

where j refers to members of the neighborhood n other than agent i. The right-hand side of Eq. (2) is a conditional probability measure that describes how agent i would form beliefs that are mathematically consistent with the model, given the conditioning variables. In particular, one should note that the distinction between one's beliefs about the choices of others and their actual choices derives exclusively from the fact that agent i observes only his or her own random payoff term, ω_i. Finding an equilibrium set of behaviors in a neighborhood is thus a fixed-point problem—that is, determining what subjective conditional probabilities concerning the behavior of others correspond to the conditional probabilities produced by the model when behaviors are based on those subjective beliefs.

To make the model more tractable for analysis and interpretation, Eq. (1) is often simplified in two ways. First, since one might expect individuals to care about the average behavior of those in the group, endogenous effects can be expressed as $\bar{\omega}_{-i} = (I-1)^{-1}\sum_{j \neq i} \omega_j$. Second, one can also remove uncertainty by setting ε_i to zero for all neighborhood members; this allows expectations and realizations to coincide in Eq. (2). When these assumptions are made, individual decisions solve

$$\omega_i = \arg\max_{\omega \in \Omega_i} V(\omega_i, X_i, Y_n, \bar{\omega}_{-i}). \tag{3}$$

From the perspective of formal theory, the interesting properties of social interactions models depend on the direct interdependences that exist between individual choices—that is, the endogenous effects that are captured by the presence of $\mu_i^e(\omega_{-i})$ in Eq. (1) and $\bar{\omega}_{-i}$ in Eq. (3). Social interactions models typically assume that these interdependences between individual choices exhibit complementarity. Intuitively, complementarity means that the relative payoff of a higher value of ω_i versus a lower value is increasing in the levels chosen by others. For the payoff function described in Eq. (3), complementarity means that if $\omega^{\text{low}} < \omega^{\text{high}}$, and $\bar{\omega}_{-i}^{\text{low}} < \bar{\omega}_{-i}^{\text{high}}$, then

$$V\left(\omega^{\text{high}}, X_i, Y_n, \bar{\omega}_{-i}^{\text{high}}\right) - V\left(\omega^{\text{low}}, X_i, Y_n, \bar{\omega}_{-i}^{\text{high}}\right) > V\left(\omega^{\text{high}}, X_i, Y_n, \bar{\omega}_{-i}^{\text{low}}\right) - V\left(\omega^{\text{low}}, X_i, Y_n, \bar{\omega}_{-i}^{\text{low}}\right). \tag{4}$$

Complementarity is a fundamental property for interdependent decision making because it leads to similarity of behaviors as high choice levels by others make it more likely that an individual does the same; similar logic applies to low choice levels. In fact, the model in Eq. (3) is an example of the class of coordination models that arise in applied game theory.

What types of properties may be exhibited by social interactions models? One important property is that of multiple equilibria. A model such as shown in Eqs. (1–3) exhibits multiple equilibria when there is more than one set of choices ω such that each individual is making the choice that maximizes his or her payoff. Intuitively, when complementarities are strong enough, it permits individuals to behave similarly in equilibrium but does not specify or require particular behavior. This introduces a "degree of freedom" in the determination of outcomes as a whole. In the social interactions context, this is important because multiple equilibria create the possibility that two neighborhoods with similar observable characteristics (i.e., distributions of X_i within each neighborhood n and levels of Y_n) can exhibit different aggregate behaviors. When will multiple equilibria occur? Clearly, one factor is the strength of the endogenous social effects. If these effects are weak, then the other determinants of individual behavior play a relatively larger role in determining individual outcomes and can lead to unique equilibria.

A second property these models may exhibit is phase transition. A model exhibits phase transition if its properties qualitatively change for a small change in a parameter value. Phase transitions are thus a way of describing when threshold effects occur in an environment. Why do phase transitions occur in social interactions models? Intuitively, phase transition is related to the multiplicity versus uniqueness of equilibria. Social interaction models often have the property that for a given specification of individual and contextual effects, there is a threshold for the strength of endogenous social effects such that if the level of endogenous effects is above the threshold, multiple equilibria occur.

Brock and Durlauf provide an explicit analysis of how the strength of different factors that affect individual behavior jointly determine the number of equilibrium behaviors that may be observed at the group level for binary choice models with social interactions that illustrate these properties. In their framework, individuals choose $\omega_i \in \{-1, 1\}$. Social interactions are determined by the expected average choice level in the group, m_n. Specifically, the payoff function is such that

$$\begin{aligned} &V(1, X_i, Y_n, \varepsilon_i) - V(-1, X_i, Y_n, \varepsilon_i) \\ &\quad = k + cX_i + dY_n + Jm_n + \varepsilon_i, \end{aligned} \tag{5}$$

where k, c, d, and J are constants, and ε_i is a scalar that is independently and logistically distributed—that is, $F_\varepsilon(z) = 1/(1 + \exp(-z))$. Brock and Durlauf show that the equilibrium expected average choice for a neighborhood must fulfill

$$m_n = \int \tanh(k + cX + dY_n + Jm_n)\, dF_x, \tag{6}$$

where $(\tanh(x) = (\exp(x) - \exp(-x))/(\exp(x) + \exp(-x))$, and dF_x is the distribution of X within n. The number of equilibrium values of m_n that is consistent with Eq. (6) is determined by the value of J, holding other factors constant. For example, if each member of n is associated with the same individual effects (i.e., $X_i = \bar{X}$), then the following results holds: For each value of $k + c\bar{X} + dY_n$, there exists a threshold J^{Thresh} (which depends on $k + c\bar{X} + dY_n$) such that if $J < J^{\text{Thresh}}$, then there is only one equilibrium, whereas if $J > J^{\text{Thresh}}$, then three different values of m_n are consistent with Eq. (6).

To date, far less effort has been dedicated to analysis of econometric issues in social interactions topics than to theoretical work. Brock and Durlauf, Manski, and Moffitt provide general treatments. These studies provide a number of important results for conducting and interpreting empirical work. The following sections review the main econometric issues that arise.

Identification

To provide an illustration of the basic identification issues in social interactions research, consider that $V(\omega_i, X_i, \varepsilon_i, Y_n, \mu_i^e(\omega_{-i}))$ produces a linear representation of individual choice; that is, choices obey the following basic regression specification:

$$\omega_i = k + cX_i + dY_{n(i)} + Jm_{n(i)} + \varepsilon_i, \tag{7}$$

where X_i denotes an r-length vector of observable individual characteristics, $Y_{n(i)}$ denotes an s-length vector of contextual effects, $m_{n(i)}$ denotes the expected value of ω_i for members of neighborhood $n(i)$, and $n(i)$ denotes the neighborhood of individual i [which allows Eq. (7) to describe individuals from different neighborhoods]. This model, referred to as the linear-in-means model, was first studied by Manski in 1993. The linearity assumption facilitates interpretation as in a linear model. To put this in the context of the previous discussion, note that all endogenous effects here work solely through expectations—an assumption that might be appropriate if the neighborhoods were particularly large. To highlight one of the key econometrics issues, we first focus on the case in which $E(\varepsilon_i \mid X_i, Y_{n(i)}, i \in n(i)) = 0$, such that identification questions are intrinsic to neighborhood effects rather than the endogeneity of the neighborhoods themselves.

To understand why identification conditions arise in this model, observe that when beliefs are rational,

$$m_{n(i)} = \frac{k + cX_{n(i)} + dY_{n(i)}}{1-J}, \quad (8)$$

where $X_{n(i)}$ equals the average of the X_i's in neighborhood $n(i)$ and appears in the regression because this average is one of the determinants of $m_{n(i)}$. Substituting Eq. (8) into Eq. (7), the individual choices may be expressed in terms of observables via

$$\omega_i = \frac{k}{1-J} + cX_i + \frac{J}{1-J}\, cX_{n(i)} + \frac{d}{1-J} Y_{n(i)} + \varepsilon_i. \quad (9)$$

Equation (9) summarizes the empirical implications of the linear-in-means model. The identification problem may thus be thought of as asking whether one can recover the structural parameters in Eq. (7) from the coefficients in Eq. (9).

To complete this analysis, one need only compare the number of regressors of Eq. (9) with the number of coefficients of Eq. (7). One can see that Eq. (9) contains $2r+s+1$ regressors, whereas there are only $r+s+2$ coefficients in Eq. (7). Although it might appear that one could recover the structural parameters from a regression of ω_i onto the various regressors, in fact the parameters of Eq. (7) are overidentified. However, this conclusion fails to account for possible collinearity between the components of Eq. (9); collinearity may potentially arise because of the presence of $X_{n(i)}$ and $Y_{n(i)}$ in the equation. For example, following the case originally studied by Manski, suppose that $X_{n(i)} = Y_{n(i)}$. In this case, the researcher would be unable to distinguish between contextual and individual effects. When this condition holds and there are only $r+s+1$ linearly independent regressors in Eq. (9), the associated coefficients for these linearly independent regressors are identified, but they cannot be uniquely mapped back into the $r+s+2$ structural coefficients in Eq. (7); identification of the structural parameters in Eq. (7) thus fails. Manski termed this failure of identification the reflection problem to capture the intuition that the identification problem relates to distinguishing the direct effect of $Y_{n(i)}$ on an individual versus its indirect effect as "reflected" through the endogenous effect generated by $m_{n(i)}$.

When does the identification problem preclude identification of the parameters of Eq. (7)? The key requirement for identification is that the vector $X_{n(i)}$ is linearly independent of the other regressors in Eq. (7), (1, X_i, $Y_{n(i)}$). For this to be so, a necessary condition is that there exists at least one element of X_i whose group-level average does not appear in $Y_{n(i)}$. Intuitively, one needs prior information that at least certain individual-level effects are present whose group-level analogs do not affect individuals.

It is important to recognize that the identification problem previously discussed is a product of the linear specification. Identification breaks down when $m_{n(i)}$ is linearly dependent on the other regressors in Eq. (7); linear dependence of this type will typically not arise when individual behaviors depend on other moments of the neighborhood behavior. More important for empirical work, this argument also implies that identification will hold for nonlinear probability models of choices. For example, the binary choice model of Brock and Durlauf is identified under weak assumptions.

Self-Selection

The discussion of identification has not addressed the issue of self-selection into groups. For contexts such as residential neighborhoods, self-selection is of course important. In fact, recent theories of neighborhood composition are driven by the presence of social interactions. From the perspective of our discussion, self-selection implies that $E(\varepsilon_i \mid X_i, Y_{n(i)}, i \in n(i)) \neq 0$.

There does not exist any general solution to the analysis of social interactions with self-selection. One approach, followed by Evans *et al.*, is to use instrumental variables to account for $E(\varepsilon_i \mid X_i, Y_{n(i)}, i \in n(i)) \neq 0$. An alternative approach, developed by Brock and Durlauf, models the self-selection explicitly. We focus on the linear-in-means model. Consider the following equation to illustrate the effect of self-selection on identification:

$$\omega_i = k + cX_i + dY_{n(i)} + Jm_{n(i)} + E(\varepsilon_i \mid X_i, Y_{n(i)}, i \in n(i)) + \xi_i, \quad (10)$$

where $E(\xi_i \mid X_i, Y_{n(i)}, i \in n(i)) = 0$ by construction. Following the classic approach to selection developed by James Heckman, consistent estimation of Eq. (10) requires constructing a consistent estimate that is proportional to $E(\varepsilon_i \mid X_i, Y_{n(i)}, i \in n(i))$, call it $\delta(X_i, Y_{n(i)}, i \in n(i))$, and including this estimate as an additional regressor in Eq. (10); that is, one in essence estimates the regression

$$\omega_i = k + cX_i + dY_{n(i)} + Jm_{n(i)} + e\delta(X_i, Y_{n(i)}, i \in n(i)) + \xi_i. \quad (11)$$

The key insight of Heckman is that once this is done, Eq. (11) may be estimated by ordinary least squares. Brock and Durlauf describe how to implement this procedure in the social interactions case using two-stage methods.

Self-selection corrections have important implications for identification. Consider two cases. First, suppose that the decision to join a neighborhood depends only on $m_{n(i)}$—that is, $\delta(X_i, Y_{n(i)}, i \in n(i)) = \delta(m_{n(i)})$. In this case, Eq. (11) is now nonlinear in $m_{n(i)}$ (since $\delta(\cdot)$ is almost certainly nonlinear given the fact that the neighborhood choice decision is made among a set of discrete

alternatives) and is thus identified outside of pathological cases. Alternatively, in general $\delta(X_i, Y_{n(i)})$ will be linearly independent of $(1, X_i, Y_{n(i)})$ since $\delta(\cdot)$ is nonlinear. As such, $\delta(X_i, Y_{n(i)})$ is an additional individual-level regressor whose group-level analog does not appear in the behavioral equation (Eq. 7). Thus, identification may be achieved.

This approach to self-selection may be criticized to the extent that the self-selection correction is constructed on the basis of parametric assumptions concerning the distribution of the various model errors. We regard this as a legitimate but not critical caveat. The analysis by Brock and Durlauf that we have described should be interpreted as demonstrating that self-selection not only does not make identification of social interactions impossible but also may, if appropriately modeled, facilitate identification. This facilitation follows from the fact that neighborhood choices embody information on how individuals assess social interactions.

An equally important new direction is the development of data sets that will facilitate more detailed analyses of social interactions. One important development is the Moving to Opportunity Demonstration being conducted by the Department of Housing and Urban Development that involves creating incentives for poor families to move to more affluent neighborhoods in order to determine how they are affected; Katz *et al.* provide valuable evidence on social interaction effects. Other efforts are promising in terms of the detailed data that are being obtained. The Project on Human Development in Chicago Neighborhoods is noteworthy for the detailed information on attitudes and outcomes that is being compiled.

See Also the Following Articles

Communication • Social Psychology • Urban Studies

Further Reading

Akerlof, G. (1997). Social distance and economic decisions. *Econometrica* **65**(5), 1005–1027.

Brock, W., and Durlauf, S. (2001a). Discrete choice with social interactions. *Rev. Econ. Stud.* **68**(2), 235–260.

Brock, W., and Durlauf, S. (2001b). Interactions-based models. In *Handbook of Econometrics* (J. Heckman and E. Leamer, eds.), Vol. 5. North-Holland, Amsterdam.

Brock, W., and Durlauf, S. (2003). A multinomial choice model with social interactions. In *The Economy as an Evolving Complex System III* (L. Blume and S. Durlauf, eds.). Oxford University Press, New York.

Cooper, R., and John, A. (1988). Coordinating coordination failures in Keynesian models. *Q. J. Econ.* **103**(3), 441–463.

Durlauf, S. (2003). Neighborhood effects. In *Handbook of Regional and Urban Economics* (J. V. Henderson and J.-F. Thisse, eds.), Vol. 4. North-Holland, Amsterdam.

Evans, W., Oates, W., and Schwab, R. (1992). Measuring peer group effects: A study of teenage behavior. *J. Polit. Econ.* **100**(5), 966–991.

Glaeser, E., Sacerdote, B., and Scheinkman, J. (1996). Crime and social interactions. *Q. J. Econ.* **111**(2), 507–548.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* **47**(1), 153–161.

Jencks, C., and Mayer, S. (1990). The social consequences of growing up in a poor neighborhood. In *Inner-City Poverty in the United States* (L. Lynn and M. McGreary, eds.). National Academy Press, Washington, DC.

Katz, L., Kling, J., and Liebman, J. (2001). Moving to opportunity in Boston: Early results of a randomized mobility experiment. *Q. J. Econ.* **116**(2), 607–654.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.* **60**(3), 531–542.

Moffitt, R. (2001). Policy interventions, low-level equilibria, and social interactions. In *Social Dynamics* (S. Durlauf and H. P. Young, eds.). MIT Press, Cambridge, MA.

Sampson, R., Morenoff, J., and Earls, F. (1999). Beyond social capital: Collective efficacy for children. *Am. Sociol. Rev.* **64,** 633–660.

Social Measures of Firm Value

Violina P. Rindova
University of Maryland, College Park, Maryland, USA

Glossary

audits Research investigations conducted by specialized organizations (information intermediaries) with regard to firm performance on preselected criteria relevant to different stakeholder groups.
awards Forms of recognition for outstanding achievement in some area of firm activities, often accompanied by a monetary prize.
comparative reputational rankings Comparative orderings of organizations that compete in an industry or an organizational field.
corporate contests Direct comparisons of the performance of firms or their products, usually made by a panel of judges.
information intermediaries Organizations that monitor and certify firm performance by collecting and disseminating information about firms, and constructing social measures of firm performance.
opinion polls Surveys of stakeholder perceptions of firm performance.
stakeholders Single individuals or groups of individuals (employees, customers, suppliers, shareholders, and others) that are affected by and affect the operations of a firm.

Social measures of firm value are explicit or implicit quantitative or qualitative evaluations of the performance and ability of firms to create value along various dimensions relevant to the firms' diverse stakeholders. Valuations may be produced through audits, opinion polls, and contests. These processes can then be used to yield comparative reputational rankings and expert categorizations and ratings.

Origins and Role of Social Measures of Firm Value in Market Exchanges

Stakeholders in a business organization control the resources that are vital for the economic success of the firm. The decisions to make these resources available to a given firm depend on the evaluations that stakeholders make about the ability of the firm to create value for them. Stakeholders are single individuals as well as groups of individuals, including employees, customers, suppliers, shareholders, and others, that are affected by and affect the operations of a firm. In choosing the extent to which they will exchange their resources with the firm, stakeholders, in turn, influence the ability of the firm to create value and achieve desired economic results. However, in order to select firms with which to exchange their limited resources, stakeholders must overcome the classical market selection problem, which relates to the concepts of adverse selection and moral hazard. Moral hazard refers to the incentives of sellers to exert less than complete effort in producing and delivering products and services, and adverse selection refers to the likelihood that buyers will pick a lower quality product or service. These problems arise from the fact that sellers of products and services have more information than buyers do about the quality of the marketed products and services. This asymmetry between the information available to a particular market player (private information) and the information available to others (public information) gives rise to market selection problems. Information asymmetries and market selection problems are reduced when firms submit to certification and monitoring by outside agents.

In order to mitigate these information problems in markets, various organizations have emerged to serve as information intermediaries. Such entities specialize in collecting and disseminating information about firms. Information intermediaries may be involved in making private information public, as in the case of financial analysts, who engage business managers in discussions regarding their firm's operations. Through such discussions, financial analysts receive and make available information that supplements the information a firm has made publicly available in its financial statements. Information intermediaries can also be involved in making otherwise publicly available information more accessible and comprehensible, by issuing summary statistics and reports. This function of information intermediaries is also important to the market exchange process because, even if information about firms is publicly available, it may not be readily accessible to all market players; information intermediaries therefore potentially decrease the market players' search costs and increase the value they may derive from a particular exchange.

Information intermediaries may also "create" new information about firms by providing stakeholders with third-party evaluations of firms, in the forms of ratings, rankings, classifications, or awards. This activity of information intermediaries produces various social measures of firm value. Such measures not only facilitate stakeholders' information gathering, but also directly aid in their decision processes. The reason for this is that social measures of firm value provided by information intermediaries often carry institutional validity, because information intermediaries, in the words of management scholar Hayagreeva Rao, are "social control specialists who institutionalize distrust of agents by inspecting their performance ... and who sustain social order." Social measures of firm value therefore are outputs of organizations that specialize in the collection, dissemination, and analysis of information about firms, in an effort to reduce the market selection problem that stakeholders face.

Financial versus Social Measures of Firm Value

Financial measures of the value of a firm are quantifications of the economic value of the assets a firm controls and the expected ability of a firm to deploy these assets in generating economic returns. Whereas a systematic overview of financial measures of firm value lies outside the scope of this article, the brief overview provided here serves as a basis for distinguishing between financial and social measures of firm value.

According to finance theory, the value of a firm is the sum of the value of all of its assets. Different definitions of firm value exist and are appropriate for different situations. For example, the liquidating value of the firm can be realized if its assets are sold separately from the organization that has been using them. In contrast, the going-concern value of the firm is its value as an operating business. Further, the accounting value of the firms' assets determines its book value, while the value at which these assets can be sold determines its market value. Because the firm has both a liquidating value and a going-concern value, the higher of the two constitutes the firm's market value. The market value of the firm as a going concern is calculated as sum of the current and projected cash flows of the firm, discounted at the firm's average cost of capital. Whereas the estimation of future cash flows and the cost of capital may involve various levels of elaboration in analyses, the basic logic of financial measures of firm value is that a firm's value reflects its potential to generate economic returns for its suppliers of capital. This logic is based on the assumption that the suppliers of capital are the residual claimants, who receive their returns after all other claimants (stakeholders) have been satisfied.

Social measures of firm value differ from the financial approaches in several ways. Financial approaches are concerned primarily with evaluation of the economic assets of the firm and their productive capacity to generate economic returns for the suppliers of capital. In contrast, social approaches to firm valuation view the firm as a social agent with whom individuals and other social entities form relationships and pursue a wide range of economic and noneconomic goals, such as quality of life, social justice, professional development, and safe and clean environments. Thus, social measures of firm value are concerned with the role of the firm as a social actor, the activities of which impact a variety of stakeholders, rather than a single stakeholder group. Consequently, social measures of firm value differ from financial measures in their concern with both economic and noneconomic outcomes associated with the operations of a firm, and with the effects of these operations on various stakeholders. It should be noted, however, that financial measures of firm value are a type of social measure as well. As Ralf Hybles has explained, "The schemas of finance and accounting present an overarching set of abstract categories and decision rules that coordinate across specialties and organizations" so that "the dual functions of finance and accounting together certify the economic legitimacy of every type of contemporary organization."

Whereas financial measures are always explicit and quantitative, social measures take a variety of forms, i.e., explicit and implicit and quantitative and qualitative. Because they are quantitative and standardized, financial measures enable a high degree of comparability of firm value in dollar terms. However, they offer limited means for evaluating aspects of firms performance that are not readily measured in dollars, such as intangible assets,

employee satisfaction, and innovativeness of research and development efforts. Social measures seek to capture these aspects of firm performance and do so through both quantitative and qualitative and explicit and implicit means. Examples of explicit quantitative social measures of firm value are various reputational rankings and ratings, such as those published by *Fortune Magazine* in its "Most Admired Corporations" surveys. Examples of explicit qualitative measures are various awards, such as the annual awards presented by *R&D Magazine* to companies making technologically significant products. Implicit quantitative measures of firm value are metrics that assess specific organizational practices, rather than the overall performance of the firm and its products. For example, the Investor Responsibility Research Center (IRRC), a not-for-profit organization, offers research for portfolio screening that focuses on firm practices related to waste generation and disposal of toxic and hazardous materials, and the environmental liabilities of a firm. Examples of implicit qualitative measures are various classification schemes that categorize competing firms in an industry, in terms such as "specialty" versus "mass" or "low-end" versus "high-end" producers. As these examples indicate, social measures of firm value can take a variety of forms, including rankings, ratings, awards, and expert and lay categorizations. Because of this diversity, social measures of firm value may be less readily identifiable and available to the general public, compared to financial measures. Further, the diversity of the forms of social measures also suggests that the term "measures" may not capture precisely the nature and forms of social evaluations of firm value. Such evaluations may be better understood as summary representations of business performance that contain explicit or implicit evaluative elements. Their importance in markets and their value to stakeholders derives from the fact that they summarize, and often institutionally "certify," information about competing firms and thereby reduce stakeholder uncertainty with regard to exchanges with these firms. Therefore, the processes through which information intermediaries collect and process information about firms to generate such evaluative representations are an integral part of the validity and usefulness of such evaluations. The processes through which social measures of firm value are constructed are discussed next.

Constructing Social Measures of Firm Value

Extant research on social evaluations of firms can be organized in three main groups, based on the processes through which the evaluation takes place. In practice, these three processes are often combined, and evaluations of firms may contain elements of all of them. However, it is useful to distinguish among the processes in order to understand the variety of social measures of firm value. The three types of processes through which social measures of firm value are produced can be broadly described as expert audits, opinion polls, and contests.

Audits

Audits are research investigations about practices and outcomes based on various standardized criteria that are used to produce comprehensive evaluations. Different information intermediaries specialize in auditing firm performance on dimensions deemed of interest to different stakeholder groups. For example, Rao documented the emergence of consumer watchdog organizations in the United States; the formation of the first nonprofit consumer watchdog organization, Consumers' Research, was in 1927, and by 1995, the number of such organizations had increased to 200. If consumer watchdog organizations that operate for profit, such as Morningstar and Lipper Analytical Services, are included this number is even higher. Other examples of organizations dedicated to conducting audits of firm activities and providing stakeholders with evaluations of these activities include those dedicated to monitoring and reporting on the socially responsible and environmentally conscious behaviors of firms. For example, The Council on Economic Priorities (CEP) is a not-for-profit organization founded with the goal to provide analysis of the social and environmental records of corporations. It has published a book called *Rating America's Conscience*, which rates 130 consumer product companies on criteria such as hiring records, charitable contributions, involvement in South Africa, and defense contracts. The CEP has also published a shopping guide, *Shopping for a Better World*, which rates 168 companies and over 1800 products, and an investment guide, *The Better World Investment Guide.*

The audit process is characterized by reliance on specialized staff; its goal is to collect and process information about firms through a combination of quantitative and qualitative methods, and to produce standardized metrics, ratings, or rankings. The audit process places an emphasis on the extensiveness of data gathering, triangulation of information, and standardization of information outputs to enable systematic comparisons among competing firms. Yet, the audit process may include relative criteria, e.g., how a firm performs relative to its industry peers, as well as data on absolute standards, such as government-mandated or industry-sponsored standards and guidelines. For example, information intermediaries screening firms for environmental performance tend to rate and rank firms for their performance on the Toxic Release Inventory (TRI), which is mandated by the Environmental Protection Agency (EPA).

Whereas organizations that employ an audit process to produce social measures of firm value seek to demonstrate the social rationality of their criteria and the rigor of their methods, they tend to approach firm evaluations from a particular ideology, which guides the selection of criteria on which firms are evaluated, as well as the allocation of resources in collecting and processing information about firms. Also, despite their use of statistical and other scientific methods in processing the data, these organizations are subject to a very limited degree of social control with respect to the validity and the reliability of the measures they generate. For example, *US News & World Report* publishes annual rankings of the top business schools in the United States. These rankings are based on information concerning input characteristics (e.g., for a student earning a Master's degree in business, the average Graduate Management Admission Test score, and the grade-point average of the entering class), throughput characteristics (such as teaching methodologies used), and output characteristics (such as percentage of students with job offers at graduation, and average starting salaries of graduates). The aggregation of such diverse data increases the noise in rankings, thereby reducing their validity as measures of underlying quality. Further, these rankings are based on self-reported information by schools, the veracity of which has come into question. Together, these arguments suggest that the validity of social measures of firm value that are based on audits can improve significantly if well-developed auditing and data analytical techniques in the field of accounting and social sciences, in general, are deployed more systematically. Further, to the degree that multiple information intermediaries undertake audits with similar purposes, they will be subject to the discipline of the market.

Opinion Polls

Opinion polls do not strive to examine and assess the practices of firms and their relative performance. Instead, they seek to capture stakeholder perceptions of firm performance. The rationale underlying opinion polls is to capture and make explicit the collective consensus about a firm's performance and abilities to create value for various stakeholders. Social measures of firm value derived from opinion polls are important in market exchanges because individual stakeholders are ultimately the ones who form expectations and who make buying and selling decisions. Further, whereas the audits of expert organizations tend to espouse a particular ideology, which influences both the criteria of the audit and the final evaluations derived from the data analysis, opinion polls can capture the diversity of stakeholders beliefs and evaluations of firms. For example, whereas some investors are socially responsible, others are not; whereas some customers are "green," others are "price-shoppers"; and whereas some employees consider the pro-bono activities of a firm an important part of its identity, others focus only on personal benefits. Therefore, the construction of social measures of firm value that account for the diversity of a firm's stakeholders is an important issue for both research and practice.

Though researchers agree about the importance and value of opinion polls as an approach to capturing stakeholder perceptions of firms, they disagree about the appropriate methodology for doing so. Research on "public opinion" has long wrestled with the problem of how the viewpoints of the general public can be captured and observed. Currently, public opinion research efforts have adopted the perspective that the public point of view is best captured through an aggregation of individual opinions. Anchored in a pluralistic worldview, this perspective adheres to the democratic principle of "one person, one vote," advocating systematic random polling as a way to unearth the opinions of a diverse polity. The critical concern when employing this approach is to create samples that are "representative" of the population at large. As Charles Roll and A. Hadley Cantril have explained, "Respondents ... are not selected because of their typicality or of their representativeness. Rather, each sampling area and each individual falls into the sample by chance and thus contributes certain uniqueness to the whole. It is only when these unrepresentative elements are added together that the sample should become representative." Much as opinion polls are used to construct a profile of "public opinion" on a particular topic, so can stakeholder evaluations of firms be uncovered by systematically polling a company's stakeholders.

Methodologically, opinion researchers advocate random sampling as the preferred means of polling a diverse constituency, because this method produces the closest approximation of all of the constituency's characteristics. This method also allows for efficient use of resources; for example, a sample of about 500 to 1500 people is considered sufficient to represent national public opinion in the United States, with a margin of error of 3–5% at a 95% confidence level. If this methodology were systematically applied, a comparably small sample could be used to capture the diversity of perceptions that stakeholders hold. In the case of the *Fortune* survey, the Most Admired Corporations, survey data are collected from some 8000 executives and analysts annually. This sample size far exceeds the sample size needed to conduct a valid poll of all of a company's evaluators. Despite its big sample size, the *Fortune* survey has been criticized for its lack of representativeness. Heeding the warning of public opinion researchers that the validity of a poll hinges on avoiding sampling from "typical individuals" and on using subgroup quotas, more valid polls of stakeholders may be obtained by targeting

a random set of external observers rather than a circumscribed group of constituents. Recently, the Harris–Fombrun Reputation Quotient (RQ) standardized survey instrument has been developed and administered to capture the perceptions of various stakeholders, including consumers, investors, employees, and key influentials. The annual RQ is a survey of the most visible companies in the United States, and it has been conducted annually since 1999. The survey is conducted in two stages. For example, in the 2001 survey, 10,038 survey respondents (4063 by telephone and 5975 online) were interviewed to nominate two companies that stand out as having the best and the worst reputations overall. In the second "ratings" phase, 21,630 randomly selected online respondents were asked to do a detailed rating of one or two companies with which they were "very" or "somewhat" familiar. As a result, in this poll, each company was rated by an average of 600 randomly selected respondents who had indicated that they were either "very" or "somewhat" familiar with the company.

Another methodological issue that has been posed with regard to measures generated through opinion polls is whether stakeholder evaluations should be reflected in terms of average scores or distribution of opinions. It has been argued that stakeholder evaluations of firms should not be measured by averaging responses by all members of a collectivity and generating a single composite measure. Instead, stakeholder evaluations should be captured as distributions of opinions that can be represented in tables and figures as frequency distributions.

Corporate Contests

A growing body of research in sociology and management has begun to focus on a third type of process through which social evaluations are constructed. This process has been characterized as corporate contests, in that it directly pitches the performances of rivaling firms against one another. Researchers view these contests as a mechanism through which actors are evaluated relative to one another and high performers are identified and made highly visible to stakeholder publics. Victories in such contests make the value of competing firms appear self-evident, because of, as Rao put it, "the taken for granted axiom that winners are 'better' than losers and the belief that contests embody the idea of rational and impartial testing." Rao provided an interesting example of the use of contests: in the early U.S. automobile industry (1895–1912), contests were used to increase the perceived value of the new vehicle—the automobile—and to establish the reputations of its producers. The first contest was the *Times-Herald* race organized in 1895, in which five of the 11 entrants actually participated and two completed the race. An example of a modern-day corporate contest is the annual competition for the Industrial Design Excellence Award (IDEA) sponsored by *Business Week*.

Corporate contests may have some limitations in the comprehensiveness of the evaluations they afford. Rao warned against such limitations: "Contests structure search in crowded and confused markets and circumvent the issue of measuring capabilities." Two types of errors associated with contests: they foster artificial distinctions between equivalent participants; and they may lead to nonequivalence of capabilities being awarded the same level of acclaim, because winners in one year may have lesser capabilities than winners have in other years. Another key issue in organizing contests is the composition of the rival group. At one extreme, all firms, without regard to size or type, can be compared with each other on common dimensions of performance, such as progressiveness of work–family practices. Such contests, however, are likely to be dominated by size, because there is much information available about large, publicly traded firms. At the other extreme, firms could be compared only with rivals producing identical products. Though this narrows the relevant set considerably, it reduces the difficulty of making comparisons of firms across product groupings.

Forms of Social Measures of Firm Value

The different processes through which the ability of firms to create value is assessed tend to generate different forms of social measures of firm value. More specifically, the outcomes of audits and opinion polls are usually comparative reputational rankings of firms. Audits also generate expert categorizations and ratings, whereas corporate contests usually identify award "winners." Just as the three processes often coexist and are combined by information intermediaries, so the three forms are often interrelated. For example, expert ratings can be used as primary inputs for granting awards, thereby converting an audit process into a corporate contest.

Comparative Reputational Rankings

Reputational rankings are comparative orderings of organizations that compete in an industry or an organizational field. Because they juxtapose rivals in an industry in a hierarchical fashion, rankings specify the prestige ordering of the industry. The ordering reflects a firm's relative success in meeting the expectations of the industry's stakeholders. Thus, rankings are useful because they combine the judgments of different individuals on uniform criteria and enable summary comparisons of firms.

Whereas reputational rankings provide convenient direct comparisons among firms in an industry, they also pose a number of questions about their validity. One set of

questions relates to the degree to which a given set of rankings represents the evaluations of a single stakeholder or multiple stakeholder groups. This issue can be more easily addressed in evaluations based on opinion polls, which, as discussed earlier, can be designed to capture the diversity of firm stakeholders. In the case of audit processes, assuring such diversity may significantly increase the costs of information gathering and may contradict the ideology espoused by the information intermediary conducting the audit.

Recognizing and addressing the issue of the perspective of which stakeholder group a set of rankings represent has important implications for the usefulness and validity of the rankings. Much of the dispute surrounding the rival rankings of business schools presented in the press may be due to the different groups they survey and the different criteria these groups apply. The rankings of business schools published by *Business Week* are based on surveys of recruiters and alumni, but exclude faculty, students, donors, and local communities. The rankings of business schools published by *US News & World Report* include survey data from business school deans, but not survey data from any other informants. Ilia Dichev analyzed the two sets of rankings issued by these two journals between 1988 and 1996 and concluded that "the correlation between concurrent changes in *Business Week* and *US News* rankings is close to zero, even in the long-run. Thus, the cross-rankings correlation results suggest that the two rankings are largely based on different information. Since both rankings seem to reflect relevant information, it appears that neither ranking should be interpreted as a broad measure of school quality and performance . . . , but rather . . . as useful but noisy and incomplete data about school performance."

The issue of stakeholder diversity and social measures of firm value is not simply an empirical issue, but also a theoretical one. Some scholars argue that stakeholder evaluations of firms are necessarily disjointed because they reflect the contradictory interests of self-interested constituents. Others argue that stakeholder evaluations converge because constituents incorporate into their assessments implicit judgments about whether the firm is meeting the interests of other key constituents. These theoretical differences clearly spell different prescriptions about the processes through which social measures of firm value should be constructed. Empirically, the *Fortune Magazine* survey asks respondents to nominate leading companies in an economic sector and to evaluate each company on eight dimensions: (1) quality of management, (2) product/service quality, (3) long-term investment value, (4) innovativeness, (5) financial soundness, (6) ability to attract, develop, and keep talented people, (7) community and environmental responsibility, and (8) use of corporate assets. Early studies based on these ratings assumed that they were eight distinct dimensions, but Charles Fombrun and Mark Shanley have found that these dimensions were highly correlated and loaded on a single factor. They therefore concluded that when respondents rated firms on these seemingly distinct dimensions, they were in fact assessing a stable underlying construct, which could be called "reputation." In their analysis of those rankings, Fombrun and Shanley also found that, although the ratings were best predicted by financial performance variables, they were also influenced by media prominence, advertising, and charitable contributions, suggesting that respondents may unconsciously factor other constituents' concerns into their judgments. In recent years, *Fortune* appears to have recognized the unidimensionality of the construct and now no longer stresses the disaggregated ratings.

To overcome some of the problems with constructing comparative rankings, researchers have also attempted to uncover "natural" forms of status orderings in industries. In the investment banking industry, for example, the hierarchical status ordering of firms in the industry manifests itself on tombstone announcements. A tombstone announcement is a listing of a pending public security offering, which identifies the investment banks participating in the syndicate that underwrites the securities. On the tombstone, banks are listed in hierarchically arranged brackets. Listing a bank in the wrong bracket is a cause for withdrawal from the syndicate, either by the bank, or by other banks in the bracket. Using tombstones, scholars have developed an index of the relative standing of over 100 investment banks.

Awards

Unlike comparative reputational rankings, corporate awards have attracted rather limited research. Fombrun provided one of the most comprehensive treatments of the subject. He identified five major types of awards (product, process, social and environmental performance, and leadership awards). Product awards are given to recognize innovative or quality products that outperform industry standards. Many product awards are sponsored by trade journals, such as *Popular Mechanics* and *Motor Trend* in the auto industry. Process awards recognize importance organizational practices. For example, *Personnel Journal* gives the Optima Awards to companies with innovative human resource management practices. In the area of environmental performance, the United Nations environmental program sponsors the Global 500 Roll of Honor for Environmental Achievement. In the United States, over 16,000 awards are given annually to individuals and organizations by more than 6000 donors. These diverse types of awards perform common functions. From the perspective of the donors and sponsors of the awards, they highlight achievements in order to encourage others to imitate them, thereby elevating the

level of performance in an industry as a whole. From the perspective of the recipients, the awards designate some firms as "winners," thereby conferring to them a special status in their industry, as well as significant visibility relative to competitors. In addition, some awards are associated with attractive monetary prizes. The pervasiveness of awards as a form of social evaluation of firms and their significant impact on stakeholder perceptions of these firms suggest that they warrant significant future research.

Expert Ratings

Expert ratings often tend to be confused with comparative rankings. The difference between the two lies in the standard of performance that is used as a reference point for making evaluations of firms. Rankings use a relative standard and evaluate the performance of a firm relative to the performance of others firms. In contrast, ratings use an absolute standard and compare the firm to a preset performance index. This difference is important, because changes in a firm's ranking result not only from changes in its own performance on the set of criteria used by the rankings, but also from changes in the performance of other firms on those criteria. In contrast, changes in ratings indicate absolute increases or decreases in performance.

Expert ratings are often used as inputs in the construction of comparative reputational rankings or in selecting winners in contests. Yet, they can also exist independently and function as a distinct form of social measures of firm value. Examples of such measures are "star" ratings of wines, hotels, restaurants, and movies. Expert ratings are a prevalent form of product evaluation in service and entertainment industries, based on the experience of using goods, the quality of which cannot be ascertained with advance inspection.

Conclusion

Overall, social measures of firm value offer a powerful way for drawing attention to the relative success of firms at meeting stakeholder expectations. Social measures of firm value are an important feature of the institutional and competitive environments of firms because they reduce stakeholder uncertainty and search costs, during the process of selecting firms with which to exchange the resources that stakeholders control. Although most social measures of firm value have some limitations with regard to their validity, they offer the tantalizing possibility that the performance of firms can be evaluated systematically, in terms of both economic and social dimensions. The growing interest in including social measures in firm valuations suggests that they are an increasingly important area of social research and management practice.

See Also the Following Articles

Polling Industry • Polling Organizations • Ranking • Socio-Economic Considerations

Further Reading

Brealey, R., and Meyers, S. (1996). *Principles of Corporate Finance.* McGraw-Hill, New York.

Bromley, D. B. (1993). *Reputation, Image, and Impression Management.* John Wiley & Sons, Chichester, UK.

Carter, R. B., and Manaster, S. (1990). Initial public offerings and underwriter reputation. *J. Finance* **65,** 1045–1067.

Dichev, I. D. (1999). How good are business school rankings? *J. Bus.* **72,** 201–213.

DiMaggio, P. (1987). Classification in art. *Am. Sociol. Rev.* **52,** 440–455.

Fombrun, C. J. (1996). *Reputation: Realizing Value from the Corporate Image.* Harvard Business School Press, Boston, MA.

Fombrun, C. J., and Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Acad. Manage. J.* **33,** 233–258.

Freeman, R. E. (1984). *Strategic Management: A Stakeholder Approach.* Pitman Press, Boston, MA.

Gardberg, N. A., and Fombrun, C. J. (2002). For better or worse—The most visible American corporate reputations. *Corp. Reputat. Rev.* **4,** 385–391.

Hybles, R. (1995). On legitimacy, legitimation, and organizations: A critical review and integrative theoretical model. *Acad. Manage. J. Best Paper Proc.* 241–247.

Martins, L. (1998). The very visible hand of reputational rankings in US business schools. *Corp. Reputat. Rev.* **1,** 293–298.

Podolny, J. M. (1993). A status-based model of market competition. *Am. J. Sociol.* **98,** 829–872.

Pollock, T., and Rindova, V. (2003). Media legitimation effects in the market for initial public offerings. *Acad. Manage. J.* **46,** 631–642.

Rao, H. (1994). The social construction of reputation: Certification contests, legitimation, and the survival of organizations in the American automobile industry: 1895–1912. *Strateg. Manage. J.* **15,** 29–44.

Rao, H. (1998). Caveat emptor: The construction of nonprofit consumer watchdog organizations. *Am. J. Sociol.* **103,** 912–961.

Rindova, V. P., and Fombrun, C. J. (1999). Constructing competitive advantage: The role of firm-constituent interactions. *Strateg. Manage. J.* **20,** 691–710.

Roll, C. W., and Cantril, A. H. (1972). *Polls: Their Use and Misuse in Politics.* Basic Books, New York.

Siedman, G. (1992). *Awards, Honors, and Prizes.* Gale Research, Detroit.

Stiglitz, J. (2000). The contribution of the economics of information to twentieth century economics. *Q. J. Econ.* **115,** 1441–1478.

Weston, J. F., and Copeland, T. (1986). *Managerial Finance.* Dryden Press, Chicago, IL.

Social Psychology

Lisa Troyer
University of Iowa, Iowa City, Iowa, USA

Reef Youngreen
University of Iowa, Iowa City, Iowa, USA

Glossary

personality Traits and/or temperaments that characterize an individual or category of individuals.
self The organization of thoughts, feelings, and/or social roles of an individual that lend meaning to the individual from the individual's and others' perspectives and that generate expectations for the individual's behavior.
social cognition How individuals perceive, interpret, remember, and recall information about the social world and information used to understand the social world.
social environment The context in which social interaction occurs, consisting of real, implicit, or imagined others, the actions taken by one's self and others, and the symbols and objects associated with one's self and others.
social interaction Any encounter (real, implicit, or imagined) between one or more persons or directed toward one or more persons by one or more others.
social perception The processes through which individuals use information to generate impressions of others and to form inferences regarding the causes of their own and others' behaviors.
social structure Cross-cutting patterns of relations between social groups that are guided by normative conventions in institutionalized domains of social life.

Social psychology is the scientific study of the interplay between social interaction, social structure, and human thoughts, feelings, and behaviors.

Introduction

The field of social psychology focuses on the interplay among three components of social life: human thoughts, feelings, and behaviors; social interaction; and social structure. Human thoughts, feelings, and behaviors include attitudes, opinions, beliefs, emotions, cognitive processes, perception, aggression, helping, persuasion, and conformity. Social interaction encompasses interpersonal relationships, collective behavior, and inter- and intragroup processes. Social structure refers to the cross-cutting patterns of relations between social groups that are guided by normative conventions in such institutionalized domains of social life as work, family, and education. Thus, social psychology can be formally defined as the scientific study of the interplay between social interaction, social structure, and human thoughts, feelings, and behaviors.

The field of social psychology is a broad one that is a common component of the curriculum in a range of disciplines, including psychology, sociology, management and organizations, education, industrial engineering, nursing, social work, marketing, and economics. In most higher education institutions, however, social psychology is represented primarily within psychology and sociology departments. Other reviews of the field of social psychology have focused on trends in the theoretical perspectives that characterize psychological and sociological social psychology. This article primarily focuses on the kinds of topics that social psychologists in the disciplines of psychology and sociology study by examining the topics around which social psychology textbooks are organized and the topics that are reflected in published reports of social psychological research. This overview of social

psychology begins, however, by assessing its professional and intellectual background.

Social Psychology's Professional and Intellectual Roots

American Psychological Association

Division 8 of the American Psychological Association, the Society for Personality and Social Psychology (whose membership is approximately 3500), portrays its focus as follows:

> *How do people come to be who they are? How do people think about, influence, and relate to one another? These are the broad questions that personality and social psychologists strive to answer. By exploring forces within the person (such as traits, attitudes, and goals) as well as forces within the situation (such as social norms and incentives), personality and social psychologists seek to unravel the mysteries of individual and social life in areas as wide-ranging as prejudice, romantic attraction, persuasion, friendship, helping, aggression, conformity, and group interaction. Although personality psychology has traditionally focused on aspects of the individual, and social psychology on aspects of the situation, the two perspectives are tightly interwoven in psychological explanations of human behavior (Society for Personality and Social Psychology Web site, 2002, http://www.spsp.org/what.htm).*

American Sociological Association

Compare the above to how the Social Psychology Section of the American Sociological Association represents its focus:

> *Our emphases have been on the effect [of the] organization of social life on people's thoughts, feelings, and behavior, and how face-to-face interaction reproduces society. The Social Psychology Section of the ASA works to keep the spirit of social psychology alive in sociology. Today we represent over 600 scholars whose interest include self-conceptions and identity, social cognition, the shaping of emotions by culture and social structure, the creation of meaning and the negotiation of social order in everyday life, small group dynamics, and the psychological consequences of inequality. Many section members also identify with other areas of sociological research. But all bring to their research and teaching a special interest in the individual as both a social product and a social force. The common desire is to understand the many connections between individuals and the groups to which they belong (Social Psychology Section Web site, 2002, http://burkep.libarts.wsu.edu/SPNews/Purpose.htm).*

Clearly, there is much overlap across the two representations. For instance, both illustrate an emphasis on human social behavior and both reflect an interest in longstanding social concerns such as aggression and inequality. Yet these representations also reflect subtle differences between psychological social psychology and sociological social psychology, as, for example, in the former's emphasis on explaining individual outcomes and the latter's additional emphasis on the individual as a reflection of society. These different foci may, in turn, reflect the different intellectual roots that led to the development of social psychology within each field.

Intellectual Roots of Social Psychology in Psychology and Sociology

Generally, psychological social psychology recognizes three important intellectual influences: psychoanalytic theory, behaviorism, and Gestalt psychology. The psychoanalytic tradition, associated with Sigmund Freud, links an individual's behavior to psychological conflicts the individual experienced as a child within his or her family. Thus, (1) early social interaction plays a critical role in subsequent adult behavior, and (2) psychological conflicts that arise in these interactions generate motivations for subsequent behavior. The role of individuals as determinants of a focal actor's behavior is also found in behaviorism. Behaviorism, associated with John Watson and B. F. Skinner, demonstrates how behaviors are learned, primarily through the reinforcing and punishing responses of others to an individual. This line of research illustrates the important role that external stimuli (including others) have in eliciting behavior and portrays individual behavior as responsive to external forces. Finally, Gestalt psychologist emphasize how individuals think about people and objects. Central to Gestalt psychology is the notion that actors experience and understand the world in holistic, dynamic, subjective terms, rather than discrete, objective events and units. That is, the context of a unit contributes to how it is understood and experienced. While this perspective has had a strong impact on cognitive psychology (in terms of influencing theories of perception, memory, and recall), it has also led psychological social psychologists to attend to the effects of the environment (including other social actors) on human behavior. In this regard, the influence of Kurt Lewin and his field theoretical approach, emphasizing the interdependence of physiological, psychological, and social forces on social behavior, has been far reaching. In sum, these intellectual traditions each convey the importance of the social environment in determining an individual's behavior.

In contrast, although sociology is often associated with the study of macro social issues, early sociologists also

noted that social patterns reflected in individual actions, like deviance, could be attributed to social forces rather than individual motivations. These insights encouraged attempts to document effects of social forces and patterns of individual actions and emphasis on collective outcomes and group processes by early researchers such as Georg Simmel, who argued that interactions generate social reality. Social interaction among individuals, according to this work, produces society; an insight shared by Charles H. Cooley in his portrayal of society and the self as equivalent. Along similar lines, a unique form of behaviorism, social behaviorism, was being conveyed by George H. Mead through lectures in the philosophy department at the University of Chicago. The influence of social behaviorism on sociological social psychology can be contrasted with the influence of Watson's and Skinner's psychological behaviorism, described earlier. The influence of psychological behaviorism within sociological social psychology led to the development of social exchange theory, which portrays social outcomes as the product of actors' rational attempts to maximize rewards (both social and material) and minimize costs in their interactions with one another. This line of work also draws heavily on seminal theorizing on social structure by Georg Simmel, who emphasized the importance of group compositions and the configuration of relations among group members. Early sociometric research, pioneered by Jacob Moreno and developed today in the form of social network analysis, has also been an influential intellectual tradition. Social network analysis examines social outcomes as the product of relations between individuals (e.g., exchange relations) and how relations within collectives of individuals are configured. By comparison, social behaviorism casts human action as the outcome of membership in social groups, or more specifically, as the result of particular roles one assumes in different groups, and the interactions with others that those roles engender. Action, in turn, can alter the social environment. Thus, human behavior is a consequence, cause, and reflection of the social environment developed through interaction with others. This orientation toward human social experience led to the development of an important sociological perspective, symbolic interactionism, which traces human thoughts, feelings, and behavior to interaction with others in which meanings and expectations are conveyed through the exchange of symbols (e.g., words, gestures, signs). From this perspective, social environments are situations that are proactively constructed by individuals, rather than situations to which individuals react. As a result of both strains of behaviorism, sociologists gained interest in micro-level dynamics of social interaction, leading to a distinct orientation toward social psychology that emphasizes the effect of social structure on patterns of interaction between individuals and ultimately human behavior.

As this brief overview of a few key influential intellectual streams reveals, psychological social psychology and sociological social psychology can both be traced to shared acknowledgment of the embedded nature of human behavior. Yet, psychological traditions tended to steer the study of this embeddedness toward a focus on how it motivates human behavior, while sociological traditions began to treat the embeddedness as indicative of how human behavior reflects society. As a result, it is not surprising that social psychology is represented in slightly different ways by psychological and sociological professional associations. There are similar subtle differences in the organization of the curricula in the two disciplines.

Organizing Topics in Social Psychology

The landscape of social psychology is reflected in the organization of how it is taught. Table I depicts the organization of three textbooks that were commonly used by faculty teaching undergraduate social psychology courses in top-ranked sociology and psychology departments and psychology departments in 2001 and 2002. Some of the differences across the two orientations are not surprising. For instance, all three sociology textbooks give explicit attention to collective behavior, group behavior governed by norms that spontaneously emerge within the group (rather than the norms of the society in which the group is embedded). In contrast, this topic was not an organizing topic for psychology textbooks. This, however, appears to be consistent with the more macro orientation of sociological social psychology. Likewise, deviance, which is defined as behavior that violates the norms of a group or society and is thus a societal concern, is treated explicitly through dedicated chapters in textbooks used in sociology, but not those used in psychology.

Perhaps more surprising, at first glance, is that sociology textbooks dedicate chapters to communication, social structure and personality, socialization, and social power—seemingly micro-level areas of interest—while these are not given focal attention as organizing areas of the field in psychology textbooks. Also, psychology textbooks differentiate the study of social cognition and social perception, according each topic separate chapters, while sociology textbooks tend to treat both topics within a single chapter. These patterns correspond to a key difference between sociological and psychological treatments of social psychology. Each focuses on a different mechanism as a critical determinant of human experience. Throughout the sociological discussions of social psychology, the concept of social structure is given a prominent role, while throughout psychological discussions of social psychology, the concept of information processing is accorded prominence.

Table I Organization of Three Commonly Used Social Psychology Texts in Social Psychology Courses in Psychology and Sociology Departments[a]

	Textbooks used in psychology			*Textbooks used in sociology*		
Chapter topics	*Aronson, Wilson, and Akert (2002)*	*Taylor, Peplau, and Sears (2002)*	*Franzoi (2002)*	*Michener and DeLamater (1999)*	*Wiggins, Wiggins, and Van der Zanden (1994)*	*Stephan and Stephan (1990)*
Intro to social psychology	1		1	1		1
Theory		1	1	1	1	2
Methodology	2	1	2	2	1	3
Social cognition	3	2	5	5	6	9
Social perception	4	3	4	5	6	9
Communication				7	5	
Symbols/language			(2), (3)	7	5	(2)
Self	5, 6	4	3, 4	4, 9	7	5
Social structure and personality				18	7	
Socialization		(4)		3	2	4
Lifecourse				17	3	(4)
Attitudes	7	5	6	6	6, 8	10
Conformity	8	7	9	8	9	7
Deviance				19	15	8
Social influence	[b]	7	9	8		(7)
Persuasion	7	5	7	8		10
Social power					13	14
Group processes	9	10	10	13, 14, 15	4	14
Intergroup relations				16	15	15
Collective behavior				20	16	16
Interpersonal attraction	10	8	11	12	11	11
Close relationships	10	9	12	12	11	11
Pro-social behavior	11	12	14	10	10	12
Social exchange	(11)	(9)	(12)		10	
Aggression	12	13	13	11	12	13
Prejudice	13	6	8	16	6	15
Discrimination	13		8	(16)	14	
Health	14	14		18		17
Environment	15					
Politics	16	15		20	(16)	(17)
Sex/gender		11		17		6
Applications		[c]			[c]	17

[a] Numbers in table cells correspond to chapter numbers. Top-ranked programs were those listed in *U.S. News and World Report's Guide to Graduate Programs*, 2001. Three textbooks were used by 90% of the top-10 programs in each discipline. Parentheses indicate that the topic is given cursory treatment in the chapter.

[b] Entire textbook is developed around a theme of social influence (i.e., individuals are affected by others).

[c] Applications of social psychology are given specific explicit treatment in each chapter.

The influence of these mechanisms is also reflected in differences in the conceptualization of common terms. For instance, sociological social psychologists conceptualize personality somewhat differently than psychological social psychologists. Within sociological social psychology, personality refers to psychological attributes and values that differentiate groups of people. Within psychological social psychology, personality refers to traits and motivations that determine individual behaviors. As noted by House in 1990, sociological interest in personality emerged from cross-national comparisons in which researchers found that values, beliefs, and experiences of people were relatively consistent within a nation state, yet varied across states. Furthermore, within a nation state, systematic differences in the values and attitudes of groups are associated with social structural

positions. For example, researchers using a social structure and personality perspective have found that the occupational positions, socioeconomic status, and educational attainment of individuals (i.e., the categories individuals occupy in the social structure) account for their attitudes and beliefs. In summary, while the term personality reflects a group-level phenomenon in sociological social psychology, it is largely considered an individual-level phenomenon within psychology. In fact, two of the three psychological social psychology textbooks present social psychology as an alternative to personality psychology for explaining human thoughts, feelings, and behaviors. Moreover, psychologists tend to study personality as an independent variable that determines behavior, whereas sociological social psychologists tend to view personality as a dependent variable, arising from one's position within the social structure.

The topics of communication, social power, and socialization, unique to sociology textbooks, are also linked to the concept of social structure. For instance, communication, the exchange of ideas, feelings, and information between individuals, is studied in terms of (1) how meanings are constructed within and vary across cultural groups, and (2) how communication styles vary according to one's position within the social structure. Social power, the ability to induce compliance on the part of another, even when the person resists the attempt, is understood as a function of an individual's control over resources (i.e., valued commodities). Resource control, in turn, depends on one's position within a social structure; more specifically, the extent to which one is socially connected to others who possess resources, as represented by social exchange theory. Sociological social psychologists also study socialization, the process by which individuals learn social norms, values, and beliefs, from a structural standpoint. Socialization is a function of interaction with others, and interaction is determined by one's position in the social structure. Indeed, sociological social psychologists recognize socialization as a recurring lifelong process, as indicated by unique attention given to the topic of the life course in sociology textbooks. The life course is the study of patterned changes in the positions an individual occupies in the social structure that lead to the adoption of new norms, values, beliefs, and patterns of behavior emanating from these changes throughout one's life.

Additionally, as noted previously, the topics of social cognition and social perception are treated separately within textbooks used with psychology programs, whereas they are treated within the same chapter in textbooks common to sociology programs. The emphasis of psychologists on these two components of information processing may reflect the intellectual roots of this strain of social psychology in gestalt psychology, as was described above. Social cognition refers to how individuals perceive and process information related to the social world, while social perception refers more specifically to how information is used to form impressions of other people and generate inferences regarding the causes of others' behaviors. Within the field of cognition, psychological processes of attention, memory, and recall are central topics of concern, and researchers offer detailed accounts related to how information is mentally organized (e.g., categorically through schemata) and remembered (e.g., through heuristics, or mental shortcuts). These processes, in turn, generate biases that affect how the information is used. Thus, the study of cognition has important implications for social perception insofar as information that is biased through cognitive processes can become the basis for impressions that individuals form of others and inferences that individuals make regarding the causes of their own and others' behaviors.

Despite these differences, however, it is clear that there is substantial overlap in the topics that sociologists and psychologists study in the context of social psychology, as represented by the organization of textbooks. In particular, the nature and content of the self is a central concept in social psychology. However, consistent with differences in the two strains of social psychology, the self assumes a different theoretical role. The self as the embodiment of a person's knowledge of who he or she is (i.e., the self-concept) is a variable to be explained within psychological social psychology. In contrast, drawing primarily on the intellectual traditions of Mead and Cooley that were described earlier, sociologists conceptualize the self as a capacity for reflexive action. Reflexive action, behavior that is guided by one's ability to view oneself as an object in the same way one is viewed by others, represents an important cause of social behavior within sociological traditions.

The remaining overlapping topics share common conceptualizations across psychologists and sociologists. Both psychological and sociological social psychology textbooks portray attitudes as relatively enduring evaluations of people, objects, or events. Attitudes have been a topic of ongoing concern among both psychological and sociological social psychologists, dating back to Louis Thurstone and Rensis Likert's important work on attitude measurement in the 1920s and 1930s, and subsequent work by Leon Festinger and others on attitude change. Conformity is conceptualized in both sociology and psychology as behavior that is determined by the norms (i.e., rules for behavior) that are set by others. Persuasion involves the intentional use of information to change the beliefs or attitudes of another. Conformity, persuasion, and social power (discussed earlier) are subsumed by the more general concept of social influence, which refers to the effect of others on an individual's thoughts, feelings, or behavior.

Interpersonal attraction, another topic common to both psychological and sociological social psychology

textbooks, involves the study of factors that lead individuals to develop strong positive attitudes toward others that foster a desire for increased interaction with the others. The related topic of interpersonal relationships refers to the study of interactions that are based on a high degree of interdependence and strong emotional bonds between the individuals. Two behaviors of particular interest to social psychologists, aggression (behavior intended to harm another) and pro-social behavior (behavior intended to benefit another) receive primary attention in both sociological and psychological social psychology textbooks, as does prejudice (enduring negative attitudes toward members of a social group based on their membership in the group).

Not only do social psychologists in both sociology and psychology share common topics of concern; they also share a variety of research methods. As noted in the textbooks, experimental methods are a common research strategy in social psychology, but their use tends to be more dominant among psychologists, while survey methods are the most common research methodology used by sociological social psychologists. Also, observational methods receive attention in the discussion of how social psychological research is conducted, though observation is less common to research conducted in both psychological and sociological social psychology.

In summary, psychological and sociological social psychology textbooks used in top-ranked departments suggest substantial commonality with respect to the topics that are studied by both psychologists and sociologists. Differences arise primarily in the focus of each discipline on different mechanisms that explain human social experience: while psychologists focus on information processing, sociologists focus on social structure. These differences are also reflected in research articles that appear in the top psychological and sociological social psychology journals.

Research Areas in Social Psychology

Areas of research in social psychology were studied by examining the subheadings listed in the PsycINFO database for articles in the social psychology journals sponsored by the American Psychological Association (*Journal of Personality and Social Psychology*) and the American Sociological Association (*Social Psychology Quarterly*) from 1997 through 2001. The subheading assignments for some articles were more specific than others. For example, one article might be categorized under the subheadings social schema, attention, and human information storage, while another article might be categorized under the single subheading social cognition. Consequently, the categorizations of all the articles were examined, and the more specific subheadings were reclassified under the more general ones (details regarding the procedures used for this reclassification are available from the first author on request). In addition, despite the clear distinction between social cognition and social perception that characterized psychological social psychology textbooks, these subheadings appeared to be used interchangeably in the categorization of articles for both *JPSP* and *SPQ*. Thus, articles with subheadings referring to either social cognition processes or social perception processes were reclassified under the more general term social cognition and perception. The top-10 subheadings in terms of their commonality for *JPSP* and *SPQ* are given in Table II.

Overall, the list reveals striking similarities and a few notable differences. Just 12 subheadings capture the top-10 subheadings of both journals, suggesting a strong degree of overlap across sociology and psychology in the areas that garner the most research attention. Note also that the topic of social cognition and social perception is the most common for articles in both *JPSP* and *SPQ*, but it is about twice as common in *JPSP* as in *SPQ*. In fact, *JPSP* appears to be dominated by the top three or four topics, while differences in the frequency of one topic to

Table II Rankings of Ten Most Common PsycINFO Subheadings for Articles in *Journal of Personality and Social Psychology* and *Social Psychology Quarterly* from 1997 through 2001 (Commonality in Parentheses)[a]

Subheading	*JPSP (number of articles = 906)*	*SPO (number of articles = 122)*
Social cognition and perception	1 (0.430)	1 (0.205)
Personality	2 (0.253)	17 (0.074)
Self and identity	3 (0.207)	3 (0.197)
Social group differences	4 (0.199)	8 (0.148)
Emotions	5 (0.194)	9 (0.131)
Attitudes	6 (0.179)	1 (0.205)
Interpersonal relationships	7 (0.108)	5 (0.164)
Social interaction	8 (0.099)	10 (0.115)
Group processes	9 (0.097)	7 (0.156)
Sex and gender	10 (0.089)	3 (0.197)
Status	31 (0.028)	5 (0.164)
Social power	72 (0.006)	10 (0.115)

[a] Commonality is the frequency of the subheading divided by the total number of articles in the journal from 1997 to 2001. The analysis of *SPQ* articles was replicated in the Sociological Abstracts database using the descriptors assigned by this database to articles. For nearly every article, the subheadings assigned by PsycINFO were also assigned as descriptors by Sociological Abstracts. Thus, the analysis using Sociological Abstracts did not generate different results with respect to the rankings of the different topics.

the next are smaller in *SPQ*. There are a few other differences. For instance, differences arise in the prominence of status in *SPQ* (ranked fifth), and the lack of attention this area received during the same period in *JPSP* (ranked 31st). Status, the degree of social value (usually assessed in terms of prestige and esteem) accorded to actors, is generally linked to an actor's position in the social structure and is an important precursor of social influence. The processes linking status to influence (status generalization processes), detailed in status characteristics theory, describe how attributes valued in the larger society in which a group is nested become important determinants of reward allocations, competency expectations, and action within the group. Social power (which ranked 10th in *SPQ* and 72nd in *JPSP*) is also tightly coupled with the concept of social structure, and is a central concept in social exchange theory. Personality is ranked second in *JPSP*, but only 17th in *SPQ*. Its high ranking in *JPSP* likely reflects the fact that the journal is a forum for both personality psychologists and social psychologists. Given that sociological social psychologists consider personality in the context of social structure (a major area within sociology), the relatively low ranking in *SPQ* may seem curious at first. However, an examination of articles in *SPQ* adopting the social structure and personality perspective reveals that most were classified in terms of social group differences without reference to the term personality. This is not surprising, because the sociological researchers who adopt this perspective rarely invoke the term personality and instead refer to social groupings and the attitudes and interaction patterns that correspond to different social groups.

One area that characterizes a substantial amount of research, emotion, was not accorded a central organizing role in textbooks, perhaps reflecting its emerging importance in social psychological research. Emotions, the feelings of actors toward themselves, others, and events, affect and are affected by a wide array of social psychological processes. For instance, both psychologists and sociologists have proposed that emotions are socially constructed and governed by social norms. Also, sociologists have proposed that emotions play an important role in the development of one's sense of self. Yet both sociologists and psychologists recognize the biological and physiological dimensions of emotions, in addition to their social psychological dimensions.

Conclusion

Social psychology is a far-reaching field with implications for every facet of human life that gains attention in any discipline concerned with social and behavioral sciences. The fact that both psychologists and sociologists accord social psychology a prominent place in their disciplines indicates the interdisciplinary nature of the topic. Additionally, it suggests that bringing different intellectual orientations to the study of social psychological phenomenon can generate important insights for the social and behavioral sciences. Some reviews of social psychology have decried the lack of integration of psychological and sociological perspectives and research. However, as noted in our assessment of both textbooks and research articles, there is an indication that such integration is occurring, as represented by the striking overlap in research topics. Furthermore, all three sociology textbooks and one of the psychology textbooks explicitly note that social psychology is a branch of both sociology and psychology.

The strategy used here for systematically assessing the field in terms of sociological and psychological orientations may have unintentionally led to the illusion that there are more differences than similarities across these approaches to social psychology, or that any differences may be counterproductive to the growth of social psychological knowledge. To the contrary, the differences contribute important new insights and critical avenues of integration, which may lead to an enhanced body of knowledge that informs the social and behavioral sciences. Moreover, we believe that shifts in the technological and research landscape are likely to catalyze integration across the two disciplines. For example, as online databases such as PsycINFO continue to archive both sociological and psychological sources, researchers are more likely to stumble upon useful insights of others from a different orientation. Additionally, print and online journals with explicit interdisciplinary orientations (such as *Journal of Social Psychology* and *Current Research in Social Psychology*) may also facilitate the flow of knowledge across disciplinal boundaries. Likewise, increased attention to the importance and value of interdisciplinary research by higher education institutions and organizations that support research in the social and behavioral sciences may motivate further collaborations by researchers trained in different orientations. These promising shifts make it all the more important to continue to represent the interdisciplinary nature of social psychology and the topics that are covered in both sociological and psychological research. As noted by Michener and DeLamater in 1999 (p. 5; italics in the original), "*Social psychology* bridges the gap between sociology and psychology . . . As we might expect, this leads them to formulate different theories and to conduct different programs of research. Yet, these differences are best viewed as complementary rather than conflicting. Social psychology as a field is the richer for them."

See Also the Following Articles

Behavioral Psychology • Cognitive Psychology

Further Reading

Aronson, E., Wilson, T., and Akert, R. (2002). *Social Psychology,* 4th Ed. Longman, New York.

Franzoi, S. (2002). *Social Psychology*, 3rd Ed. McGraw Hill, Boston, MA.

House, J. (1977). The three faces of social psychology. *Sociometry* **40,** 161–177.

Jones, E. (1998). Major development in five decades of social psychology. In *Handbook of Social Psychology* (D. Gilbert, S. Fiske, and G. Lindzey, eds.), 4th Ed., Vol. 1, pp. 3–57. McGraw Hill, New York.

McMahon, A. (1984). The two social psychologies: Postcrises directions. *Annu Rev. Sociol.* **10,** 121–140.

Michener, A., and DeLamater J. (1999). *Social Psychology*, 4th Ed. Harcourt Brace Orlando, FL.

Stephan, C., and Stephan, W. (1990). *Two Social Psychologies*, 2nd Ed. Wadsworth, Belmont, CA.

Taylor, S., Peplau, L., and Sears, D. (2000). *Social Psychology*, 10th Ed. Prentice Hall, Upper Saddle River, NJ.

Wiggins, J., Wiggins, B., and Vander Zanden, J. (1994). *Social Psychology*, 5th Ed. McGraw Hill, New York.

Social Work

Stephen M. Marson
University of North Carolina, Pembroke, North Carolina, USA

Glossary

action system A social entity (micro, mezzo, or macro unit) that participates in an effort of planned systematic change for a client system.
change agent system A social worker or other social entity that spearheads a planned change for a client system.
client system A social entity (micro, mezzo, or macro unit) that establishes a contract for a positive change with a change agent. Client system is often abbreviated with the term "client."
Council on Social Work Education (CSWE) The organization held responsible by the Council for Higher Education Accreditation for establishing and maintaining educational standards for professional degrees in social work.
formative measures Usually a qualitative-based measurement or observation that attends to the process of a change.
operationalization A process by which a social worker or researcher moves from the abstract (concepts) to the concrete (variables).
single system design Based on statistical concepts found in control charts, it is the systematic measurement of change over time that usually includes a statistical conclusion regarding the effects of an intervention.
social worker A person who has successfully completed a baccalaureate or master's degree from an academic program accredited by the Council on Social Work Education.
summative measures Usually a quantitative-based measurement that attends to the outcome of a change.
target system A social entity (micro, mezzo, or macro unit) that is the focus of a change by a change agent and other social systems.
total institution An organization that mandates rigorous interaction patterns among its participants. Total institutions are particularly effective at maintaining accurate records that can be used for measuring baselines. The term was coined in Erving Goffman's *Asylums*.

Social work has a rich history upon which social measurement is an important foundation. Although the general public often perceives social work as the delivery of services to individuals, it is much more than that. Graduates of Bachelor of Social Work (BSW) programs and most Master in Social Work (MSW) programs receive instruction in providing a wide range of services to highly diverse client systems. The term client system is used to stress the notion that clients can be individuals, social groups, or organizations. Social measurement is a critical dimension of all social work practice, regardless of the sizes of client systems (micro, mezzo, or macro). In studying social measurement in the history of social work, it can be seen that the emphasis 100 years ago was placed on all types of client system problems. In the last three decades, micro and/or clinical practice dominate the literature of social work measurement. This recent trend does not suggest that social measurement fails to be a critical issue in macro practice; it merely indicates that less is written in the area. Most importantly, recent trends and future projects hint that social workers will see more social measurement literature with an emphasis on macro practice.

Introduction

Since the beginning of social work in the 19th century, assessing change with client systems has been an integral aspect of professional practice. Two dimensions of measurement are at the heart of assessing social problems. First is process. Here, social workers must gain insight into the steps involved to resolve a social problem. The measurement of a social problem can be addressed within

the client system, change agent, or target system. Currently and historically, process has been the most problematic issue to address in terms of measurement protocols. More creativity among social work practitioners and academicians is required within this arena.

Second is the issue of outcome. Unlike process, outcomes are easily conceptualized in terms of quantity. Thus, measurement of outcome is less problematic than the more "qualitative" process. The tools for demonstrating effective outcome measurements are available and learnable by social workers. This information comes to social work from psychometrists and their literature of tests and measures. In essence, social workers apply psychological principles of reliability and validity to the measurement of change within social problems.

History (Measurement Themes)

When first envisioning social work, one does not immediately think of social measurement; rather, one is most likely to picture the dissemination of welfare checks or removal of children from an unsafe environment. More recently, the delivery of psychotherapy as part of an agency service or in private practice may be envisioned. However, none of these visions captures the historical foundation of social work. The birth of professional social work practice can be found in social research and social measurement.

Three phases or themes in the historical development of the profession's linkage to social measurement exist. In the first theme, the measurement of social problems was the hallmark of social work activities. In the late 19th and early 20th centuries, pioneer practitioners could not conceive of their budding profession without the systematic measurement of social problems at its heart; this was a unifying theme. Second, as academic institutions became the focus for the education of the social worker, the departmentalization of knowledge arose. Splitting or dividing a curriculum into educational components or sequences has always been thought to make the educational experience more palatable for students and more manageable for faculty. Thus, the importance of social measurement and research was conceptually disconnected to the delivery of social services to needy clients. Third, the final phase includes the realization that the conceptual disconnection between social work practice and social measurement is a fatal flaw in the education of social work professionals. This is the current stage in which the profession finds itself. Today, professionals are beginning to realize that social work must reestablish itself to promote the idea that social measurement and social work practice must go hand-in-hand. Thus, social workers are beginning to realize that we must return to the original vision espoused in the later part of the 19th century. Each of these three themes is briefly discussed in this article.

Unified Theme (Amos Warner and Mary Richmond)

To understand the historical relationship between measurement and the emergence of professional social work practice, the contents and contributions of the first social work text books, three in particular, must be reviewed. The first social work textbook was Amos Warner's *American Charities*, published in 1894. Warner received his Ph.D. in economics from Johns Hopkins University. His background led him to create a classification system for establishing priorities for the delivery of social services based on statistical measurements. Thus, this first widely used textbook adopted by the first social workers was empirically based on and emerged from the systematic measurement and analysis of social problems.

The second widely used textbook was Mary Richmond's 1898 *Friendly Visiting Among the Poor*. Compared to *American Charities*, Richmond's book is considered to have had less of a social science influence. Richmond questioned the reasons for frequent failures found in social work intervention. Her effort was to systematically review failures on a case-by-case basis and draw conclusions to improve the delivery of social work services.

This effort led to Richmond's landmark social work textbook of 1917, *Social Diagnosis*. Within the pages of *Social Diagnosis*, Richmond shifted her priorities and took a strong stand on the use of social science inquiry to identify and resolve social problems. Specifically, she advocated the systematic measurement of social problems. She warned her readers of the problems of social measurement (e.g., illiteracy of clients, cultural differences that produce different meanings for the same item on a measurement scale). However, she unambiguously contended that social measurement is a critical tool for the social work practitioner. To emphasis this point, she included a wide range of measurement protocols that could be employed by the social worker for the identification and resolution of social problems, including general family issues, the immigrant family, widows with children, neglected children, unmarried mothers, blind persons, homeless men, persons with intellectual limitations, and persons with a mental illness.

Disunified Theme (Academic versus Practitioners)

Although Richmond can be seen as a pivotal figure in emphasizing social measurement for social work practitioners, she is also a pivotal figure in a movement to disengage from measurement as an integral aspect of

social work practice. With the publication of *What Is Social Case Work?* in 1922, Richmond neglected to note the importance of measurement, but rather placed emphasis on casework as a method of practice. Why do we see this major shift in Richmond's approach?

The answer to this question may lie in the unspoken prestige that existed in the academic community at that time. For example, during the first meeting of the American Sociological Society in 1905, a discussion on whether to prohibit membership to "practical sociologists" (social workers) can be found in the minutes. In the minutes of the second meeting, there is a continuing discussion of liabilities and merits of allowing social workers to join. Eventually, social workers were permitted to join, but they took a subordinate role in the society. At this point in social work history, lines of division between academic and practicing social workers began to form.

Lurie continued this theme in the *Social Work Year Book*, published in 1929. Research completed by social work practitioners was criticized for merely focusing on specific needs of the agency. In addition, the quality of the information generated by the representative agency was based on its prestige within the community rather than "measurement methods, process and results" (p. 418). Lurie also contended that some of the worst studies ever published came from practitioners rather than academicians, stating that such studies were "statistically dubious and showed an amazing ignorance of logic and of the scientific method" (p. 418). These strong words unintentionally led to a rift between the academic community and the community of practicing social workers.

The issue becomes more apparent as one examines the type of research being published by academicians. The research questions relate to social problems, but do not capture the essence of what was needed by practitioners. The central problem, of course, was an issue first introduced by Richmond—measurement of social problems. Academicians were selecting research questions that included elements in which measurement protocol achieved social scientific standards. Practitioners wanted research in areas in which concepts were difficult, if not impossible, to measure. For example, Lurie cited a wide range of research contributions produced by academicians that were thought to be important to practicing social workers. All of these cited studies offered a degree of theoretical value, but offered little to no use for the typical social worker, who asked the question, "What should I do with this client [system]?"

Unified Theme (Integration of Academic and Applied)

Issues of social measurement appear to be at the heart of the schism between practice and scholarship. However, three significant pieces of writing began to change the direction of both social work practice and social work scholarship. The first is a landmark textbook entitled *Social Work Practice: Model and Method*, written by Pincus and Minahan in 1973, which introduced a major paradigm shift in conceptualizing social work practice. Not since the publication of Richmond's *Social Diagnosis* has a text had such a dramatic effect on the practice of social work. In addition, Pincus and Minahan gained international attention and influenced the conceptualization of service delivery in both clinical psychology and psychiatry. In terms of measurement, the central focus of this textbook was outcome. The authors noted a distinction between client system process and client system outcome. In addition, they suggested that the system process is not measurable. Thus, the authors asked social workers to concede that some of the central ideas of measurement and evaluation introduced by the founders of the profession were misdirected.

Although Pincus and Minahan's observations do not seem dramatic by today's standards, their framework rejuvenated intellectual excitement within social work circles. In terms of measurement, these authors gave the profession a coherent direction to follow. Following this lead, *Evaluating Practice: Guidelines for the Accountable Professional*, by Bloom and Fischer, was published in 1982. These authors began to systematically apply concepts introduced by Pincus and Minahan. By employing single system designs, they produced a tight focus on the systematic measurement of outcomes in social services. Single system designs gave practitioners what they needed. First, these designs enabled practitioners to systematically assess outcomes, resulting in a common standard of successful outcomes. Such a standard never existed in the history of social work. Second, subjective impressions of successful outcomes were stripped away from the social worker and/or supervisor. In the past, the successful change of a target system was primarily based on perception of the change agent. With single system designs, successful outcomes were based on rejecting a null hypothesis. The scientific dimension of social work practice was no longer merely lip service. Change agents were given a tool to apply the scientific method to social work practice.

There are several serious drawbacks, however, in the employment of single system designs. First, measurement tools are necessary for the employment of single system designs. Although counting problematic behaviors is an appropriate approach for measuring, counting certainly cannot be considered the only option available for practitioners; rather, more sophisticated methods are required. Social scientific standards related to reliability and validity must be met. These standards are necessary not only for proper identification of a social problem, but also as a basis for ethical intervention. Social work

practitioners do not have the time, energy, or resources to develop a measurement that complies with social scientific standards. This problem can be seen within Richmond's *Social Diagnosis*; she was aware that her proposed measurements lacked scientific rigor.

To address this problem, in 1987 *Measures for Clinical Practice: A Sourcebook* was published. The authors, Corcoran and Fischer, searched the literature for instruments that demonstrated practical and research applications. They studied and reported on the calibration issues for each instrument that included scoring, sampling, reliability and validity. They offered enough information for the change agent to answer the question "Should I use this measurement for my client?" To support this critical practice question, most social work research textbooks include sufficient instruction for BSW and MSW graduates in the area of reliability and validity analysis. Although there are numerous monographs that achieve the same goal as the work of Corcoran and Fischer, their work included instruments that both are directly relevant to social work practice and research and have a great deal of practical application. Social workers have demonstrated such strong support that the book is in its third edition; it now offers approximately 342 instruments for clinical practice.

In terms of measurement, the introduction of single system designs for social work practice has two major drawbacks. First, the most worthy single system designs require a baseline measure. In the real world of social work practice, baselines may be either unethical or not possible. For a victim of severe depression, the change agent would be irresponsible to institute a baseline measurement. Clearly, such a strategy would be a foundation for a malpractice lawsuit. Second, real measurement (this excludes *ex post facto* or reconstructive measures) is rarely available for agencies that operate on an out-patient basis. On the other hand, these designs and associated measurements are clearly appropriate and most effective in total institutions, such as schools, nursing homes, prisons, and hospitals.

Regarding the state of the art of measurement in social work, the profession today seems to be facing a measurement problem nearly identical to the one faced at the beginning of the 20th century. In addition, one important conclusion from measurement in social work history can be drawn. The profession has made little to no contribution to the social measurement knowledge base. Essentially, social work researchers/academicians and change agents have been adopting measurement ideas (mostly from psychology) and applying these concepts to social work research and practice. However, in projecting from the past and examining current trends, it appears that social work is on the threshold of making a significant contribution to the social measurement knowledge base. Perhaps this is the beginning of another paradigm shift.

Current Standards of Practice and Scholarship

Currently, there are two trends related to measurement in the professional education of social workers. These trends focus on research methods and the educational outcomes for BSW and MSW graduates as articulated by the Council on Social Work Education. On the BSW level, the central focus is twofold. First, BSW social workers are trained to be consumers of research. BSW graduates are expected to use research findings of others to advance their skills as a change agent. Thus, students are introduced to social science research vocabulary and concepts such as reliability and validity. Second, they are expected to employ social science methods to the evaluation of practice. Evaluation is measurement. On the MSW level, we also see a twofold focus. First, MSWs are trained to be research producers and are considered leaders of the profession. Second, like the BSW students, they are expected to apply social science knowledge to practice evaluation. If research professors are earnest in their efforts, the profession will witness huge cohorts of budding professionals developing strategies for the measurement of social problems.

Operationalization

In teaching research methods with the focus described above, professors stress the concept of operationalization. In practice situations, social workers rarely intellectualize on the concrete or variable level. However, funding sources and record audits are demanding measurable outcomes. In nursing homes, failure to comply with this standard can lead to a penalty (fine). Thus, measurement of social problems is a critical issue, and social work professors attempt to address this issue by using models similar to Fig. 1. Here, students are taught the relationship between theory and research, concepts and propositions, and variables and hypotheses, and how to move from abstract thought processes to concrete measurable social problems. Without the discipline of thought processes, social workers cannot demonstrate that client systems are improving.

However, a critical problem remains. The focus of the social work research curriculum assumes that social problems must be quantified to be measured. In the real

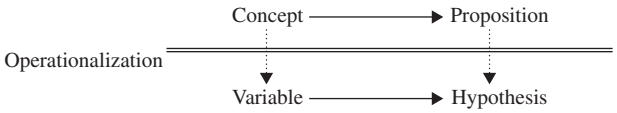


Figure 1 Central problem of measurement for social work practice.

world of social work practice, this assumption is seriously flawed. Two strategies from the academic side of social work attempt to address the issue. In the first, DePoy and Gitlin in 1998 introduced a non-traditional relationship between quantitative and qualitative research methods. They stated that the difference is not discrete, as most other authors suggest, and took the position that there is an interactive property. These authors made an obvious point that is often disregarded by most social work researchers: The nature of the research question guides the method of measurement. Taken to its extreme, one can assume that the various research strategies outlined in their book will have an impact on conceptualizing measurement of client system processes. As stated earlier, the measurement of process was conceived as out of the realm of possibility. Social workers need a new comprehensive framework.

In 1991, Alter and Evans provided such a framework, as shown in Fig. 2. This figure illustrates a common perspective shared with DePoy and Gitlin. Alter and Evans discarded the notion that outcome is the only measurable entity; rather, the change agent can measure introspectively in the realm of process and outcome (Fig. 2, right column). In addition, the change agent can measure client and target system change in terms of both process and outcome (Fig. 2, left column).

Alter and Evans advocated two different approaches to achieve their goal. First, they endorsed the position of DePoy and Gitlin. Here, they suggested that the issue of the systemic analysis of qualitative information should be revisited. Both DePoy and Gitlin and Alter and Evens stressed that social workers have not spent adequate time addressing the importance of qualitative analysis. However, Alter and Evens provided a slightly different twist when they contended that qualitative and quantitative data are not discrete entities; rather, they fall on a continuum. Second, they made systematic efforts to quantify qualitative information, advocating the use of target problem scaling and goal attainment scaling as methods of measuring process. Both of these methods have the unique characteristic of placing a numerical value on qualitative data (usually ordinal, but sometimes nominal) in an effort to measure change over time. The great strength of Alter and Evens and DePoy and Gitlin is that their work has strong implications for measurement for all social work practice—not just clinical and/or micro practice.

		Central focus of measurement	
		Client system or target system	Change agent system
Measurement strategies	Formative	Observations that monitor process or activities of client or target systems (usually qualitative)	Observations that monitor the actions/ performance of the social worker (usually qualitative)
	Summative	A measurement of the outcome of intervention (usually quantitative)	A measurement of the level of success achieved by the social worker (usually quantitative)

Figure 2 Measurement options in social work practice. Adapted from Alters and Evens (1990), p. 29.

Current Trends in Measurement

Three patterns of measurement strategies are commonly employed among practicing social workers and social work academicians: consultation, construction, and counting.

Consultation

Thousands of instruments are available and published today. The developers of such instruments have calibrated them to reach respectable levels of reliability and validity. Social workers are trained to identify when an instrument is usable for social work practice. Most importantly, many of these instruments can be found in books and on the Internet with a minimal investment of time, effort, and cost. If traditional social work citations fail, the *Mental Measurements Yearbook* can be explored. It is rare for a social worker to employ a concept that has not been operationalized.

Construction

Although it is unlikely that a social work practitioner or researcher cannot locate an instrument that is needed, that event is a distinct possibility. In addition, an instrument may be available, but the level of reliability and validity may be unacceptable. Reliability and validity of instruments become a critical issue for judges during a hearing. In such a case, the social worker must design an instrument. Instrument construction is an academic enterprise. Under normal circumstances, it takes well over a year for an instrument to reach a threshold of reasonable level of reliability and validity. Construction of new instruments is not recommended for full-time practitioners, but in some cases, no other alternative will be available.

Counting (Monitoring Designs)

For decades, social workers have been counting observations over time. Summaries of the reliability and validity of this strategy can be found in the behavior modification literature. Counting or monitoring is completed by the change agent (includes agency staff), the client system, or a combination of both. Although counting is normally an exercise to assess an outcome, if creative, a change agent

can employ monitoring to address issues of process. Following is a case illustration:

A nursing home patient was referred to a social worker because of a severe and life-threatening weight drop for which medical staff could not identify a cause. The social worker completed a psychosocial assessment that included the geriatric depression scale. There was no indication of depression or terminal drop. First, as illustrated in Fig. 3, the social worker examined the pattern of weight loss over time. From the data, it is clear that significant weight loss occurred between March 3 and April 4. To assess eating patterns, Fig. 4 was constructed. Several graphs were developed prior to Fig. 4. The earlier versions were difficult to read because of the huge amount of data. In Fig. 4, the mean percentage of food consumed per day is presented (an example of data reduction). From Fig. 4, it can be seen that March 22 and March 31 are the last dates on which acceptable levels of food were consumed.

Examining every event that occurred within the time frame (March 22–31) eliminated a psychosocial cause for the weight loss. Every note in the patient's chart was examined. Finally, the staff discovered that the patient received a new prescription to reduce blood pressure. *The Physician's Desk Reference* stated that the drug worked as an appetite suppressant for some patients. The critical weight loss problem was resolved by simply changing medications. No one realized that the medication was the cause of the life-threatening problem until food intake and weight were measured over time.

This example illustrates that in some cases, the measurement of client process is a fruitful endeavor. However, many of the rules for graph making were violated. In Fig. 3, for example, the *x* axis does not include equal intervals of time. A student would have received a poor grade on such a graph. However, in the non-academic world, data is not clean. Despite the problematic data housed in the graph, the measurement was helpful in solving a real problem.

Measuring Social Work Competence Nationally

Currently all 50 states, Puerto Rico, the District of Columbia, and Canada regulate the professional practice of social work. Most of these political entities employ the use of an instrument to ensure that these professional social workers attain a minimum level of competency. Since 1983, the Association of Social Work Boards (ASWB) has been developing and maintaining respectable levels of validity and reliability of such an instrument. ASWB has four social work examinations that test BSW graduates, MSW graduates, and MSWs with two years of post-graduate experience, both generalist and clinical. For each exam, ASWB employs a national job analysis to determine relevant skills and knowledge of currently practicing social workers. From the job analysis, a blue-print for items is developed. Items are formulated in the proportion and frequency as indicated by the blueprint. From there, each item undergoes five to eight stages in which content validity is assessed. The minimum standard for establishing a respectable level of content validity costs the agency approximately $900 per item. With a test bank that includes several thousand items, attaining respectable levels of reliability and validity (for any measurement) is not only an intellectual enterprise, it is also a costly one. ASWB does the most thorough job of addressing measurement issues of reliability and validity for the practice of social work.

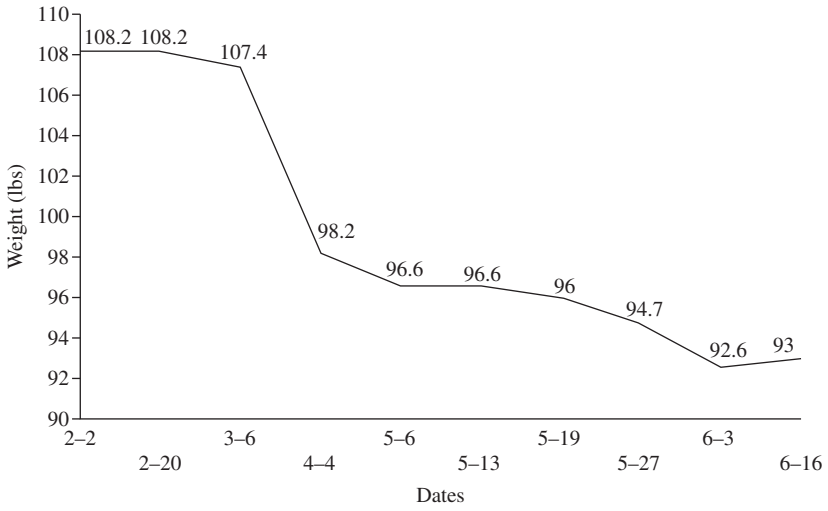


Figure 3 Weight change.

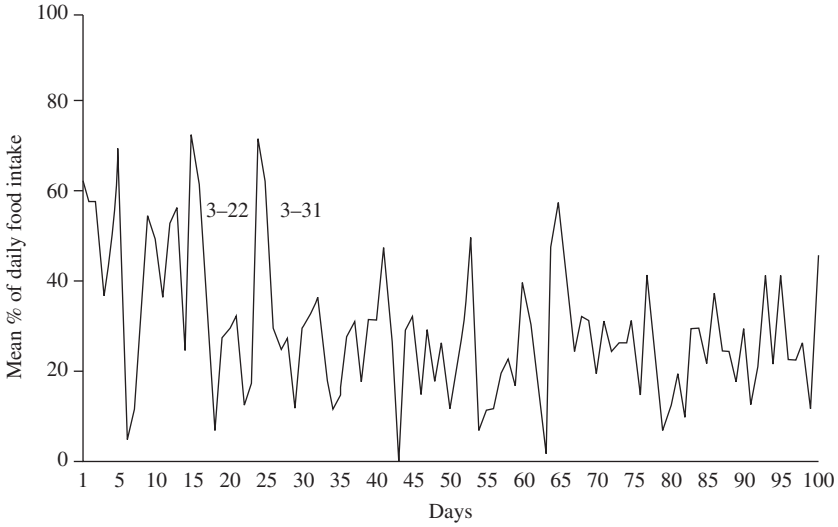


Figure 4 Food intake.

Other Issues

Issues of reliability and validity will always emerge in social work practice. For example, there is a surprisingly skewed distribution among 50 surveys addressing client satisfaction. It would be expected that the distribution of satisfaction among social service clients would be normally distributed. Among these 50 studies, the mean percentage of satisfaction is 78%. Essentially, this means that all 50 social service agencies are doing an excellent job. While this outcome is certainly possible, it is highly improbable. To most observers, such a finding is highly unlikely and probably is a result of the wording of the items. This, of course, is an issue of face validity. It should be acknowledged that the stakeholders are constructing the measures.

To address this and other issues related to measurement, the profession is moving to examine measurement in a more systematic manner. Two strategies are currently employed. First, the Council on Social Work Education is requiring BSW and MSW accredited programs to focus on the evaluation of practice. Such a program objective traverses the social work curriculum and facilitates resolving the problem noted earlier in our discussion of history of measurement in social work. With each succeeding cohort of graduates, greater insight into measurement problems will be solved.

Second, education is not enough. Professors and practitioners must have a forum to systematically address issues related to measurement. In this respect, social workers lag behind other academic disciplines and professions. During the fall of 2000, the inaugural issue of *The Journal of Social Work Research and Evaluation* was published. This journal focuses on issues of measurement and instrument development related to the delivery of social work services. The editors emphasize the importance of accepting manuscripts that include both quantitative and qualitative themes. If the journal remains true to its mission, there will be advances in the quality of measurement and in turn, the quality of service delivery.

Future of Measurement in Social Work

The best manner in which to assess the future of measurement in social work is to look at the profession's history. Several themes in this history are apparent. The first theme emerges from the work of Richmond. At the end of the 19th century and the beginning of the 20th, Richmond was well aware of the importance of accurate measurement in the effective delivery of social services to the indigent. Initially, she did not departmentalize measurement skills until the academic community demonstrated the lack of scientific rigor found in the measurements developed by practitioners. Although tension between the academic and practice arms of the profession still exists, the advent of inexpensive personal computers is diminishing its effects. However, unlike in the past, quantitative analysis of measurement may not be the central issue in social work.

The problem of social work measurement rests in the systematic examination of process. This includes client system process and change agent process. Over the past 100 years, very little work has been accomplished in this critical aspect of social work measurement. In fact, process only began to receive serious consideration during the 1990s. At this point, there is a consensus among researchers. Qualitative analysis is a legitimate approach for scientific

inquiry. The profession will move forward with increased interest and energy measuring social processes by employing qualitative methodologies. One critical area of analysis is the social history. The study of the social worker's social history is a desperately neglected area of study. The value of the qualitative measurement must receive greater scrutiny among practitioners and academics.

In terms of quantitative methods, social work will continue on its current path. Academicians will pursue social work concepts and operationalize them for use in practice, and practitioners may do the same. The Council on Social Work Education must continue its standards in the area of evaluation of practice. Social work scholars must provide a venue for the discussion and dissemination of measurement research.

See Also the Following Articles

Measurement Theory • Research Ethics Committees in the Social Sciences

Further Reading

Alter, C., and Wayne, E. (1990). *Evaluating Your Practice.* Springer, New York.

Corcoran, K. (1995). Psychometrics. In *The Encyclopedia of Social Work* (R. Edwards, ed.), Vol. III, pp. 1942–1947. National Association of Social Workers, Washington, D.C.

Corcoran, K., and Fischer, J. (2000). *Measures for Clinical Practice. Volume 1: Couples, Families and Children.* Free Press, New York.

Corcoran, K., and Fischer, J. (2000). *Measures for Clinical Practice. Volume 2: Adults.* Free Press, New York.

Cortina, J. M. (1993). What is coefficient Alpha? An examination of theory and applications. *J. Appl. Psychol.* **78,** 98–104.

Fischer, J., Bloom, M., and Orme, J. G. (2002). *Evaluation Practice.* Allyn and Bacon, Boston, MA.

Kyburg, H. E. (1984). *Theory and Measurement.* Cambridge University Press, London, UK.

Plake, B. S., Impara, J. C., and Spies, R. A. (2003). *Mental Measurements Yearbook.* 15th Ed. Buros Institute of Mental Measurements, Lincoln, NE.

Sajatovic, M., and Ramirez, L. F. (2001). *Rating Scales in Mental Health.* Lexi-Comp Inc, Hudson, OH.

Tripodi, T. (2000). The contemporary challenge in evaluating social services: An international perspective. *J. Soc. Work Res. Eval.* **1,** 5–16.

Socio-Economic Considerations

Barnett R. Parker
Pfeiffer University, Charlotte, North Carolina, USA

Arnold Reisman†
Sabanci University, Istanbul, Turkey and
Reisman and Associates, Shaker Heights, Ohio, USA

Glossary

accuracy Validity or correctness of the measurement.
hard data Data that can be obtained through direct observation by measurement or counting.
metrics Measurements of interest to the entities—parts of the universe/environment/population—to be used for gaining insight, decision-making, policy formulation, performance evaluation, etc. In some cases, metrics and raw data coincide, e.g., they are the same. In general, however, metrics are the result of some analytic procedure or algebraic relationship/combination involving two or more raw-data variables/elements.
precision Obtaining a measurement value within a given range in a large number of observations. Precision is sometimes indicated by a margin of error expressed in (+/−) percentage points.
raw data Facts and figures that are, or have been, collected/extracted from some relevant entity—a part of the universe/environment/population—and stored to be used for providing information needed to gain insight, for decision-making, policy formulation, performance evaluation, etc. Most data are numeric.
socioeconomic aspects of social measurement Consideration of any and all, intended or unintended, environmental impacts on society as a whole, or any relevant subset thereof, resulting from the dissemination and/or use of a given measurement.
soft data Data representing estimates or subjective judgment extracted from individual persons or from groups.

This article addresses the socioeconomic aspects of social measurement. Raw data are differentiated from metrics (standards of measurement) as is collection/measurement done on an *as-needed* versus on a *just-in-case* basis. Precision is differentiated from accuracy with all its ramifications. The economic and social consequences of collecting the "wrong" data, or evaluating the "wrong" metrics are discussed. The article then addresses the need for serious deliberations regarding the end-uses of data/measurements invoking, but not limited to, the use of systems analysis techniques. These allow one to clarify such issues as: the potential uses of collected data/metrics, the needed levels of aggregation and precision, the stakeholders involved, and "soft" data as surrogates for "hard" data. Several real-world examples are invoked to illustrate all of the above. The article concludes with a number of admonitions regarding the socioeconomic aspects of social measurement.

Introduction

Considerations of social measurement encompass raw data to be collected as well as metrics (standards of measurement) to be used. As will be shown, the two may well be quite different. Socioeconomic data/metrics are collected as-needed, on a just-in-case basis/philosophy, or both. When raw data are collected or metrics are measured, calculated or evaluated as they are needed, the client/customer is typically identified and the end-use is fairly well defined. This leads more easily to defining the raw-data/metrics sought and the processes to be used. However, many questions must still be asked and answered. These require much deliberation with the need to set aside conventional wisdom and/or intuition. The consequences of collecting the wrong data or

evaluating the wrong metrics may prove to be costly in economic, and disastrous in social, terms. The remainder of this article will address most, if not all, of these issues.

Importantly, the "just-in-case" situation presents, or should present, cause for serious deliberations including, but not limited to, some serious systems thinking/modeling, where such thinking can be of either the "soft" or "hard" variety.

Discussion

Unlike data, metrics may well be multi-dimensional, involving several attributes. They may be simple ratios of two attributes, for example, Labor Productivity, or they may involve a multiplicity of dimensions as in the World Competitiveness Index (WCI). It should be noted albeit parenthetically that the WCI is jointly produced and maintained by the World Economic Forum (Geneva) and the Institute for Management Development (Lausanne). It ranks selected countries in terms of their competitiveness. It depends on a number of qualitative factors, which, in turn, depend on perceptions of executives. Moreover, in 2002, Ulengin *et al*. statistically replicated the WCI by using publicly available socio-economic data for selected countries. Their data represented 61 different objective attributes grouped as; demographics, health, education, environment, technology/infrastructure, economy, and military power.

Metrics may be more directly responsive to questions of evaluation, and this may lead to executives' compensation, corporate/institutional resource allocation decisions at the microeconomic level, and to national policy at the macro level. Given human nature as it is, and when such is the case, great thought must be given to what is being measured. The various civil and criminal indictments and judicial decisions involving executives of the Enron, WorldCom, Health-South, and Tyco corporations as well as senior auditors at Arthur Andersen LLP, bankers at Credit Suisse–First Boston, and influential analysts at a number of "prestigious" Wall Street firms during the early years of this century poignantly attest to that.

The resulting measurement should direct the behavior of all actors involved in a desired direction. Deciding on what is desired is a nontrivial question as it often depends on a multiplicity of stakeholders and their respective goals and objectives. These may well prove to be (often are) conflictual. So, metric(s) considerations must, at a minimum, include answers to: Who seeks (or might seek) the information? Why is (or might be) the information needed and toward what end will it be used? What is to be measured, and at what level of aggregation? How it is to be most meaningfully measured?

Some Well-Known Metrics with Serious Untoward Consequences

The hard-data-based, easily understood, widely disseminated (US Department of Labor among others), accepted, quoted and used *labor productivity* statistics are misleading indeed; particularly so, when used to make comparisons over time or across borders or industries. One can easily make the case, for example, that the problems of America's steel industry during most of the 20th century were at least as much due to the quality of managerial plant investment decisions as they were to the quality of American labor.

The payback period is an easily understood rule for investment decisions. It is widely accepted and used by managers in plant or process modernization decisions. It has many variants but they all seek a response to the question of how long will it take to recoup the money invested. The wisdom used in applying this metric in decision making is "the shorter the payback period, the better." Any project showing payback in more than a year or two is often "axed" from further consideration. But, this rule says nothing about the magnitude of incomes, simply cumulated or discounted and then cumulated following the payback.

One can make the case that reliance on the payback period as a guide to plant modernization investments by several generations of America's steel industry executives was as much the cause of the industry's demise as were the labor unions. Needless to say, over many decades, its use in management decision-making contributed to maintaining the American steel industry's low levels of labor productivity statistics. Since payback period is a prospective decision rule, the inputs (metrics) typically used are more or less subjective estimates.

Social Measurements

Hard Data/Soft Data Tradeoffs

Some data/metrics can be secured in hard form, i.e., transactions that are counted or measured. Examples include patients serviced, personnel trained, unit costs of personnel, space, and supplies, the number of people staffing a particular department, the number of departments or clinics, and the investment in plant or equipment.

Importantly, it is not uncommon to find that such data/metrics are uncollectible and/or unavailable in hard form. Some typical examples include the cost of patients' waiting time, quality of social services, and performance of government agencies.

In between the preceding two extremes, there is a broad spectrum of data classes that can be collected in hard form, but at a cost. The costs range from trivial

to prohibitive. Also, there are classes of data which, although collectible at some cost, require time for collection. The time span may be prohibitive from the points of view of the project and/or the organization for which the data are needed. Then, there are data or metrics which, by their very nature, simply cannot be counted, measured, etc., irrespective of time and cost. Yet, these inputs may very well be key. In this category are the relative or absolute values that an organization places on various performance criteria, for example, budget balance, service level, and public image.

Also in this category are evaluations of "softer" intangible criteria, for example, institutional drawing power for quality providers and patients, and "sex appeal" (public image), as well as less soft, pseudo-tangible criteria, e.g., the reliability, maintainability, safety, and operator comfort of plant and equipment. The analyst should thus be quite flexible and broad based in designing the data-collection process.

Subjective Judgments

When data are, by nature, not collectable, then recourse is made to obtaining subjective judgments from the most knowledgeable available sources. One approach is to solicit the view(s) of a single expert. An alternative path seeks a formal consensus from a well-designed panel of knowledgeable people. This could employ some formal, interactive, consensus-seeking process, such as the Delphi method. Clearly, interviews, questionnaires, etc., taken on either a census or a survey basis, are midrange possibilities. Moreover, these highly subjective inputs that cannot be measured, nor counted, can be quantified depending on the research instruments used.

In 1979, Reisman presented various techniques of data collection within the objective/subjective spectrum for purposes of system description, evaluation, and/or valuation, forecasting demand for services, cash flows, and so on. For purposes of evaluating the outputs of services provided by a network of disparate human service agencies, several types of data, and the sources for each, are required. Some of the data can be obtained in hard form from agency records. Others require estimates, evaluations, or valuations. A real-world case involving all of the above was also provided.

Stakeholders

Stakeholders are those individuals, institutions, and other entities that have a vested interest in the issues/matter for which the data are to be collected. They have, in one way or another, some leverage and influence on the use of the data/metrics to be collected. The success or failure of data usage thus depends on the attitudes and behaviors of relevant stakeholders. In a sense, stakeholders are the data's clients. Consequently, it is important to identify the stakeholders for the data/metrics to be collected. This, as a process itself, will likely generate highly pertinent information about the perceptions and values of "clients" regarding the original problem situation.

In 1981, Mason and Mitroff reinforced this point in saying that identifying stakeholders is an easy way of generating the prevalent assumptions about a problem situation for, "while it could be difficult to 'see' assumptions, most people can rather easily generate a set of stakeholders that bears on their perspective. From the stakeholders, it is but a short step to assumptions." Identifying the stakeholders thus appears to be a prerequisite for developing models having acceptable levels of conceptual and operational validity, hopefully leading to successful model implementation.

Levels of Aggregation

As previously indicated, socioeconomic data and/or metrics are collected/measured to serve a variety of needs. Among such end-uses might be policy formulation, perfomace evaluation, resource allocation, and compensation. There may well be a number of intermediate steps that would involve models, decision rules, and related quantitative or quantified dimensions. These should be as simple as posible, yet, often, completeness contradicts simplicity. The problem is thus to decide what level of detail to include in the overall model, and what quantities to approximate and incorporate as aggregate components. This encounters a tradeoff between manageability and descriptiveness. However, in order to be complete, one needs the formulation of the model to be as detailed as possible. On the other hand, this tends to make models large and complex.

Furthermore, it is critical that the model be implementable, that is, a good representation of the situation, as well as communicable in order to enhance chances for successful implementation. (The term *representation* is emphasized as it incorporates the notion of approximation).

No model will be 100% accurate. However, the difference between a large, complex 90% accurate model and a simpler 80% accurate model may be large in terms of implementability. Experience indicates that one should err, during the earlier phases of the study, on the side of descriptiveness. However, it is often necessary in later phases to combine variables and aggregate parameters/metrics to make models manageable and, therefore, more implementable. This aggregation process should not be viewed as a compromise action. On the contrary, it requires a high level of competence. The end result's cost/benefit ratio is improved. This aggregation process often takes place as a result of data collection attempts. The latter should not be interpreted as the data are unavailable, but rather, that

the data are too costly to collect, or that earlier tests indicate the model/decision rule is insensitive to such data.

For instance, in the early phases of a systems study involving a federation of a number of disparate communal agencies, seven client groups (based on age) were identified. However, after preliminary data collection it was discovered that the differences between both the values and evaluations of the services offered to high school and junior high school groups were not sufficiently great to justify the separation of these two groups. Consequently, the two age groups were combined. Similarly, it was found that one of the agencies operated on a minimal budget as compared with all others. It was therefore decided to eliminate this particular agency from the study.

In summary, the level of aggregation employed in a model/decision rule may well be a decisive factor in the success of its usage. Simpler models are more easily managed, communicated, and understood. The user thus feels more at ease with the resulting model structure and solutions, thereby favoring the chances for successful implementation.

Levels of Precision

If the distance between point A and point B, provided within a kilometer or two, is satisfactory to the end user, there is no need to use a measuring tool accurate to plus or minus one millimiter. The additional precision comes with an increase in cost and time required for collection. This also applies to socioeconmic data/metrics.

Precision with which data are acquired is often directly related to the cost and time required to collect that data. At the same time, the precision of data/metrics must not be confused with the accuracy of the end result. If data/metrics groupings collected at differing levels of precision are then combined, the resulting entity can only be claimed to be as precise as that of the least precise grouping.

Although this concept is well established in engineering, and even more so in land surveying, accountants often count petty cash down to the penny and add such precise and hard data to rough estimates of other asset values in the millions of dollars so as to establish corporate balance sheets. Yes, over many years, a world-renowned audit firm called Arthur Andersen did indeed show that the assets and the liabilities of one of their clients, a company called Enron, did balance, to the very last cent. Great *precision* indeed. Most of the stakeholders bought into it; at least up to the point that officers of both firms began to be criminally indicted for "megafudging" on the *accuracy* of these very precise-looking figures. History has recorded the fact that this "fudging" seriously impacted many pension funds for many years, and many retirees for life, and that it was the result of unmitigated greed not only among Enron executives but also at a number of very prestigious banks, brokerage houses, and at least one law firm. None of this would have happened had more meaningful metrics been in place for evaluating performance and rewarding personnel at each of these interrelated enterprises. One could go further and query the various relevant regulatory agencies and institutions as to what metrics were being used as sensors given that such monumental and widespread wrongdoing was missed.

Levels of Accuracy

Often, though clearly not always, it is possible to check the accuracy of data or information obtained by having individuals or groups make subjective judgments against the actual events that occur. Some examples include estimates of costs and time to completion, person-hours required, task times, height and weight, as well as predictions of usage of supplies, demand for services, and manpower availability. Whenever such judgments can be validated, clearly they should be. (Some methods of validation are discussed in the work of Reisman.)

Surrogates

Measures

Often in socioeconomic systems/situations it is impossible by any means or at any cost to obtain values for the most direct measure of a variable. Consider for example, the quality of social or health care services provided. It is necessary to evaluate such phenomena using surrogate measure(s) that can be more readily quantified, albeit subjectively. Surrogate measures for the quality of food might, for example, be looks, freshness, texture, and taste. Surrogate measures for the quality of health care delivery, if measurable, might be mortality and morbidity rates, length of stay in hospital, laboratory tests performed, specialists consulted, and availability of trained nurses. Some surrogates for social welfare might be number of clients seen by professional workers, "service time," or the number of hours workers spend with, or on behalf of, clients. Others include effectiveness of services as perceived by clients, workers within the agency, and outside evaluations. Selected surrogates for education include number of trainees completing the program, starting salaries of graduates, and number of graduates going on for further studies in quality institutions.

Sources

Similar to those problems encountered in obtaining the most direct measures of a variable, it is often difficult, if not impossible, to obtain a measure, be it hard or soft (that is, subjective), from the most direct source. It is thus impossible to query certain patients regarding the quality of social service or health care received. Some patients, for example, may be too young to evaluate what they get,

while some may not be sufficiently coherent to respond. In either of the preceding cases, however, it is possible to obtain some expression of quality from the client's natural or legal guardian.

It may also be difficult to have members of the board of trustees and/or the lay community leadership to invest the time required to reach consensus on issues on which they, and only they, should act. In such cases, it has been shown possible to use surrogate panels acting on behalf of recognized leaders in seeking consensus on questions of concern using such techniques as Delphi.

Multiple Data Sources

Whenever it is possible and economical to obtain data, especially subjective estimates, from several independent sources, attempts should be made to do so. Moreover, the results of such collections should be compared and/or statistically analyzed for similarities and differences. In the case of social welfare services, for example, a high correlation was found between the level of professionalism among workers providing a service and their own estimates of effectiveness of their services vis-à-vis estimates by outside experts and clients. This finding was obtained within a highly professional institution (most, if not all, workers held advanced degrees in their specialty, and their supervisors were acknowledged leaders in the profession) for educating emotionally disturbed youngsters. Workers rated their effectiveness in the 60 to 70 range on a 0 to 100 scale. The same services were rated in the 70 to 80 range by outside experts, and by those acting on behalf of the patients, that is, parents or legal guardians.

However, in another part of the same study, a supplementary parochial school run by a husband, wife, and daughter team showed great disparity in effectiveness ratings, for example, 95 to 100 by "workers," 70 to 75 by students, and 40 to 50 by outside experts.

Checks on the Quality of Data

As indicated above, in the communal federation study, surrogate panels acted on behalf of recognized leaders in obtaining consensus on questions of concern via the Delphi method. In that study, a stratified random sample of the questions was put to the true leaders for consensus. These results, and those obtained from the surrogate panels, were then statistically analyzed to determine the level of agreement. It was found that similarly designed panels of knowledgeable people provided statistically similar responses.

An Expanded Illustrative Example

The following retrospection will illustrate many of the above issues. Lessons learned are outlined following presentation of the example (which is based on the work of Pollack-Johnson *et al.*, and Reisman).

Background

Ph.D.-granting programs in American universities became a focus of national concern in the post-Sputnik era. The number of doctorates granted during the 1960s exceeded the total production prior to 1960. During the late 1960s, however, the sitting U.S. Commissioner of Education found himself in a serious predicament. He had charge of the National Center for Educational Statistics (NCES). At the time, this was the world's most comprehensive database on educational statistics. The NCES staff statisticians included Ph.D. holders from the world's best graduate programs. The NCES continued turning out rosy projections for the demand for Ph.D.s in the sciences, and for elementary and secondary school teachers. This drove many universities to expand their graduate programs and to start new ones. The same was true for schools or departments of education. At the very same time, the *Washington Post* reported that over 30,000 Ph.D.s were unemployed or driving cabs. To make matters worse, it was reported that many an open position for grade-school teachers attracted over 400 applicants.

In consequence of the above, several efforts to forecast the supply and demand for doctorates in the sciences through the year 1980 and the supply and demand of elementary and secondary school teachers were undertaken in 1970 and 1971. Such studies were performed by government related agencies such as the NCES itself and the National Science Foundation (NSF), by private foundations such as the Commission on Human Resources and Advanced Education (CHRAE) of the Russell Sage Foundation, and by independent faculty members.

After reliable data on the actual outcomes became available, it was discovered that unlike the others, one of these forecasts was extremely accurate. Moreover, this was also true on a year-by-year basis.

It is rare to have so many concurrent forecasts of the same phenomenon for which the recency, predictability, and time horizon are all comparable. Such a similarity of conditions reduces the number of independent variables involved and makes a comparison of methodologies more meaningful. Consequently, a study of the various methodologies used and the differences between them was undertaken to see what lessons might be learned.

First, all but one of the forecasting teams relied strictly on hard data and employed various objective statistical extrapolative methodologies. One team, however, enriched what could be learned from the hard and, therefore, historical, data with subjective forward-looking inputs from a panel of knowledgeable and diverse

individuals using Delphi. The panel assigned subjective probabilities to each of a set of mutually exclusive future scenarios defined by several socioeconomic and demographic dimensions. Whatever hard data were found to exist were used to make projections under each of these scenarios. The expected outcome was then calculated by summing the products of the respective scenario-specific outcomes and their probabilities.

Identification of Data Needs

Questions of data needed and data available were addressed by first creating a relatively deaggregated feedback model to track those people flows into, and through, the educational system network. Feedback, at various levels of educational attainment, and between the educational system and the rest of society, was facilitated by having a counterpart network. Together, the two interrelated networks comprised 60 nodes. These nodes specified the educational levels and corresponding employment segments. The network model is shown in Fig. 1. Significantly, the model was established without regard to data requirements. It was based on the 1965 Bolt *et al*. framework, a forerunner to Reisman's.

Sources of Data

Extensive efforts were made to determine where and how the data required for the model of Fig. 1 could be collected. Among the agencies contacted were the National Center for Educational Statistics, National Research Council, National Academy of Science, National Science Foundation, Bureau of Labor Statistics, Bureau of the Census, National Education Association, Institute of

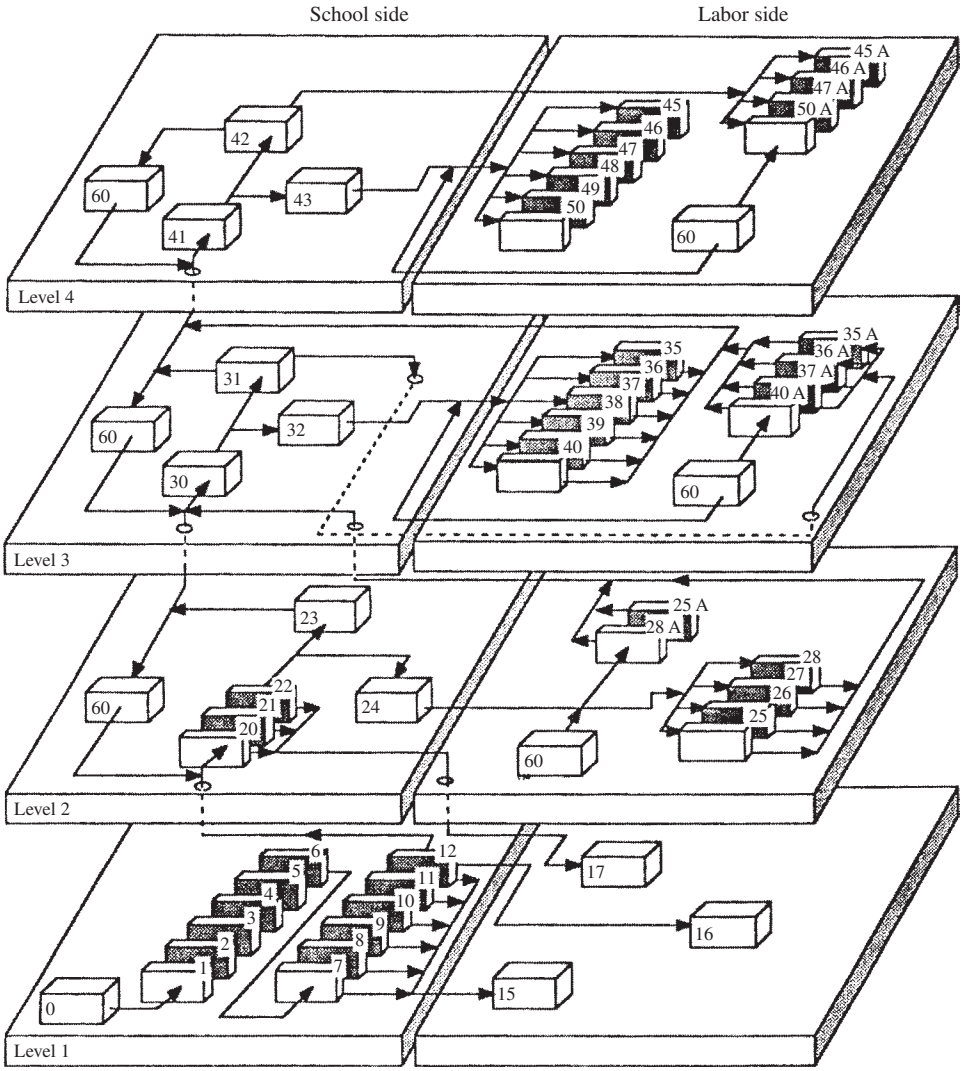


Figure 1 A highly deaggregated model of the education and use of human resources.

International Education, U.S. Department of Defense, U.S. Department of Justice, U.S. Department of State, John Mitner & Associates, Association of American University Professors (AAUP), and departments of education in a number of states. These agencies were truly helpful and generous with the data that they had in hand. On some occasions, if the data were found to be nonexistent within the agency's files, referrals to other agencies or organizations were made. Figure 2 illustrates the nodes that communicate with each other. If node i communicates with node j, either a "1" or a "0" (zero) is placed in the (i, j)th cell. A "1" indicates that flow data are available, whereas a "0" indicates that no flow data are available.

For each communicating pair of nodes, historical flow data are required. For example, node 23 (senior college students) communicates with nodes 23 (itself), 25A (labor force with at most a college degree and without teacher certification), 28A (people not in the labor force with at most a college degree and without teacher certification) and 30 (first year master's degree students). This implies that the model requires historical data (yearly) on the number of college seniors that (a) do not graduate but repeat their senior year, (b) move on to the labor sector, (c) move into the "not in the labor force" sector, and (d) move into graduate school.

Data are typically not available in flow form. Rather, the aggregated numbers of people at each node are available. Indeed, much of the aggregate numbers are often available by sex, age, or other characteristics. Using these data, the flows can, in a few instances, be reconstructed. In most cases, however, the flows have to be estimated. The "0" entries in Fig. 2 indicate that either (a) the flow data were not available, or (b) the aggregate numbers themselves could not be broken down fine enough to reconstruct the flows. As can be seen in Fig. 2, a large number of nodes have "0" entries.

The Model

The model shown in Fig. 1 required as input school-age populations. Each cohort was then pushed through the network. Using historical data, the transition probabilities

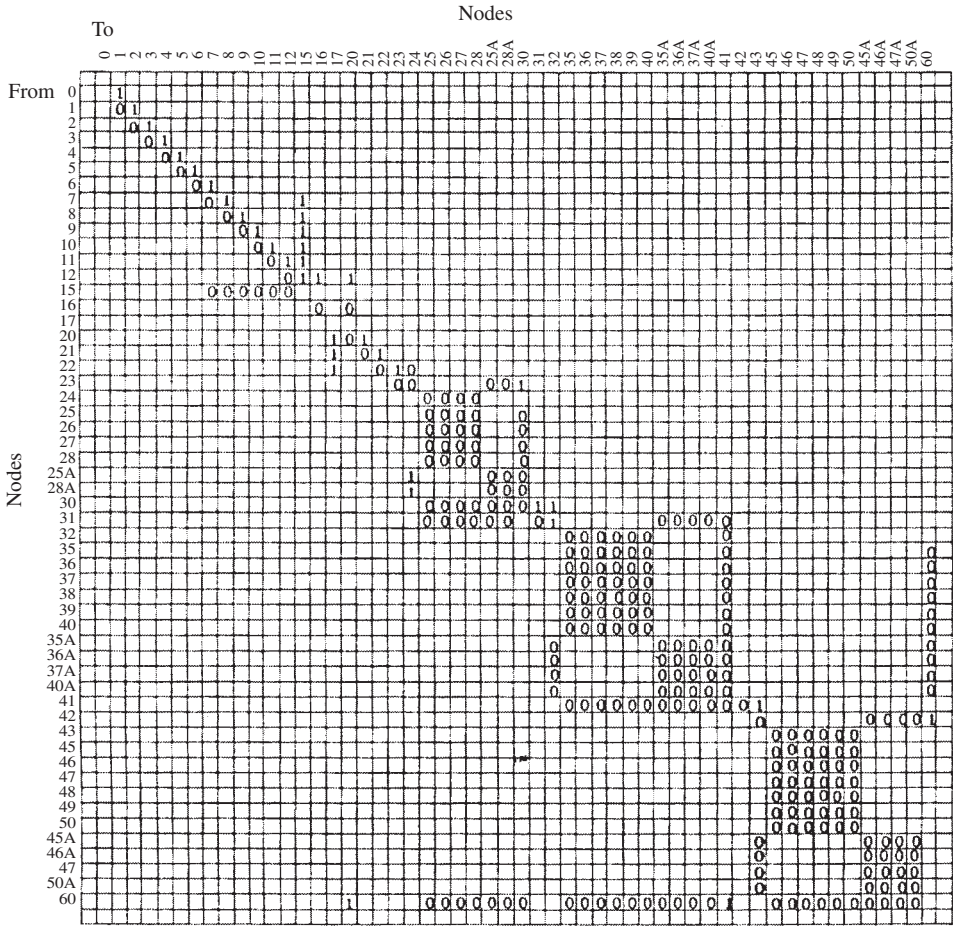


Figure 2 Data needs: data availability matrix for forecasting the education and use of human resources in the United States.

were estimated. For example, consider the node, grade 7. For a given cohort group, the proportion of students who go to the next level (in this case, grade 8) or repeat grade 7, or drop out, was determined. By tracking the transition probabilities over time and at each node, it was possible to characterize the aggregate behavior of individuals throughout the network. As a second example, consider the nodes at the master's degree production level. Enrollment in the master's degree program would come from foreign students, recent college graduates, and/or from working people with college degrees. Using the transition probabilities and the probability distribution of the time it takes to obtain a master's degree, it was possible to predict how many master's degree students reach their final year, graduate, go on for a Ph.D., go back to the labor force, and/or apply for a teaching certificate.

Nodes at the lower educational levels are more homogeneous with respect to age than are those at higher levels. Also, the choices for future movement are greater at higher levels. The statistical methodology for estimating transition probabilities differed for nodes at different levels. Estimation of such probabilities at the lower level was based on the push-based Markov chain theory. Thus, given a successful investment in education at a particular grade, the probability of moving into the next grade is well defined. In contrast, the estimation procedure of transition probabilities at the higher educational levels was based on the pull-based Renewal theory. That is, opportunities for advancement are based primarily on vacancies becoming available at the next higher level.

Clearly, the Markovian and Renewal approaches are not the only methods for modeling the flow of people through an educational network. However, to be useful, the Markovian and Renewal approaches require historical data on the flow of people among communicating nodes over time. Typically, these types of data are not directly available. In practice, data usually reflect a count of the number of individuals at given node. These data must be converted into flow data. In many instances, this is simple. For example, knowledge of the number of individuals in grade 7 in year 1984 and grade 8 in year 1985 can be used to determine the number of students who moved from grade 7 to grade 8. If the number of ways of entering grade 8, or if the number of destinations for students leaving grade 7, were high, then estimation of the flows from grade 7 to grade 8 would be difficult. To be useful, the Markovian and Renewal models require extremely disaggregated data. If flow data are not available, transition probabilities cannot be estimated and nodes must be redefined (aggregated).

Other representations of the educational network can be postulated. For example, nodes characterized by age and or sex could be important, especially beyond the bachelor's degree. In this study, separate networks were developed for males and females.

Lessons, Both General and Specific, That Can Be Learned from This Example

This example demonstrated that:

1. Even under the best of circumstances not all hard data needed are available.
2. In the matter of long-range socioeconomic forecasting, usage of subjective metrics outperformed usage of hard data.
3. In the matter of long-range socioeconomic forecasting, usage of "[non]conventional wisdom" techniques outperformed conventional techniques.
4. Use of a representivive panel/sample of stakeholders in securing subjective metrics via consensus-seeking techniques provided good metrics.
5. Usage of systems thinking and/or modeling defined the data needed as well as their availability.
6. Usage of systems thinking and/or modeling defined the levels of aggregation of the data needed as well as their availability.
7. A number or practical suggestions for long-range forecasters, especially those working in the fields of education and human resource policy and planning, resulted from this retrospective review. They can be summarized as follows:
    a. First and foremost is the importance of core assumptions: being aware of them, avoiding "assumption drag" by making sure that assumptions are current, testing them when possible, and reconciling them with known theory from disciplines beyond the immediate topic at hand.
    b. Avoiding biases due to one's institutional affiliation.
    c. Of great importance is the question of possible structural changes in the system. The form of forecasting approaches used should depend on the likelihood of achieving feasible levels of system stability. Overall, a forecasting method should be as objective as possible, as causal as it can be, and broken down into segments where appropriate. It should be as simple as needed to model the most crucial factors, and as eclectic as possible in order to be robust and to synthesize as much information as possible.
    d. In forecasting doctorate production, and other forms of human resource forecasting, it is imperative to recognize the importance of economic forces on enrollments and degree production. (Pollock-Johnson *et al*. discuss these issues at some length.)
8. Unfortunately, there are no large decision support systems for human resource planning in the educational arena.
9. Current data are dispersed across a large number of agencies.

10. Typically, data exist in the form of simple time-series, although, in some cases, cross-tabulated data are available.

11. Little attention was given to the management of this information. As a result, the nature of information varied by network level. Detailed information existed at some levels, with fundamental information missing at other levels. Census information at the lower levels, for example, was available. Survey information, however, was not generated on a regular basis. Further, the results were not monitored in longitudinal studies. Missing and conflicting data are not uncommon occurrences. Moreover, nonuniformity or changing definitions of time-series were likely problems.

12. Human resource policy formulation objectives at the national level vary from administration to administration. However, a commitment to collect and manage needed information for a decision support system (with respect to information content and quality) is in order, and long overdue. Without such attention, sophisticated educational planning models at the national level will be less well accommodated.

13. Finally, the study served to illustrate the data voids in national education statistics. It did not attempt to segregate educational disciplines, e.g., science versus humanities, nor professions, e.g., engineering versus law; and, within the teacher model, no attempt was made to distinguish between specialities, e.g., science versus mathematics versus physical education. This kind of segregation in, or de-aggregation of, the data is clearly needed to aid in the formulation of policy at the national level.

A Recent Methodology for Forecasting Social or Political Outcomes

The most recent methodology for forecasting business success as well as social, political, and foreign policy outcomes and, yes, perhaps even coups d'etats, involves Internet-based marketplaces for futures trading. This approach is certainly resonating in the business world and is beginning to be used by large corporations. As with earlier Delphi-based methods, forecasting the model's efficacy depends on inputs from knowledgeable individuals. Unlike before, however, this methodology provides financial incentives to those individuals in position to influence the outcome. Therein, one might argue, lay a serious danger.

Developments in Converting Data into Useful/Meaningful Metrics

As indicated earlier, simple metrics used for setting policy and for operational decision making are often simplistic and may result in significant, adverse, socioeconomic effects over time. Though they are intuitively satisfying and easy to use, more often than not, their long-term effects are self-evident only to those who look or want to look beyond what is immediate and obvious. In recognition of the fact that unidimensional, or even bi-dimensional (ratio form) metrics generally do not capture the complexity of the systems involved, the literature concerned with analysis for decision-making has recorded a number of multi-attribute or multi-criteria methodologies. These include, but are not limited to *Decision Tables*, and, more recently, *Data Envelopment Analysis*.

Data Envelopment Analysis

Data Envelopment Analysis (DEA) is a technique that allows for measurement of relative efficiency of organizational units. The methodology's main strength lies in its ability to capture the interplay between multiple inputs and outputs, a process that cannot be satisfactorily probed through traditional ratio analysis.

Significantly, applications of DEA have a high rate of implementation. DEA applications of record range, sector-wise, from: banking to the not-for-profits; from welfare agencies to the military; from health services to manufacturing; from education to policing. The objectives served include: organizational design, organizational effectiveness, credit evaluation, privatization, insurance underwriting, benchmarking, productivity analysis, modernization policy analysis, scale and performance measurement, physician report cards, environmental regulation, pollution prevention, facilities/equipment planning, and evaluation of macroeconomic performance.

The following web site at Warwick University, UK, provides state-of-the-art information on data envelopment analysis: www.DEAzone.com.

Data Mining

Advances in data collection and storage technology have made it possible to store large collections of data. This is especially true regarding customer and business transactions in retailing, e.g., at the point of sale. Data mining research deals with the extraction of useful and valuable information that cannot be otherwise (via standard querying tools) uncovered from large collections of data. The tools thus created allow uncovering of interesting patterns deeply buried within the data. Such patterns facilitate the making of strategic decisions.

In the context of data mining itself, the interesting problems can be categorized as classification and clustering. These partition the data items into disjoint groups or classes such as associations, which seek correlations among data items, and sequences which, as the name

implies, find the sequencing among data items. Data mining thus offers many possibilities for learning actual behavior at a point in time. Clearly, if collected from the same sources over extended time periods, and correlated with more macro data/parameters, it may offer tangible evidence of the effects of, say, business cycles on customer behavior. The following website, at the National Center for Data Mining, University of Illinois, Chicago, provides state-of-the-art information on this subject: www.ncdm.uic.edu

Concluding Remarks

There are many admonitions regarding the socioeconomic aspects of social measurement. Among these are:

1. "*Be careful what you measure' cause what you measure you will get*." Humans adjust. Some examples justifying this admonition have been provided above. Some others follow:
   a. The stock market's, and many pension fund managers', preoccupation with short-term profitability (based on various metrics) has shortened many a corporate lifespan and, with it, the jobs paying into the pension funds.
   b. Teachers, schools and, indeed, school districts (under pressure from parents and public officials) have, over time, adopted what was being taught to what was being measured on the output side, e.g., Massachusetts' statewide school-leaving test, a School Leaving Certificate examination in Nepal, and the nationwide college entrance examination in Turkey (taken at the same time by over a million students each year). Obviously the exam does not include any essays, compositions, etc. Hence, Turkish high schools, both public and private, hardly emphasize essay writing.
   c. Many a hospital has avoided accepting very serious cases after mortality statistics began appearing in community news media.
2. "*Statistics don't lie. Statisticians do*." Statistics are often used to divert attention from real problems. As an example, a president of a university in serious academic decline announces that applications are up when he should have been saying that freshmen go elsewhere, enrollments are down, and, so, let's look for internal problems.

Summary

End uses of socioeconomic data/metrics must often rely, at least in part, on data which are not measurable, countable, or, in general, "hard." Even when such data are obtainable, there are situations when they can only be secured at great monetary or human resource costs.

There are, of course, many situations where hard data cannot be obtained at any cost; hence, subjective or "soft" data may be used. The literature has recorded a number of techniques for securing such "soft" data. It behoves the data collector or metrics evaluator to choose the "right" technique to do this. The collector/evaluator must therefore identify the feasible alternative techniques, state the benefits and costs associated with each in the context of the problem at hand, and then choose the best one based on some cost-benefit analysis. In data collection or metric evaluation, consideration must be given to the level of aggregation, to cost-representative tradeoffs, as well as to end-use sensitivity to the measurements' precision. However, experience shows that it is easier to aggregate data in hand than not to have collected it in sufficient detail in the first place. So, once again, "*Be careful what you measure, cause what you measure you will get*."

See Also the Following Articles

Aggregation • Data Collection, Primary vs. Secondary • Data Mining • Measurement Theory

Further Reading

Avkiran, N. K. (2002). *Productivity Analysis in the Service Sector with Data Envelopment Analysis*, 2nd Ed. N. K. Avkiran, Camira, Queensland (www.uq.edu.au/financesite/aboutbook.htm).

Forsythe, R., Nelson, F., Newumann, G. R., and Wright, J. (1992). Anatomy of an experimental political stock market. *Am. Econ. Rev.* **82**(5), 1142–1161.

Hulse, C. (2003). Pentagon prepares a futures market on terror attacks. *The New York Times* July 29, 2003.

Mantel, S. J., Jr., Service, A., Reisman, A., Koleski, R. A., Blum, A., Dean, B. V., Reich, R., Jaffee, H., Rieger, H., Ronis, R., and Rubinstein, J. (1975). A social service measurement model. *Oper. Res.* **23**(2), 218–240.

Mason, R. O., and Mitroff, I. I. (1981). *Challenging Strategic Planning Assumptions—Theory, Cases and Techniques.* Wiley, New York.

Miser, H. J., and Quade, E. S. (1988). *Handbook of Systems Analysis: Craft Issues and Procedural Choices.* Elsevier, New York.

Pollack-Johnson, B., Dean, B. V., Reisman, A., and Michenzi, A. (1990). Predicting doctorate production in the USA: Some lessons for long-range forecasters. *Int. J. Forecast.* **6**(1), 39–52.

Reisman, A. (1979). *Systems Analysis in Health-Care Delivery.* Lexington Books, Lexington, MA.

Reisman, A. (1989). A systems approach to identifying knowledge voids in problem solving disciplines and professions: A focus on the management sciences. *Knowledge Soc.: Int. J. Knowledge Transfer* **1**(4), 67–86.

Reisman, A. (1992). *Management Science Knowledge: It's Creation, Generalization and Consolidation.* Quorum, Westport, CT.

Reisman, A. (1994). Technology management: A brief review of the last 40 years and some thoughts on its future. *IEEE Trans. Eng. Management* **41**(4), 342–346.

Reisman, A. (2003). *Data Envelopment Analysis.* Working Paper SUGSM 03-03, Sabanci University, Istanbul, Turkey. [*Socio-Econ. Planning Sci.*, in press].

Reisman, A.Oral, M. (2003). *Soft Systems Thinking: A Context within a 50-year retrospective of OR/MS.*

Reisman, A., Eisenberg, N. C., and Beckman, A. (1969). Systems analysis and description of a Jewish communal system. *J. Jewish Communal Service* **46**(1), 70–92.

Reisman, A., Ritchken, P. H., Pollack-Johnson, B., Dean, B. V., Escueta, E. S., and Li, G. (1986). On the voids in US national education statistics. *J. Econ. Soc. Measurement* **14**(4), 357–365.

Ulengin, F., Ulengin, B., and Onsel, S. (2002). A power-based measurement approach to specify macroeconomic competitiveness of countries. *Socio-Econ. Planning Sci.* **36,** 203–226.

Sociology

Joseph W. Elder
University of Wisconsin, Madison, Wisconsin, USA

Glossary

aufhebung German term used by Marx to refer to pressing contradictions to achieve transcendence, through the sublation of error and preservation of truth.
geisteswissenschaft German term for the disciplined study of products of human consciousness.
ideal type Unambiguously defined, internally consistent, meaning-directed mental construct used heuristically to compare with social behavior.
idiographic explanation Interpretation of a unique event.
nominal definition Statement that a word or group of words is synonymous with another word or group of words.
nomothetic explanation Interpreting an event by describing it as one of a set of regularly recurring events.
noumena Subjects/events as they are in themselves.
operational definition Statement defining a set of empirical operations to be used to identify a word or group of words.
phenomena Subjects or events as they appear to observers.
qualitative research Drawing on the subjects of study to define the terms of the study.
quantitative research Applying researchers' categories systematically to large numbers of subjects.
reliability The extent to which procedures of measurement produce the same results in repeated trials.
validity The extent to which measuring procedures actually measure what they purport to measure.
verstehen German term for understanding; sympathetic reliving of the experiences of others.

The 18th-century founders of sociology envisioned a "science of society," discovering "laws" that would enable social engineers to produce improved societies, just as other engineers produced improved machinery or effective cures for diseases. Concepts and the measurement of concepts lie at the heart of the sociological enterprise. Complicating those concepts and their measurement is the fact that societies (and other social groups) are collectivities of individuals. Is referring to societies as collectivities actually engaging in reification? Do collectivities have identifiable attributes as collectivities, or are their attributes basically the sums of the individual parts? If collectivities are meaningful units of analysis, how can they—or their causal processes—be defined and measured? And how is the accuracy or generalizability of those causal processes established? Over the years, sociologists have devised a number of strategies of social measurement to address these questions. Although the 18th-century vision of a law-based science of society soon faded, the idea of observing society and social behavior scientifically has persisted, generating the academic discipline of sociology.

Sociology and the Measurement of Collective Phenomena

One ultimate philosophical problem that must be addressed when dealing with the empirical world is the problem of definitions. What terms are used to describe what is perceived? In most instances, the everyday language within a family provides individuals with their initial definitions of the world. The loosely structured sets of near synonyms and antonyms of the many everyday languages around the world define the empirical world in diverse ways for the speakers of each language. When clarification of a specific word or words is called for, the word or words to be clarified (the *definiendum*) may be matched with a word or set of words (the *definiens*) describing what the original word or words mean. In such cases, a nominal definition is created—i.e., a word or group of words is declared to be synonymous with another word or group of words. Nominal definitions are neither true nor false; they are merely

statements of proposed equivalency. A word may have several nominal definitions. For example, a Marxist's nominal definition of social classes might be "groups of people occupying the same position in relation to the means of production." W. Lloyd Warner's nominal definition of social classes might be "groups of people ranked together as having the same social status." For purposes of clear communication, it is important to know which of the nominal definitions is being used. Once that has been established, the major requirement for an intelligent dialogue is consistency. As a field of knowledge expands, the number of nominal definitions within the field typically increases. In the early days of urban sociology, Louis Wirth defined cities as communities characterized by large and heterogeneous populations living within relatively small areas. A few years later, Roderick McKenzie, extending the field of urban sociology, identified four different types of cities based on their central economic functions: primary-service cities, distributive cities, industrial cities, and cities lacking any specific economic base. Sociologists could then direct their efforts at discovering similar and different patterns of human interaction in the four different types of cities.

When evidence is to be gathered from the empirical world, a nominal definition is more useful if it is accompanied by a set of empirical operations measuring that nominal definition. Ethnic groups, capitalists, the working poor, criminals, peasants, homosexuals, revolutionaries, etc. can be nominally defined. But until the identity can be determined, through some measuring procedure, precisely who out there in the empirical world is and who is not an ethnic group, a capitalist, the working poor, a criminal, etc., it is hard to make unchallenged statements describing what they are doing or what is being done to them. The set of empirical operations identified as synonymous with a word or group of words is called an operational definition (or a corresponding definition). Taking the nominal definition of social classes, for example, to be "groups of people ranked together as having the same social status," it is necessary to find a way to operationalize such groups. For example, everyone in a community could be asked to rank everyone else in the community as having higher, lower, or about the same social status, compared to themselves. The information gathered through the operational definition would enable a sociologist to group together people who felt they shared the same social status—in other words, to group people according to their social classes.

The history of sociology is a history of nominal definitions, operational definitions, and strategies for gathering data according to the operational definitions. All three are elements of sociological measurement. But none of them overcomes an ultimate philosophical problem: the impossibility of establishing an absolute identity between nominal definitions and operational definitions. Validity has been defined as the degree to which an empirical procedure actually measures what it purports to measure. Validity has also been described as the *sine qua non* of measurement; without it, measurement is said to be meaningless. Validity has been regarded as constantly evolving, as a matter of degree, as ascertained only indirectly. In the final analysis, no way has yet been found to establish some ultimate form of validity. Words and actions are not identical. Nominal definitions and operational definitions are not the same. The best that can be achieved are collectively agreed-upon approximate matches between operational and nominal definitions, recognizing that the agreement regarding the extent of the approximate match may vary between collectivities and over time.

Another ultimate problem that must be addressed when dealing with the empirical world is the problem of particulars and universals. Particular events are observed at particular times in particular places. However, the scientific exercise often requires generalization from particulars (bound by time and space) to universals (not bound by time and space). Links between particulars and universals remain open to debate. Was the 1992 Los Angeles violence following the release of the officers who beat Rodney King a particular instance of a race riot, an antipolice protest, or a class uprising? Universals are often socially defined. For example, some sociologists have stated that behavior, in order to be criminal, must be labeled criminal by an officer of the law. Does that mean that a particular theft on a particular day is not an instance of criminal behavior until it is labeled criminal behavior by an officer of the law? What happens if, after the behavior has been labeled criminal, the thief is found not guilty? Is that theft no longer a particular instance of criminal behavior? At times, it is difficult to establish linkages between particulars and universals other than collectively agreed-upon inclusion or exclusion of a given particular within the scope of some designated universal.

One way sociologists have dealt with the problem of universals and particulars is to begin with the universals (e.g., all families with annual incomes less than $20,000; all women under the age of 65) and then to take samples of particulars from the pool of universals. By operational definition, this solves the problem of linking particulars with universals. Sampling is one of the most frequently used techniques by which sociologists measure collective phenomena.

Founders of Sociology and Problems of Social Measurement

A French essayist, Auguste Comte (1798–1857), is often credited with naming and launching sociology. In his writings, Comte proposed identifying social processes of

resemblances and succession in order to identify laws of social statics and social dynamics. By establishing policies based on these laws, Comte believed that, in time, societies could be organized according to rational principles rather than religious doctrines. He envisioned the day when a hierarchy of sociologists would replace the church hierarchy and would follow empirically based processes of inquiry and application. As an alternative to calling this new process of inquiry and application social physics, Comte called it sociology.

Although the idea of administering societies scientifically according to laws of human behavior soon faded, the idea of studying societies scientifically did not. Another Frenchman, Emile Durkheim (1858–1917), advanced the new discipline of sociology by defining its parameters. Sociology was the study of "social facts" that emerge from individual facts. Social facts are exterior to individuals, they constrain individuals, and they cannot be modified by individual will. Group rates (e.g., marriage, divorce, and suicide rates) are social facts, in that they exist apart from the individuals comprising the group. Durkheim believed that social laws exist in the empirical world and that they can be discovered, like other laws, by using scientific concepts (i.e., nominal definitions created by sociologists rather than by lay people). Durkheim wrote about a society's collective conscience (its shared beliefs and sentiments) and what happened to that collective conscience as societies moved from "mechanical solidarity" (characterized by low division of labor and high collective conscience) to "organic solidarity" (characterized by high division of labor and low collective conscience). For his operational definition of collective conscience, Durkheim observed the extent to which a society enforced "repressive" laws or "restitutive" laws. "Repressive" laws inflict suffering or loss directly on violators of a society's rules. To Durkheim, this reflected a strong collective conscience with widely shared beliefs and sentiments. "Restitutive" laws merely require violators to return things to what they were before the violation. To Durkheim, restitutive law reflected a weak collective conscience and an absence of widely shared beliefs and sentiments. According to Durkheim, laws are collective phenomena (i.e., social facts); therefore, they can be used to measure other collective phenomena, such as a society's collective conscience.

One of Durkheim's most well-known studies explored relationships between three collective phenomena: suicide rates, group integration, and within-group regulation. Durkheim's operational definition of group integration included membership in a religious community (according to Durkheim, Jews had the highest group integration, followed by Catholics, Protestants, and the religiously unaffiliated). Durkheim's operational definition of group integration also included marital status (according to Durkheim, membership in a family with children provided the highest group integration; absence of family ties provided the lowest group integration). Durkheim's operational definition of within-group regulation included stability or termination of marital bonds and stability or change in economic well being. Using these operational definitions, Durkheim identified four different types of suicide: egoistic, anomic, altruistic, and fatalistic, each configured by different combinations of social facts. Although, over the years, details of Durkheim's studies (including his operational definitions) have been criticized, his approach to the study of social behavior has played a significant role in the history of sociology as a discipline that attempts to measure and explain collective behavior in the empirical world.

One of Europe's most-recognized 19th century intellectuals was Karl Marx (1818–1883). According to Friedrich Engels, who spoke at Marx's graveside, Marx had discovered the "law of development of human history" just as Darwin had discovered the "law of development of organic nature." Marx, while a student in Berlin, had been affected by the philosophical writings of Georg Friedrich Hegel. Hegel maintained that human thought provided the dynamic for constant change in human history. Hegel's unit of analysis was an entire nation, or *geist* (culture, spirit, cultural phenomenon, product of human consciousness). According to Hegel, history was a movement of cultures from "pure being" (devoid of human thought or consciousness) to the "absolute idea" (the entire historical unfolding of complete consciousness). At the core of this movement of cultures was the dialectical process whereby one concept generates an alternative concept, from which emerges a new concept. The German term for this meeting of concepts was *aufhebung* (pressing contradictions, leading to transcendence through the sublimation of error and preservation of truth). According to Hegel, ideas embedded in the *geist* generated evolving institutions, such as the sovereignty of the monarch and the state.

Marx's reading of Ludwig Feuerbach profoundly reordered his thinking. Feuerbach argued that humans and their material conditions—not abstract ideas—are the source for change. Monarchy, sovereignty, and the state were not abstract ideas emanating from the *geist*. Instead, monarchs and states existed in the material world; they sanctified their existence by concepts such as sovereignty. According to Feuerbach, Hegel had mislabeled the antecedent and the predicate. Through a "transformative method," the antecedent and consequent should be reversed.

Marx found Feuerbach's arguments convincing. He accepted Feuerbach's idea that the momentum for historical change comes not from abstract ideas but from the material conditions of existence (humans must have food and shelter before they pursue politics and religion). Marx, however, retained Hegel's concept of

historical movement through pressing contradictions. Instead of ideas generating alternative ideas from which emerged new ideas, Marx now held that material conditions generated alternative material conditions from which emerged new material conditions of existence. Using this as his framework, Marx produced a corpus of materials identifying the underlying components of the material conditions of existence and tracing their changes through history, especially European history. Key components of these material conditions included the primacy of productive forces, the formation and existence of classes, class consciousness, class conflict, exploitation, the labor theory of value, pressing contradictions, and radical transformations. As Marx introduced these concepts, he generally provided nominal definitions. Indeed, in his lifetime, he produced a lexicon of terms specifically associated with his frames of reference. However, Marx provided few operational definitions to accompany his nominal definitions.

A German contemporary of Durkheim, Georg Simmel (1858–1918), tried to define and justify the field of sociology as Durkheim had done. Simmel acknowledged that sociology's claim to be a science was challenged on the grounds that "societies" do not in reality exist, only individuals exist. Simmel argued that intellectual abstraction is an essential part of scientific thought. In reality, for example, only color molecules, letters, and particles of water exist. Yet scholars usefully study their abstractions, i.e., paintings, books, and rivers. According to Simmel, reality is studied through socially constructed categories; the choice of categories depends on the intentions of the observer. Turning to the abstraction of "society," Simmel reasoned that society emerges from multiple processes of sociation (interaction) between its component parts. Therefore, sociology might usefully pay less attention to society and more attention to the forms of sociation that occur in wide varieties of groups. Forms of sociation include competition, subordination, representation, solidarity, exclusiveness, etc. Simmel held that the relationship between sociation and social behavior paralleled the relationship between linguistics and language. Sociologists should proceed like grammarians, identifying the underlying, often "invisible" regularities in human interaction. According to Simmel, through studying sociation, criminologists might learn about the psychology of mass crimes by observing the behavior of theater audiences. Students of religion might learn about the willingness of individuals to sacrifice for the group by studying the behavior of labor-union members. Simmel introduced no special ways of measuring sociation other than commonsense observations. He deduced his most well-known descriptions of sociation not from observations, but from aspects of their definitions. In his essays, he addressed questions such as how group size affects group behavior, the point at which several conniving rogues become a delinquent gang or a social gathering becomes a party, and how three-person groups differ in essential ways from two-person groups.

In Germany in the latter half of the 19th-century, scholars were debating whether those methodologies used to study nature (*naturwissenschaft*) could be used to study cultural phenomena (*geisteswissenschaft*). The German sociologist Max Weber (1864–1920) concluded that cultural phenomena differ from natural phenomena in several fundamental ways: cultural phenomena have meaning, but natural phenomena do not. Natural phenomena occur regularly and repetitively; cultural phenomena do not. Furthermore, cultural phenomena will continue to change as long as humans are capable of raising new questions to the "eternally inexhaustible flow of life." Because cultural phenomena differ so fundamentally from natural phenomena, the methodology used to study natural phenomena cannot be used to study cultural phenomena. Students of cultural phenomena must develop their own methodology.

In his essays *The Meaning of "Ethical Neutrality" in Sociology and Economics* and *"Objectivity" in Social Science and Social Policy*, Weber outlined a methodology for studying cultural phenomena. His methodology required social scientists to understand (*verstehen*) human conduct within the cultural expressions (*ausdruck*) of its times. *Verstehen* could be achieved by placing oneself empathetically in the positions of subjects being studied and trying to relive their experiences. According to Weber, after social scientists understood subjects' behaviors and attitudes, they had a responsibility to explain those behaviors and attitudes interpretively to others. This was not a responsibility of scholars dealing with the natural sciences. Continuing his methodology for studying cultural behavior, Weber suggested that sociologists must create unambiguously defined, internally consistent "ideal types" of those cultural phenomena. Ideal types are not hypotheses (although they may help construct hypotheses). They are not descriptions of reality (although they provide unambiguous means to help describe reality). They are not an average. Ideal types are constructed by one-sidedly accentuating one or more aspects of a cultural phenomenon while synthesizing many other aspects of that phenomenon. A sociologist's own interests and value preferences determine which aspects of the phenomenon will be accentuated. Sociologists with different interests and value preferences are able to construct markedly different ideal types of the same cultural phenomenon. Ideal types (as unambiguously defined, internally consistent, meaning-directed constructs) exist nowhere in reality. Ideal types have only a heuristic value. They provide limited artificial models with which historical realities can be compared. Weber, during his career, constructed ideal types of a wide variety of cultural phenomena, including

social action, church, sect, the Protestant ethic, the spirit of capitalism, modern capitalism, classes, status groups, parties, and imperatively coordinated groups (*verband*), including those groups legitimized on the basis of "rational authority," "traditional authority," or "charismatic authority." Weber also wrote that the terms Karl Marx used to describe the material conditions of society were most useful if they were regarded not as descriptions of reality, but as ideal types—concepts with which historical realities could be usefully compared.

"Phenomenology" was the name assigned to a school of philosophy initiated by several German scholars in the first third of the 20th century. Edmund Husserl and Alfred Schutz drew on Immanuel Kant's distinctions between subjects/events as they appear to observers (phenomena) and subjects/events as they are in themselves (noumena), independent of forms imposed on them by observers' categories. Acknowledging the futility of studying noumena, Husserl and Schutz focused on the study of phenomena. According to Schutz, if social scientists are to deal objectively with the subjective meaning of human actions, they must build their constructs—including their ideal-typical constructs—on actors' common-sense constructs of their own actions. As with the *Geisteswissenschaft* movement and Max Weber, the phenomenologists held that the study of human behavior called for a methodology different from that required for the study of natural and physical phenomena.

Robert E. Park (1864–1944) is frequently identified as the founder of the Chicago school of sociology. Park's essay, *The City—Suggestions for the Investigation of Human Behavior in the Urban Environment*, presented multiple research agendas associated with the study of cities: What competing forces shape urban population patterns? What unique urban occupations emerge out of the extensive division of labor in cities? Why do crime and vice increase in cities? What excessive forms of human nature appear in cities, stimulated by the social contagion of divergent types of people who congregate in cities? The Chicago school sociologists observed that, in cities, secondary groups were replacing primary groups. Primary groups were small, face-to-face groups characterized by mutual trust and cooperation. Secondary groups were larger, more hierarchical groups that increasingly relied on coercion rather than cooperation for social control. According to the Chicago school, as primary groups were replaced by secondary groups in cities, social disorganization increased, indicated by family breakdown, divorce, drug use and alcoholism, youth gangs, vice, and criminal behavior. Terms such as "primary groups," "secondary groups," and "criminal behavior" called for nominal definitions and operational definitions. Data needed to be gathered, and statistics were increasingly called for to measure collective phenomena and test hypotheses. Sociologists specializing in sampling and survey techniques and focusing on demographic and ecological variables began developing new methodologies to meet their research needs.

Sociology and the Measurement of Causality

Measuring collective phenomena is one thing; measuring causal relationships between collective phenomena is another. The Scottish philosopher David Hume (1711–1776), in his *An Enquiry Concerning the Human Understanding*, challenged the concept of causality when he presented the following metaphysical premises: Humans derive all their ideas from their perceptions of the world in which they live. Many regularities exist in that world. Nevertheless, humans cannot observe cause and effect. At most, humans perceive spatial and temporal contiguity, temporal succession, and constant conjunction. To say they observe cause and effect adds nothing but a mystical quality to their otherwise straightforward observations. Therefore, efforts by humans to establish causality are doomed to failure.

The German metaphysician Immanuel Kant (1724–1804) maintained that reading David Hume "awakened" him from his "dogmatic slumber" and led him, ultimately, to write his *Critique of Pure Reason*. In his *Critique*, Kant distinguished between phenomena (things as humans perceive them) and noumena (things as they actually are), noting that noumena can never be known except as they are perceived. Kant observed that humans imposed conceptual categories (such as time, space, quantity, quality, relation, and modality) onto things they perceived. Kant noted further that such categories were not obtained by reflecting on the empirically given. Instead, such categories existed before, and were separate from, experienced events. They were pure intellectual constructs—universals. Individuals had to understand the universals before they could understand the particulars. Furthermore, it was by means of such categories that humans could perceive the empirically given and communicate with each other about it. But they could never actually know the empirically given. Reality would always be filtered through perceivers' conceptual categories.

Georg Friedrich Hegel (1770–1831) endorsed Kant's view that empirical events, including historical events, are filtered through human consciousness. According to Hegel, history is shaped by what humans think. One concept generates an alternative concept from which emerges a new concept through the process of *aufhebung* (pressing contradictions, leading to transcendence through the sublimation of error and the preservation of truth). The world is constantly changing. The past will never recur; the present will always generate a new

future with new ideas, until there is no further room for new ideas. Karl Marx (1818–1883), drawing on (but inverting) Hegel's perspective, applied the concept of *aufhebung* (pressing contradictions) to the material world. According to Marx, the world of material production generates pressing contradictions capable of producing changed material conditions. But such changed material conditions do not occur on their own. Humans must initiate them through praxis (the transformation of collective social perceptions into collective social action). Marx held that through the power of the owning classes, most people acquired a false consciousness of their historical conditions—historical conditions in which they were oppressed and exploited. This false consciousness led people to interpret their historical conditions as inevitable or unchangeable and to engage in social actions that reproduced, rather than transformed, the historical conditions that exploited them.

Marx identified a complex series of interrelationships between deterministic and voluntaristic elements in human history. He wrote that "no social order ever disappears before all the productive forces for which there is room in it have been developed." According to Marx, humans must evolve accurate interpretations of the social order in which they live before they can change that social order. He further argued that the juxtaposition of a social order ripe for change and accurate collective social perceptions transformed into collective social action (praxis) could produce an *umschlag* (an abrupt inversion) of the previous order and the introduction of a new social order. To test the accuracy of social perceptions, humans had to engage in praxis. Only praxis would reveal if the social order were indeed ripe for change, i.e., if all the productive forces for which there was room had been developed. If all the necessary causal factors were in place, praxis would generate an *umschlag*. If not, praxis would fail to generate an *umschlag*. In either case, causal verification required collective social action.

As a young scholar in Germany, Max Weber, found himself and his mentors surrounded by a methodological controversy: should the search for causal relationships between cultural phenomena follow the same, or different, methodologies as the search for causal relationships between natural phenomena? In addressing this question, Weber identified two different approaches to the study of causal relationships: nomothetic and idiographic. The nomothetic approach (used extensively in the study of natural phenomena) observed recurrent instances of particulars in order to develop general laws about universals. Empirical observations (including time sequences) were the core of the methodology, as was the use of conceptual or empirical controls. Following the nomothetic approach and the establishment of general laws, causes of given events were explained by "covering laws" (the reason that X was followed by Z was that X was a particular instance of Y, and Y was always followed by Z). The covering law was that Y was always followed by Z.

According to Weber, the idiographic approach (used extensively in the study of cultural phenomena) tried to explain unique historical events that could not be explained by covering laws. This called for a different methodology. In order to understand why a person acted in a certain way at a given moment in history, a sociologist had to relive empathetically what that person was experiencing at that moment. This required *verstehen* (reliving or reexperiencing) by the sociologist. If the person were living in a different culture or time period from that of the sociologist, the sociologist had to take special care to understand the expressions (*ausdruck*) of that person's time and culture in order to provide a meaningful explanation of his or her behavior. In such instances, there was no ultimate way in which it was possible for a person to validate an individual causal explanation. A description could be offered to peers about how a causal explanation was derived, but the validity of the explanation depended on how widely the explanation was or was not accepted by peers.

Max Weber also described how ideal types could be used to provide causal explanations for historical events. By definition, ideal typical actors are constructed to be clearly motivated, rationally directed non-real persons. Like a jurist in a court case, it would be possible to hypothesize how such clearly motivated non-real persons would behave in a given historical situation. Then, how actual people behaved in that same historical situation could be examined and comparisons made concerning the actual people's behavior and the non-real ideal-typical actor's behavior. The degree to which the people's actual behavior matched the non-real ideal-typical actors' behavior would suggest how closely the real meanings of the actual people matched the constructed meanings of the non-real ideal-typical actors. For example, if Max Weber's non-real ideal-typical predestined Calvinists engaged in worldly activities in order to obtain a hint of their souls' salvation, and actual historical Scottish Calvinists inaugurated extraordinary industrial activity, it could be inferred that a cause of the actual Calvinists' extraordinary industrial activity was their desire to obtain a hint of their souls' salvation. According to Weber, "In order to penetrate to the real causal interrelationships, we construct unreal ones."

Max Weber identified two different kinds of ideal types: historically individual complexes and developmental sequences. Historically individual complexes were created by accentuating certain distinctive features of some empirical phenomenon and making it into an analytic construct (e.g., church, sect, feudalism, individualism). Developmental sequences were created as a series of ideal-typical events conceptualized in space, time, and causal relationship (e.g., the routinization of charisma, the dynamics of

bureaucratization). The main purpose of both kinds of ideal types was heuristic—to compare with empirical reality. Regarding developmental sequences, Weber warned against a popular belief that sociology's goal was to produce laws of human behavior. Developmental sequences looked like such laws. They might even be mistaken for such laws. Were such to happen, ideal types would cease being heuristic and would become pernicious.

The English philosopher and economist John Stuart Mill (1806–1873), in his *A System of Logic*, outlined five "canons of inductive inference" that he believed enabled observers to agree on causes and effects in the empirical world. These canons included the method of agreement, the method of differences, the joint method of agreement and differences, the method of residues, and the method of concomitant variation. Mill based his canons on the ways in which observers found laws in the physical and natural world. According to Mill, these worlds were most effectively studied through "hypothetico-deductive" procedures that involved proposing, empirically testing, and continually refining hypotheses regarding regularly recurring events.

Considerable debate occurred, primarily in Europe, about the applicability of hypothetico-deductive procedures to the study of human behavior. "Logical positivism" was the name given to the philosophical positions associated with the Vienna Circle of the 1920s and 1930s that included such luminaries as the economist Otto Neurath, the physicist Philipp Frank, and the philosopher Moritz Schlick. The basic tenets of logical positivism included a desire to develop precise and unambiguous languages incorporating mathematical and logical symbols to refer to the empirical world, and the requirement that any proposition, causal or otherwise, must provide a method for its own verification or falsification. Logical positivists insisted on the unity of all sciences, because all sciences study the empirical world, components of which they try to link together by identifying relations of similarity and contiguity. According to the logical positivists, sociologists should use the same methodologies as physicists, astronomers, and biologists.

By the 1940s, the logical positivist movement had begun to decline, at least in part because some of its assumptions led to conclusions that were incompatible with its assumptions. One such assumption was the requirement that any proposition must be capable of verification or falsification. Efforts to establish unambiguous criteria for verification or falsification ran into serious problems. The philosopher Karl Popper, in *The Logic of Scientific Discovery*, argued that, on philosophical grounds, neither verification nor falsification could ever be attained. In their place, Popper substituted the vaguer (but possibly more attainable) criterion of "corroboration." Subsequently, Thomas Kuhn, in *The Structure of Scientific Revolutions*, described how, in the earliest stages of any field, the accumulation of knowledge was *ad hoc* (or preparadigmatic). However, as knowledge of any field accumulated, collectors of that knowledge began to operate within a paradigm (a multiple set of problems and solutions accepted by a community of scholars). Over time, those scholars would ignore an increasing number of negative instances challenging the validity of their paradigm. Finally, the negative instances would overwhelm that paradigm's credibility. At that point a new paradigm would be identified, one that accounted for many of the anomalies that had accumulated (and been ignored) under the old paradigm. In time, the majority of the scholarly community would come to reject the old paradigm and accept the new paradigm. A paradigm shift would occur, producing a scientific revolution. The new paradigm (solving some of the problems unsolved by the old paradigm but lacking some of the capabilities of the old paradigm) would generate a burst of new scholarly activity. According to Kuhn, scientific revolutions had occurred when Copernican astronomy replaced Ptolemaic astronomy, Newtonian dynamics replaced Aristotelian dynamics, and Darwinian evolution replaced teleological evolution.

Responses by sociologists to Kuhn's concepts of paradigms and scientific revolutions varied widely. Some sociologists doubted the existence of any shared sociological paradigm and suggested that sociology was still preparadigmatic. Others held that sociology had several paradigms, each with its own methodologies and causal explanations. Some scholars distinguished between revolutionary science and normal science and suggested that most scientists (including sociologists) practiced normal rather than revolutionary science. During the final decades of the 20th century, the focus of inquiry shifted from the philosophy of science to the history and sociology of science. What principles of rhetoric did Galileo use to convince his critics that their Earth-centered view of the universe was wrong and his Sun-centered view of the universe was right? How do any scientists (including sociologists) state their claims regarding facts and causal relationships, and how do they convince other scientists that their claims are valid? From a philosophical perspective, the methodology of sociology decreasingly meant strategies used by sociologists to validate their propositions and increasingly meant strategies used by sociologists to convince their peers to accept their propositions.

Sociology and Methods of Research

Sociological Training

Most undergraduate and graduate programs in sociology in the United States insist that students complete one or

more courses in statistics and research methods. The required statistics courses typically cover topics such as means, medians, modes, correlations, regressions, two-way tables, sampling distributions, confidence intervals, tests of significance, statistical inferences, probability theory, logistic models, likelihood functions, goodness of fit, binomial distributions, analyses of variance, strategies for handling missing data, and procedures for conducting special forms of analysis (such as path analysis, network analysis, factor analysis, and time-series analysis). The required research methods courses typically cover topics such as operationalizing variables, constructing hypotheses, identifying independent and dependent variables, recording observations, interviewing, drafting questionnaires and interview schedules, constructing scales, testing validity and reliability, sampling, and cleaning data after they have been collected but before they are processed. The requirement that sociological training will include courses in statistics and methods reflects a widely shared perspective in the United States that methodologies used to study natural or physical data can be applied equally well to sociological data.

The Use of Secondary Data

Because gathering data can be difficult, time consuming, and expensive, sociologists often use data that have been collected by other researchers. In 1962, the Inter-University Consortium for Political and Social Research (ICPSR) was established in Ann Arbor, Michigan to acquire, preserve, and make available to researchers wide-ranging bodies of social science data, many of them "social facts" as identified by Emile Durkheim. These data include national and international census, election, and financial information (e.g., gross domestic products, per-capita incomes, and income-distribution inequalities), opinion polls, household surveys, health information, longitudinal surveys, and statistical abstracts. During the first 40 years of its existence, more than 500 institutions joined the ICPSR. Benefits that member institutions received included access to the ICPSR's ever-expanding archives, its abilities to transfer data into new storage media to keep pace with technological changes, and the training it provides in quantitative methods of data analysis. Sociologists studying crime and criminal behavior can benefit from secondary-data sources such as uniform crime reports, prison records, and the national crime victimization survey. Sociologists interested in demography can obtain mortality and morbidity data from a wide range of secondary sources, including state and national census offices, the Centers for Disease Control and Prevention in Georgia, and relevant branches of the United Nations and the World Health Organization. As informational websites become increasingly available through the Internet, sociologists have expanding access to secondary data, with the caution that website data have not necessarily undergone professional or peer review.

The Use of Primary Data

Qualitative Research

As American sociologists generally define it, qualitative research differs from quantitative research in its greater dependence on the subjects of study (rather than on the sociologists studying the subjects) to define variables, historical sequences, and causal relationships. Qualitative research tends to focus on human agency (and hence widespread individual variability) rather than on structural agency (and hence recurrent comparability). Qualitative research is often used in case studies, including comparative case studies and extended case studies. Qualitative research is also sometimes used in pilot studies and the early stages of inquiry, when research problems and relevant variables are being initially identified.

Interviewing is one of the principal methodologies of qualitative research. Like anthropologists and psychologists, sociologists engaged in interviewing are generally trained to be sensitive to response bias and to be aware of how their rapport with their subjects, their ways of phrasing initial and follow-up questions, and their own verbal and nonverbal cues can distort the information they receive. Other methodologies used in qualitative research involve collecting oral histories, other kinds of oral productions, written materials, and visual productions. Further methodologies can include phrase-completion exercises, recording "natural" conversations, forming focus groups, and conducting various types of observations (structured and unstructured, participant and non-participant, obtrusive and unobtrusive) of individuals' behavior. Because of potentially sensitive information being obtained from individuals through such qualitative methodologies, sociologists engaged in primary research are often required to conform to certain ethical standards, including the informed consent of the participants and their own guarantees of record destruction to protect the anonymity of individuals.

Data gathered through qualitative research methodologies often require special forms of storing and retrieving, as well as of processing. Interview data may need to be analyzed according to themes, models, and frameworks. Unspoken materials, as well as spoken materials, from interviews should be reviewed. Premises and logical structures of narrative and performance data may require content analyses as well as cultural and historical contextualizations and criticisms of sources. Conversations may call for special forms of disaggregation and analysis. The skills required for dealing with qualitative data correspond in many ways to the skills of *verstehen* and *ausdruck* mentioned by Max Weber in his essay *The Meaning of "Ethical Neutrality" in Sociology and Economics*. The

perception is shared among many sociologists that qualitative researchers are more likely than quantitative researchers to question, as Max Weber and the phenomenologists did, the applicability of physical/natural science methods to the study of social behavior.

Quantitative Research

For sociologists, one of the main purposes of quantitative research is to make statements with some degree of confidence about sizable human groups. Such statements typically accept the hypothetico-deductive premises of regularity and recurrence seen to be present in the natural and physical sciences. Quantitative research often begins with the researcher selecting the population (or universe) to be studied, based on the nature of the questions to be answered. The selected universe could be all the citizens of a designated nation, laborers in a sweatshop, or the homeless in a large urban center. It could be women chief executive officers of *Fortune 500* corporations, or children raised in homes with same-gender parents. If the universe is large, the researcher might wish to select a manageable-sized sample from which to gather data. The sample could be selected in such a way that every unit in the universe has an equal likelihood of being selected (a probability sample). Or the sample could be structured in a way that guarantees the inclusion of certain subsectors (e.g., genders, ages, locations, ethnic groups) of the universe. An initial sample could be selected on the basis of convenience (a nonprobability sample) and then expanded by "snowball" referrals to additional units of the sample. The manner in which the sample is selected directly affects the confidence with which findings from the sample can be generalized statistically to the universe. A researcher can introduce additional degrees of sophistication by engaging in time-series sampling, paired-comparative sampling, and longitudinal panel studies.

The hypothetico-deductive approach calls for examining statements of purported invariant relationships (hypotheses) between antecedent (independent) variables and consequent (dependent) variables, presented so that the relationship is capable of falsification. In order to be capable of falsification, such statements must be nontautological and without empirically continuous antecedents or consequents, and they must specify the relationships between the antecedents and consequents (e.g., as necessary conditions or sufficient conditions). If, after the data have been collected and examined, the hypotheses have not been falsified, it can be stated that the data demonstrate the hypotheses. But it cannot be stated that the hypotheses have been "proved." "Proving" invariance requires evidence that has yet to be gathered.

In order for a statement of invariance to be examined, the variables in the statement must be operationalized (i.e., made capable of observation and measurement). If, for example, the statement refers to domestic violence, democracy, or intelligence, instructions must be provided for how data are to be gathered as evidence of domestic violence, democracy, or intelligence. Furthermore, the case must be made that the data obtained by following the instructions are valid indicators (i.e., actually provide evidence) of domestic violence, democracy, etc. This case must be made rhetorically, because it cannot be made empirically.

Comparisons are at the heart of sociological research: comparisons of sociological phenomena at two points in time, comparisons of dependent variables following changes in independent variables, and quasi-experiments in which treatment effects on dependent variables are compared with control effects on comparable dependent variables while accounting for experimenter effects. Quantitative research tends to focus on structural agency (rather than human agency) and hence recurrent comparability. The goal of such research is to demonstrate empirically supported sociological relationships, often in sizable human groups. In order to gather systematic information from large numbers of people, sociological researchers frequently conduct surveys for which they prepare data-gathering instruments, such as questionnaires (for written responses) and interview schedules (for oral responses). The questions asked in these instruments may be close ended (with limited fixed choices) or open ended. The major advantage of close-ended questions is the ease of final tabulation. The major advantage of open-ended questions is the possible acquisition of useful unanticipated information. If replies to open-ended questions are to be used systematically, they must be coded. Raters who code the open-ended answers must be trained so as to minimize differences in their coding patterns and to maximize their interrater reliability.

Prior to using questionnaires or interview-schedules in the field, it is essential to pretest them with a sample of respondents who will not be included in the final survey. Pretests may uncover problems of question clarity and accuracy. Because the phrasing of questions can alter respondents' answers, efforts are usually made to ask each respondent identically worded questions. At times, however, wording equivalency may be more important than wording identity. In a survey of sexual behavior, for example, alternative words were substituted for certain sexual practices so that respondents from different social backgrounds could understand what sexual practices were being referenced. Designers of questionnaires and interview schedules can prepare scales (such as Guttman scales, Likert scales, or Thurston scales) to measure relative degrees of attitudes or practices. Interview schedules require trained interviewers to gather data in face-to-face conversations or over the telephone. One advantage of using interview schedules rather than questionnaires is the possibility of asking respondents to answer follow-up questions for clarification or additional

details. Advantages of using questionnaires generally include greater speed and lower cost of data collection.

Reliability (the extent to which questionnaires or interviews give the same results in repeated trials) is desirable in survey instruments. Asking the same question in different ways can test a procedure's reliability. In a survey of sexual behavior, for example, differently worded questions were asked at different points in the interview about the number and gender of the respondent's recent sexual partners. Afterward, the respondents' replies to each of the different questions were compared for consistency, and hence for procedure reliability. Another technique for testing a survey's reliability is to reinterview a random selection of already interviewed subjects. The reinterview can achieve two ends: it can establish the fact that the subjects were indeed interviewed, and it can identify differences between the ways subjects originally answered the questions and the ways they answered the questions the second time. The narrowness of differences between the first and second answers would be a measure of the survey instrument's reliability. Arranging for questions to be asked in multiple languages requires special care. A useful technique is to have one translator convert the English questions into Spanish (for example) and a different translator convert the Spanish questions back into English. Comparing the original English with the "back-translated" English could identify words or phrases needing to be retranslated.

The heart of the sociological enterprise involves concepts and the measurement of concepts. Over the years, increasingly sophisticated statistical procedures have refined researchers' abilities to analyze data. But sociology's fundamental philosophical problems remain the same: How are social collectivities accurately defined and measured, and how are their causal relationships established?

See Also the Following Articles

Data Collection, Primary vs. Secondary • Measurement Theory • Phenomenology • Qualitative Analysis, Sociology • Reliability • Secondary Data • Validity Assessment • Weber, Max

Further Reading

Emerson, R. M., Fretz, R. I., and Shaw, L. L. (1995). *Writing Ethnographic Fieldnotes.* University of Chicago Press, Chicago, IL.

Mills, C. W. (1959). *The Sociological Imagination.* Oxford University Press, New York.

Ragin, C. C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies.* University of California Press, Berkeley, CA.

Riessman, C. K. (1993). *Narrative Analysis.* Sage Publ., Newbury Park, CA.

Schutt, R. K. (2001). *Investigating the Social World: The Process and Practice of Research.* 3rd Ed. Pine Forge Press, Thousand Oaks, CA.

Stern, P. C., and Kalof, L. (1996). *Evaluating Social Science Research.* 2nd Ed. Oxford University Press, New York.

Software

Micah Altman
Harvard University, Cambridge, Massachusetts, USA

Glossary

algorithm A precise set of instructions, in an abstract programming language, describing how to solve a problem.
compilation The translation of instructions in a programming language into a machine language.
machine language The set of instructions directly executable by a particular type of computer hardware.
program A sequence of instructions, written in some programming language, that a computer can execute.
programming language An artificial language used to write computer programs that can be translated into a machine language.
pseudocode A notation resembling a programming language but intended for pedagogy, not translation into machine language.
software engineering The discipline engaged in systematic study and practice of designing and implementing software.
software rot The tendency of software to fail as it becomes older. Usually this is because of changes in the operating environment or limitations in the design assumptions and not because of changes to the software.

Software, generally defined, is a set of instructions designed to be executed by a computer. Practically every modern method of quantitative and statistical analysis relies on software for its execution. Despite its ubiquity, software is usually regarded (when considered at all) as necessary, but preferably invisible, infrastructure for such research. In fact, an understanding of software is essential to correct and efficient application of many quantitative and statistical methods.

Introduction

How Is Software Used?

The use of software in social science research is extensive and wide ranging. Software is used at every stage of the research process, from collecting, organizing, and analyzing information to writing and disseminating results. The mathematically demanding nature of modern statistical analysis makes the use of relatively sophisticated statistical software a prerequisite for almost all quantitative research. Moreover, the combination of increasingly powerful computers, ubiquitous computer networks, and the widespread availability of the software necessary to take advantage of both has made practical on a hitherto unprecedented scale the application of many complex methods, such as maximum likelihood estimation, agent-based modeling, analytic cartography, and experimental economics.

Brief History

The idea of the algorithm, which is basic to all computer software, dates back as far as 825 CE to the Persian mathematician Abu 'Abd Allah Muhammad ibn Musa Al-Khwarizimi. The invention of the computer program is much more recent: Ada Lovelace is usually attributed with creating the first theoretical computer programs in 1843, to be used with Charles Babbage's analytical engine (which was never physically constructed).

The modern form of "software," a set of instructions that is separated from the physical computer, originated in 1945 when John Von Neumann first proposed the "stored program." This stored program (now known as a "computer program") would comprise a set of instructions for a general-purpose computer that would be stored in the computer's memory, along with the data, rather than being physically "wired" into the computer hardware. (The Eniac, the first general-purpose, programmable, electronic digital computer, had to be rewired in order to run different programs; subsequent digital computers have followed Von Neumann's design.)

```
function Standard_deviation (X: vector ) {
  variables x_sum, x_mean, x_std, : real ; n ,
   count: integer;
  x_sum = 0;
  n = length(X) ;
  for count = 1 to n {
    x_sum = x_sum + X[count];
  }
  x_mean = x_sum/n;
  x_dev = 0;
  for count = 1 to n {
    x_dev = x_dev + (x_mean - X[count])^2;
  }
  x_dev = sqrt(x_dev/n);
  return (x_dev);
}
```

Figure 1 Pseudocode for computing the standard deviation.

In the earliest period of digital computing, from the construction of the Eniac in 1945 through the mid-1950s, software was developed either by the end user or by the manufacturer. Typically, the vendor delivered, at most, low-level utility programs. The end user would have to write whatever software that he or she needed, or copy it from another user.

A market for software contracting first developed in the mid-1950s, but application software remained entirely custom written: The vendors of the mainframes of the day would supply freely the operating systems software needed to run the system since such software was coupled so closely to that particular type of hardware that no one else had the experience or the time to develop it. All other software was written by the end user or by contractors, who would write custom software tailored to the needs of that user and application.

It was only in the 1960s that generalized software packages products first emerged and not until the late 1970s that software became a "shrink-wrapped," standardized, stand-alone, mass-market commodity. Today, much of the software used in social science research is of the shrink-wrap variety: a stand-alone package capable of performing a wide variety of functions and written in a programming language that permits portability across different types of computer hardware. However, although software has become much more standardized, social scientists sometimes find it necessary in the course of their research to write their own programs. In addition, despite the increasing standardization of software, its intrinsic complexity is such that even standardized, widely used commercial software may occasionally yield wildly incorrect or inaccurate results.

```
function standard_deviation_2 (X: vector ) {
  variables x_dev, x_sum_sq, x_sum : real; n,
   count: integer;
  x_dev = 0; x_sum = 0; x_sum_sq = 0;
  n = length(X);
  for count = 1 to n {
    x_sum = x_sum + X[count];
    x_sum_sq = x_sum_sq + X[count]^2;
  }
  x_dev = sqrt((n*x_sum_sq - x_sum^2)/n^2);
  return (x_dev);
}
```

Figure 2 A single-pass algorithm for the standard deviation.

Algorithms, Computability, and Computational Tractability

What Is an Algorithm?

The idea of the algorithm is fundamental to software. An algorithm is a sequential set of steps that can be used to solve a well-defined problem. More strictly, an algorithm is a finite, deterministic set of instructions written in an abstract syntax that, when executed, completes in a finite amount of time and solves a specified problem. An algorithm is said to solve a problem (or to be effective) if and only if it can be applied to every instance of that problem and is guaranteed to produce an exact solution for each instance.

Consider the problem of computing the standard deviation of a population, $\sigma = \sqrt{\left(\sum(x-\bar{x})^2/n\right)}$. Pseudocode for one algorithm that computes this is shown in Fig. 1. A cursory examination will show that this algorithm is effective for any vector of real numbers and will complete in an amount of time roughly proportional to the length of the vector.

This is not, of course, the only algorithm that solves the problem. The mathematical expression could be expanded and rearranged to yield $\sigma = \sqrt{\left(n\sum x^2 - (\sum x)^2/n^2\right)}$, which computes the standard deviation in a single pass, without first computing the mean. Directly translating this expression into pseudocode yields a different algorithm, as shown in Fig. 2.

Algorithmically, the two methods of computing the standard deviation are similar in structure, and the execution time for either is a linear function of the length of x. As we shall see in the next section, however, straightforward implementation of each of these algorithms may differ greatly in real accuracy and speed.

As stated previously, the term algorithm, when not otherwise qualified, denotes a deterministic, finite set of steps guaranteed to produce results with well-defined

properties. Other categories of algorithms exist: approximation algorithms, randomized algorithms, and heuristic algorithms. These qualified classes of algorithms, especially heuristics, are used most frequently to approach problems for which no tractable deterministic, finite, effective algorithm is known.

Approximation algorithms produce results that are guaranteed to be within some formally defined distance (usually given as a relative measure) of the optimal solution to a problem. For example, approximation algorithms are sometimes used for the traveling salesperson problem (TSP), which is stated as follows: Given a list of cities and the cost of travel between each, find the cheapest route that visits each city once and returns to the starting point. No algorithm solves the TSP efficiently, and it can be proved that no approximations exist for the general TSP problem. However, if the cost of travel between cities satisfies the triangle inequality, approximation algorithms (such as the minimum spanning tree algorithm) exist that are guaranteed to yield solutions that are no more than twice the optimal cost. For other problems, approximation algorithms may exist that yield solutions arbitrarily close to the optimal solution. For example, when one uses a converging infinite series, such as a Taylor series, to approximate a function, one can reduce the approximation error of the algorithm as much as desired by adding more terms.

Randomized algorithms use nondeterministic steps and have a known probability of yielding correct answers. Randomized algorithms that can sometimes return incorrect results are called Monte Carlo algorithms, whereas Las Vegas algorithms may return an indication that a solution was not found but never return incorrect results. In contrast, heuristic algorithms, often known simply as heuristics, specify sets of steps but yield solutions that do not have well-known properties. In other words, heuristic algorithms yield solutions that are not known to be correct (or even approximately correct) but are thought to be often useful in practice.

Turing Machines

The Turing machine is an abstract representation of a computer introduced by Turing in 1936 to give a precise definition to the concept of the algorithm. It is still widely used in computer science, primarily in proofs of computability and computational tractability. Turing imagined a mechanical device that moved along an infinite length of recording tape, reading and modifying symbols on that tape in accordance with a fixed internal table of actions. As a Turing machine moves along a tape, it uses its table, in combination with the current input symbol, and the contents of its internal state register to determine the next action. The table indicates to the machine whether to modify the current symbol and/or state, and also whether to move forward or backward along the tape.

Mathematically, a Turing machine is a tuple: $M = (K, \Sigma, \delta, s)$, where K is a finite set of states, $s \in k$ is the initial state, Σ is a finite alphabet of symbols, and δ is a transition function that represents the "program" for the machine:

$$\delta: K \times \Sigma \longrightarrow \left(K \cup \{halt, accept, reject\}\right) \times \Sigma \times \{left, right, stay\}.$$

A string of symbols σ from the alphabet Σ represents the input to the Turing machine. To "execute" the machine, one applies δ to the first symbol in σ: $\delta_o = \delta(s, S_0)$ and uses the output to update σ and to provide the input for the next iteration of δ.

The Church–Turing thesis, in its most common form, states that every physically possible form of computation can be carried out by a Turing machine. This thesis is generally assumed to be true, and it has some important and useful implications: all computer languages that are Turing complete, which includes all common programming language, are equivalent in what they can compute: any computation possible in FORTRAN, for example, is possible (if not necessarily equally convenient) in any other language. If one can construct a proof of the (non)-computability of a particular problem, or of the effectiveness of an algorithm, that proof applies to all other physically possible forms of computation (including quantum methods).

Computability and Computational Complexity

A problem is said to be computable (or decidable) if and only if there exists an algorithm that solves the problem. Turing, in his 1936 paper, first proved the halting problem to be undecidable. Informally, the halting problem can be stated as follows:

> *Given a description of an arbitrary algorithm and its input, decide whether the algorithm halts (yielding an answer) or runs infinitely.*

Turing demonstrated that a direct consequence of the halting problem being undecidable is that there cannot be an algorithm that, given any statement about the natural numbers, determines that statement's truth. Subsequently, many other undecidable problems have been described, and the typical method of proof has been to show that a new problem reduces to the halting problem. Remarkably, two decades later, Rice showed that given any nontrivial property of a computer program (or mathematically, a partial function), the problem of determining whether that property applies to an arbitrary computer program is generally undecidable.

It is important to note that Turing's and Rice's proofs apply to the set of algorithms as a whole, not to all individual instances. It is certainly possible, for example, to prove the correctness of some computer programs, although it is impossible for an algorithm to be able to determine the correctness of any program given to it. For individual instances of algorithms, the central roles of algorithmic analysis are to determine the correctness and efficiency of individual algorithms and to characterize the difficulty of different classes of problems.

Computer scientists use computational complexity classes to characterize the difficulty of computable problems. A complexity class comprises a model/mode of computation (e.g., the deterministic Turing machine described previously), a resource we wish to bound (e.g., execution time or storage space), and a bounding function. A problem is said to be a member of the class if some algorithm exists that, using the specified mode of computation, can solve any instance of that problem using amounts of resources limited by the given bound. Bounding functions for execution time are conventionally denoted using "big O notion," $O(f(n))$, where n is the size of the problem. Additive and multiplicative constants are omitted because these vary with the computing model used. For example, $O(2^n)$ denotes that the time it takes to execute an algorithm grows exponentially as input grows (for worst-case instances).

There is an infinite number of possible complexity classes. Two classes, P and NP, are of particular interest because they are widely used as measures of computational tractability. Both P and NP are measures of time complexity, which is proportional (by construction) to the number of the instructions that an algorithm must execute to reach a solution. The bounding function is expressed in terms of the size of the problem-instance, which is defined as the number of parameters or items (of fixed size) in the instance. (Technically, NP applies to decision problems that yield true or false as an answer. However, other types of problems can easily be converted to decision problems to determine the complexity class.)

The class P is the set of problems for which algorithms exist that can solve any instance in polynomial time. Formally, the class NP is defined as the set of problems solvable in polynomial time by a nondeterministic Turing machine, which would automatically choose the correct answer from among a finite set of possible logic branches. (This is a useful mathematical construct but not physically possible.) This class of problems is widely thought to require exponential time to compute by any real computer, for at least one instance of the problem: $O(c^n), c > 1$. NP-complete problems are thought to be the most difficult problems in NP. A problem is NP-complete if it is in NP and if every other problem in NP is reducible to it.

A problem is said to be computationally tractable if it is in P. A problem is said to be computationally intractable (also referred to as computationally complex or computationally hard) if it is at least as difficult as a problem in NP. No polynomial-time algorithm is known to exist, using conventional computers, for any problem in NP.

The most common way to show that a particular type of problem is NP-hard is to show that another problem already known to be NP-hard can be reduced to it. There are many types of reduction techniques, and one particularly straightforward type is called the Karp reduction, or the polynomial-time many-one reduction. To use this technique, one finds a known NP-hard and a polynomial-time algorithm that converts any instance of that problem to an instance of the new problem. Since any algorithm that solves the new problem would also solve the intractable problem, the new problem must be at least as difficult as the intractable problem. To show that the problem is NP-complete, one proves that a known NP-complete problem can be reduced to the new problem and vice versa.

The use of membership in P and NP to characterize a problem as tractable or intractable, respectively, has two advantages. First, it is independent of any particular computer hardware design. Intractable problems cannot be made tractable through improvements in conventional hardware technology. Second, it is independent of any particular algorithm since it is the problem itself, not a specific algorithm, that drives the requirement of exponential time. Intractable problems cannot be made tractable through advances in software or algorithmic design.

There are, however, some limitations to this characterization of tractability. First, the distinction between tractable and intractable problems is most important for instances of large size, where the exponential factors in the time requirements of these problems become dominant. If problem A is solvable in $O(1.1^n)$ time and problem B is solvable in $O(n^{2000})$ steps, B is formally more tractable than A but is more difficult to solve in practice. Second, NP-completeness is a worst-case measure of complexity: Some problems in NP may have instances that can be solved in polynomial time, and it may even be the case that the average instance is solvable in polynomial time. Third, NP applies strictly to deterministic exact algorithms. An "intractable" problem may still be "solved" by a randomized algorithm that gives a solution with high probability or an approximation algorithm that is close to the desired solution. Fourth, a small number of problems in NP (but not NP-complete) are thought to be solvable efficiently with quantum computers, should such computers ever be constructed on a large enough scale. It is currently believed, however, that no physically possible quantum computer can compute NP-complete problems efficiently.

Despite these theoretical limitations, relatively few NP-hard problems have been found to be easier than

expected in practice; few have yielded to approximations, randomization, or quantum algorithm techniques; and few have been found easy in the average case. NP-completeness remains a powerful and widely used gauge of computational tractability.

Application of Complexity Theory: How Difficult Is It to Manipulate an Election?

One key problem for voting systems is the potential for voters to manipulate the results through strategic voting. A voter is said to act strategically when he or she casts a vote that does not reflect his or her true ranking over the choices but is calculated instead to achieve a favorable outcome. For example, a voter in a presidential election might prefer the Libertarian candidate to the Republican and Democratic candidates but casts a vote for the Republican candidate because he or she believes the Libertarian has a negligible chance of winning. A voting system is said to be nonmanipulable if it is not possible for any voter to gain from strategic voting.

Strategic voting has been studied extensively in political science and economics. A powerful negative result, discovered by Gibbard and Satterthwaite independently in the early 1970s, is that any nondictatorial voting scheme is manipulable (for elections with at least three candidates). Like Arrow's theorem, on which Satterthwaite's proof drew, this impossibility result engendered some pessimism regarding the design of electoral systems.

In 1989, Bartholdi *et al*. used computational complexity theory to show that under some voting systems effective strategic voting is NP-hard. Thus, these elections systems are, if not manipulation-proof, at least manipulation-resistant. In using complexity theory to inform social choice theory, they initiated the study of the computational properties of electoral systems.

Bartholdi *et al*.'s proof is too long to present here, but another proof in the same vein is instructive. This proof, which comes from previous work by the author, shows that the problem of constructing "optimal" election districts is NP-hard. The following portion shows that constructing a district plan having the minimum possible population deviation among them from discrete nonuniform census blocs, as the law requires, is NP-hard:

1. Note that each of the n census blocs, c_i, must be assigned to one and only one of k election districts, d_j, and that the population of the district is simply the sum of all census blocs that comprise it:

$$Population = P(d_i) = \sum_{j \in d_i} c_j.$$

2. Define the measure of population deviation for a redistricting plan to be the difference in population between the most and least populous district:

$$\textit{Population deviation score} = \mathrm{PD}(\mathbf{c}) = \max_{i \in k}\left(\sum_{j \in d_i} c_j\right) - \min_{i \in k}\left(\sum_{j \in d_i} c_j\right).$$

3. The optimal districting plan is thus a division (strictly a partition) of the n census blocs into k districts such that PD is minimized.

4. Finally, we show that we can convert any instance of 3-partition,a known NP-complete problem, into the redistricting problem:

3-partition

Input: Set A of $3m$ elements, a bound $B \in Z^+$ and a size $s(a) \in Z^+$ for each $a \in A$ such that $B/4 < s(a) < B/2$ and such that $\sum_{a \in A} s(a) = mB$.

Solution: Determine whether A can be partitioned into m disjoint subsets such that for $1 \leq i \leq m$, $\sum_{a \in A_i} s(a) = \mathrm{B}$.

To convert an instance of 3-partition into an instance of optimal redistricting,

a. For each element $a_i \in A$, create an artificial census bloc with a population equivalent in size, $c_i = s(a_i)$.

b. Take the solution for the optimal redistricting problem, PD*, using the artificial census blocs created in step (a). If $\mathrm{PD}^* = 0$, the answer to the corresponding 3-partition is "true"; otherwise, the answer is false.

Thus, any algorithm that solves the optimal redistricting problem can also be used to solve 3-partition. Since the 3-partition problem is computational intractable, optimal redistricting is computationally intractable as well.

Implementation: Bugs, Verification, and Accuracy

Although algorithms are at the conceptual core of all software, computers execute not algorithms but programs—implementations of an algorithm written in some real programming language, and executing within a particular computing environment. The same algorithm may be expressed using various computer languages, use varying encoding schemes for variables and parameters, rely on arithmetic operators with varying levels of accuracy and precision in calculations, and run on computers with varying performance characteristics. Three problems can occur in the gap that arises between the formal algorithm and its actual implementation: bugs, inaccuracies, and performance bottlenecks.

Bugs and Verification

Any computer program of reasonable size is sure to have some programming errors or "bugs," and there is always a possibility that these errors will affect research results. In practice, software used in research will be tested but not proven correct. As Dahl *et al*. famously noted in 1972, program testing can be used to show the presence of bugs but never to show their absence. It has also been observed generally that even software that worked well when first written tends to encounter problems as changes occur to the computing environment within which it runs, such as the operating system, system libraries, and computing hardware. (This phenomenon is known colloquially as "software rot".)

Errors in mathematically-oriented programs are often particularly difficult to detect since the software may return plausible results rather than failing entirely. Well-replicated studies of experienced spreadsheet programmers performing standardized tasks have demonstrated that undetected bugs are the rule, not the exception. Although extensive testing will substantially reduce the rate of errors that remain in the final version of a piece of software, much caution is still warranted when creating one's own software.

In limited circumstances, it is possible to prove software correct, but it is exceedingly unlikely that any particular software package used by social scientists will have been subject to formal methods of verification. Until recently, in fact, such formal methods were widely viewed as completely impractical by practitioners, and despite increasing use in secure and safety critical environments, usage remains costly and restrictive.

Computer Arithmetic, Numerical Accuracy, and Stability

Mathematicians, social scientists, and other humans perform arithmetic symbolically: computers do not. The difference between symbolic and computer arithmetic can lead to inaccuracies, and to avoid these inaccuracies we need to understand how computers do math. All computer hardware, and practically all software, performs arithmetic by representing every number as a fixed-length sequence of 1s and 0s, or bits, b. Integers are often represented as a single sequence of bits, each representing a different power of two, with a single bit indicating the sign. Under this representation, arithmetic on integers operates according to the "normal" (symbolic) rules of arithmetic, as long as the integer operands nor the results are too large ($>2^{b-1}-1$), leading to an (possibly undetected) overflow error. For example, usually $b=32$, so the number 2147483648 ($1+2^{32-1}-1$) would overflow and may actually roll over to -1.

Real numbers are represented using floating point arithmetic. Floating point numbers are represented by two sequences of bits, with one sequence representing a mantissa (m) and the other representing an exponent (e): $\pm m \times 10^e$. An additional bit indicates the sign. The specific details of floating point arithmetic operations vary across different computing platforms. (The algorithms for performing floating point arithmetic encompass some subtle technical details, which are beyond the scope of this article.)

All floating point representations are subject to overflow, underflow (when the true number is smaller than the smallest value capable of being represented), and rounding errors. Rounding and other numerical problems can lead to inaccurate results, even when every step of an algorithm is correctly followed. (The accuracy of the solution is, roughly, the distance between the results that are actually produced and the correct answers when computed using infinite precision.) One source of rounding error arises directly from storing data in this representation: some numbers cannot be exactly represented using this scheme. An example is the number 0.1, which has an infinitely repeating binary representation using this technique. The infinitely repeating floating point transformation of "0.1" must be rounded to m bits, resulting in a slight loss of accuracy when performing subsequent calculations. A second source of rounding error occurs when a number is added to (or subtracted from) a very much smaller number. This type of rounding error can occur even when both operands are exactly represented. In the extreme case, the result simply rounds to the very large number.

Underflow, overflow, and rounding have many implications for accurate computing. One implication is that summations are more accurate when performed on a list of elements that is sorted in order of increasing magnitude. Therefore, the algorithm in would produce a more accurate result if we modified it as in Fig. 1. Standard proofs of the correctness of particular algorithms usually ignore the underlying arithmetic implementation and the effects of rounding.

The limits of computer arithmetic, and the variations in it across different platforms, have three implications for replicability and accuracy. First, a computer program can produce different answers when run on different computers, run on the same computer using a different operating system, or recompiled with different options.

```
function Standard_deviation_3 (X: vector ){
  variables X_sort: vector;
  X_sort = sort_least_to_greatest(X);
  return (Standard_deviation (X_sort));
}
```

Figure 3 A more accurate standard deviation.

Second, numeric errors can accumulate within an algorithm or can (rarely) cancel each other. Third, inaccuracies in floating point arithmetic can interfere with the formal mathematical properties of elementary and nonelementary functions. For example, the associative law of arithmetic does not hold universally for computer arithmetic: Where $\oplus$ is the floating point addition, $(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$. Practically, floating point inaccuracies can cause a range of problems, from small inaccuracies in the results to results that are returned without error but are completely inaccurate or even failure of an otherwise correct algorithm to halt. Careful analysis of the accuracy of each numerical implementation, in its entirety, is necessary in order to ascertain the level of accuracy that can be associated with a particular solution.

Surprisingly, accuracy alone is not enough to ensure that implementations of algorithms produce usable results. Because of the rounding of data when stored initially, and because of the possibility of measurement error in much of social science, a reliable implementation must be numerically stable as well as accurate. An algorithm for computing a function is said to be numerically stable if small errors in input cause only small errors in output; that is, $\hat{y} + \Delta y = f(x + \Delta x)$, where x is the true input, $\hat{y}$ is the true output, and $\hat{y} + \Delta y$ is the computed value. Δx is error that enters into computations through, for example, converting a decimal number into a binary number with a finite degree of precision. Less formally, a stable algorithm gives "almost the right answer to almost the right problem."

Performance Tuning and Bottlenecks

Algorithms determine performance at the grand scale. For sufficiently large values of n, an $O(n)$ sort algorithm will finish before an $O(n^2)$ algorithm. Nevertheless, implementation is important: It is not uncommon for a well-tuned implementation of a particular algorithm to run an order of magnitude faster and use an order of magnitude less resources than a naive implementation of the same algorithm. This arises most commonly because of performance bottlenecks. In all computing architectures, executing a program involves accessing implicitly a variety of heterogeneous resources, such as floating point units for arithmetic calculations, memory chips and storage devices for access to data, and networks and busses for communication. These resources have different performances characteristics, and a bottleneck may occur when the calculations of the program are interrupted to wait for some slower (or temporarily unavailable) computer resource. It is rare that formal algorithmic analysis delves down to this level of detail.

Programmers and compilers can use profiling tools to analyze the empirical behavior of a particular computer program by monitoring its performance and the resources its uses as it runs. By rearranging the order of operations, changing the pattern of access to data, and substituting note equivalent (but more efficient) sets of computing instructions, the program can be made to run faster. Low-level performance tuning, however, is often tied closely to a particular hardware configuration, programming language, and operating environment, and is at odds with portability, clarity, and maintenance. Moreover, it is usually counterproductive to tune a program without an empirically generated profile of its performance. Thus, performance tuning is best done after the software is designed, written, and tested.

Application: Benchmarking a Statistical Package

Consider the two algorithms for computing the standard deviation discussed previously. If the vector $\mathbf{X}$ is very large, it may be expensive to read values from it. In this case, despite both algorithms having approximately equivalent time complexity, the program implementing the single-pass algorithm in Fig. 2 could be much faster than a program implementing the algorithm in Fig. 1, which requires that every element of $\mathbf{X}$ be read twice. However, the algorithm in Fig. 2 is, in practice, much more susceptible to rounding errors when $n\sum x^2$ and $(\sum x)^2$ are both large. (In this case, the tradeoff between performance and accuracy is avoidable; more accurate one-pass algorithms exist.)

In practice, programs are often so complex that numerical benchmarks, rather than formal analysis, are used to gauge their accuracy. Numerous benchmarks are available for testing simple statistical functions and models, such as univariate descriptive statistics, cumulative distribution functions, linear regression, analysis of variance, and nonlinear regression. One particularly popular and useful set, the Statistical Reference Datasets (StRD), is maintained by the National Institute of Standards and Technology (NIST).

Each StRD problem is composed of either data taken from published research or data specially generated to stress computational capabilities. For each problem, the data are accompanied by values certified by NIST to be correct. These values are obtained analytically where possible, or by using supercomputers to compute approximate results using two independent algorithms and exceptionally high-precision floating point arithmetic.

As an example of a benchmark, consider a simple univariate descriptive statistics problem. Compute the mean, standard deviation, and one-observation lag autocorrelation for the first n digits of π. The StRD provides both the input data and the result values. These have been calculated on a supercomputer using very high-precision arithmetic and the results rounded to 15 significant digits.

```
> options(digits=22)                                    # set number of digits to display
> x<-read.csv(file=''PiDigits.txt'' header=TRUE)        # read test data
> x[1:10,1]                                             # check data, digits of pi
[1] 3 1 4 1 5 9 2 6 5 3
> xm<-mean(x[[1]])                                      # calculate mean of data
> xm
[1] 4.534799999999720046
> xsd<-sd(x[[1]])                                       # calculate standard deviation
> xsd
[1] 2.867543699108215271565
> xacf<-acf (x[[1]], lag.max=1, type=c(''correlation''),plot=FALSE)
> xac<-xacf [[1]][2]                                    # calculate autocorrelation
> xac
[1] -0.003683261271785385603666
> -log(abs((xm-4.5348)/4.5348))                         # calculate LRE for mean, sd, ac
[1] Inf
> -log(abs( (xsd-2.86733906028871)/2.86733906028871))
[1] 35.0175962904394
> -log(abs( (xac+0.00355099287237972)/-0.00355099287237972))
[1] 34.8496905126905
```

Figure 4 Testing a statistics package for accuracy by using the digits of pi.

The resulting values are said to be certified by the benchmark.

To gauge the reliability of a particular statistical package using the StRD, one loads the data into the program, runs the analysis, and compares the results generated by the software to the certified values provided by NIST. Although no benchmarking method can prove that a piece of software is accurate, performing well on the StRD benchmark provides evidence that a software package is accurate for that domain of problems.

The transcript in Fig. 4 shows the results of running the NIST "Pi-digits" benchmark in a popular statistics package. In this example, we calculate the number of correct digits in the results produced by the statistical software using the log relative error (LRE). More formally, the LRE is:

$$-\log_{10}\left(\left|\frac{result-certified}{certified}\right|\right), \quad certified \neq 0.$$

The previously discussed statistics package agreed with the certified benchmark results to at least 34 digits. Since the accuracy of the benchmark is only certified to 15 digits, we can infer that the statistical package was accurate to at least 15 digits for these calculations.

Software Development

Developing software remains a complex and difficult activity. Many practitioners suggest that this complexity is unavoidable. The complexity of the problem domains to which software is applied, the inherent malleability of software, and the problems that characterize the behavior of discrete systems with large state-spaces all contribute to the overall difficulty of developing software. Regardless of its source, the difficulty of writing software has a number of important consequences, and much of the work in the field of software engineering is devoted to managing the complexity of software and the risks that result from such complexity.

Implications of Complexity for Programming

One well-known consequence of software's complexity is that there is wide variation in the quantity and quality of work produced by individual programmers: Early findings by Sackman *et al.* found differences of more than 20 to 1 in the time required by experienced programmers to solve the same problem. This result has been widely replicated, with findings of large differences in productivity and defect rate across programmers and programming teams.

A second consequence of the software's complexity is that as the desired functionality of a program increases, a monolithic program providing that functionality eventually becomes too complex to understand and manage. As a consequence, most programming techniques seek to manage the complexity of programs by decomposing them into smaller, simpler components. All

decompositions attempt to form abstractions to represent each of the components so that some details about the implementation can be encapsulated, or hidden from, the other components. For each component, a set of interfaces are created that define the way each component is used, so that components can be integrated to form the program. If the decomposition is successful, the implementation of each component can be changed over time and can later be adapted to different computing needs. Also, as long as the interface behavior is preserved, the program will continue to function correctly as a whole.

The first generation of programming languages, such as FORTRAN I, was designed to support decomposition of programs into mathematical expressions. It was quickly realized that this method of decomposition was too limiting. Modern programming languages and methods typically support one of five strategies for decomposing problems. Procedure-oriented methods aim to decompose programs directly into sets of algorithms. Object-oriented methods aim to decompose programs into classes, objects, and behaviors, the last of which encapsulates the specific problem-solving algorithms used. Logic-oriented, rule-oriented, and constraint-oriented methods aim to decompose programs into sets of goals, if–then rules, and constraints, respectively. These three methods use generalized algorithms to solve for or optimize against the specific sets of goals, rules, or constraints.

For example, consider the sort_least_to_greatest() routine in Fig. 3. This function is an example of a procedure-oriented decomposition of a problem. The decomposition of our program separates the sorting algorithm from the algorithm used to compute the standard deviation. The implementation of the sort procedure is unspecified, but it could be any one of a wide variety of sort algorithms, depending on the programmer's desire to save space, time, or effort. Whichever algorithm is used, the program as a whole remains correct as long as the same procedural interface is provided with each algorithm.

There is no method of decomposition that is universally better for all applications. For example, logic-oriented programming techniques and languages are considered to be particularly well suited for some applications in artificial intelligence. For general applications, however, the procedure and object-oriented methods are most commonly used, and modern best practices and programming languages emphasize the object-oriented approach. Moreover, research is active in the area of extending and augmenting the object-oriented paradigm: Techniques such as component-based programming and design patterns provide methods for grouping objects into larger abstractions, and new paradigms such as aspect-oriented programming aim to augment object-oriented approaches with alternative, concurrent decomposition strategies.

Software Development Life Cycle

Caution is warranted when a software program is large enough to involve more than one author. In 1975, Brooks discussed the widespread difficulties of software development and formulated his well-known, and well-studied, "law": "Adding manpower to a late software project makes it later." This results from the fact that adding additional programmers often increases the communication and coordination costs involved in a project faster than it decreases the remaining work.

Surveys of practice in industry continue to demonstrate the difficulties of software development. A significant minority of projects are never completed, and the vast majority of the remainder finish significantly over budget, long past the original deadlines, and often with reduced or impaired functionality. Although the exact percentages are debated, it is widely recognized that the majority of software projects of moderate size fall short in some serious way.

Projects falter for many reasons, but the two causes most commonly diagnosed are poor schedule estimation, which is nearly universal, and problems deriving from excessive, changing, unclear, or incomplete requirements. Models for development have been proposed to address these common problems. The first model for developing software, the "waterfall" model, was articulated (although not advocated in its pure form) by Royce in 1970 and quickly gained dominance. It advocated that software development proceed in five phases: (i) requirements specification, in which the functionality of the software is described in detail; (ii) design, in which the overall structure of the software is designed and categorized into subcomponents with well-defined interfaces; (iii) implementation, in which the code for each component is written and tested individually; (iv) integration, in which individual components are integrated into a complete system that is then tested together; and (v) operation and maintenance, in which the software is delivered to the customer, modified to meet changing requirements, and repaired as bugs are discovered.

Although the waterfall model can still be found in modern use, it is now regarded as somewhat naive. A decade after the model was developed, it was widely criticized as it became clear from experience that it did not adequately address either design risks or requirements risks. Risks that requirements were not properly anticipated, were misunderstood, or needed to change over the course of the project and risks that the design of the project was flawed, incomplete, or misunderstood in implementation. Recognition of these risks has led many to abandon the waterfall model in favor of incremental and iterative models of software development, in which multiple versions of a project are developed and delivered to users over the life cycle of the project. Rapid development, testing,

and delivery of smaller (and possibly incomplete) increments, prototypes, and/or versions of a software product allow for the ongoing incorporation of feedback from users into the requirements and design phases of future increments. Although there are ongoing and vigorous debates about which processes are best, it is widely recognized that some form of incremental approach is necessary to successfully complete a project.

Software Distribution

After software is created, it is most often distributed in one of two forms. Software can be distributed as source code, which provides the high-level, human-readable set of instructions comprising the software, or the source code can be compiled (converted) into low-level instructions, commonly known as object code, which can be distributed alone and executed directly by the computer.

Object code is difficult, although not impossible, for humans to inspect or modify, and it inherently obscures the design and algorithms used in the software. Distribution of software as object code is the norm for commercial software products because it helps to protect the intellectual property embodied in the software. In addition, this intellectual property is protected by copyright law and often by some combination of patent, trademark, and trade secret law. Under current law, large civil and/or criminal penalties can be levied against those who make illicit use of source code or even simply reverse engineer the object code.

On the other hand, source code distribution, when done properly, is conducive to software reuse and extension. Source code distribution is the method of choice in academic research and noncommercial projects because distribution of the source code enables others to learn from and potentially improve the software. Moreover, an innovative family of licenses, known as "open source" licenses, have been designed to provide an incentive for others to learn from and improve software by guaranteeing that the software may be freely used for any purpose (including commercial purposes) and that any future modifications or extensions of the software by anyone will also be freely available in source form. The open source license parallels the academic norm of openness, which requires that publicly recognized (published) research be replicable so that others may verify and extend it.

What Software to Write and What Software to Use?

In most research, especially that involving standard methodologies, it is usually more appropriate to use a standard software package than to write one's own. Since all software contains bugs and can produce inaccurate results in some circumstances, one should choose a package that is as open to inspection as possible. The software should document the algorithms used, especially those relevant to data processing and analysis. The documentation should also include information regarding the expected range of inputs for each algorithm and the accuracy of the results within this range, and it should explain the warnings or errors that the software will produce when it encounters problems. Software packages that provide source have an advantage in this area because users can inspect the code directly. However, the availability of source code is not a substitute for thorough documentation.

Choose code that is also well tested, particularly for the tasks for which you intend to use it. The developers of the software should have clearly explained the methodology used to test their software and provide a complete record of all changes to the software, including previously reported bugs and subsequent fixes. Mathematical programs should document the accuracy of any functions and routines available to the users and provide test results using standard benchmarks.

Sometimes, however, one may not be able to find well-documented and thoroughly tested software that uses algorithms appropriate for one's problem. When this occurs, one must weigh carefully the potential for inaccuracies or inefficiencies to arise from applying an algorithm to a problem for which it is not well suited against the considerable effort required to develop software and the prevalence of bugs and inaccuracies in newly written software.

See Also the Following Articles

Computer Simulation • Computer-Based Mapping • Computer-Based Testing • Computerized Adaptive Testing • Computerized Record Linkage and Statistical Matching

Further Reading

Altman, M., Gill, J., and McDonald, M. (2003). *Numerical Issues in Statistical Computing for the Social Scientist.* Wiley, New York.

Bentley, J. (1982). *Writing Efficient Programs.* Prentice Hall, New York.

Booch, G. (1994). *Object-Oriented Analysis and Design with Applications.* Addison-Wesley, Reading, MA.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1990). *Introduction to Algorithms.* MIT Press, Cambridge, MA.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns.* Addison-Wesley, Reading, MA.

Hennesy, J. L., Patterson, D. A., and Goldberg, D. (2002). *Computer Architecture: A Quantitative Approach*, 3rd Ed. Morgan Kaufman, New York.

Higham, N. J. (2002). *The Accuracy and Stability of Numerical Algorithms,* 2nd Ed. SIAM, Philadelphia.

Knuth, D. E. (1998). *The Art of Computer Programming*, 2nd Ed. Vols. 1–3, Addison-Wesley, Reading, MA.
McConnell, S. C. (1996). *Rapid Development: Taming Wild Software Schedules.* Microsoft, Redmond, WA.
Papadimitriou, C. H. (1994). *Computational Complexity.* Addison-Wesley, Reading, MA.
Royce, W. (1998). *Software Project Management: A Unified Framework.* Addison-Wesley, Reading, MA.
Skienna, S. S. (1997). *The Algorithm Design Manual.* Springer-Verlag, New York.

Spatial Autocorrelation

Daniel A. Griffith
University of Miami, Coral Gables, Florida, USA

Glossary

auto-model A statistical model whose associated probability density/mass function contains a linear combination of the dependent variable values at nearby locations.
correlation A description of the nature and degree of a relationship between a pair of quantitative variables.
covariance matrix A square matrix whose diagonal entries are the variances of, and whose off-diagonal entries are the covariances between, the row/column labeling variables.
estimator A statistic calculated from data to estimate the value of a parameter.
geographic connectivity/weights matrix An n-by-n matrix with the same sequence of row and column location labels, whose entries indicate which pairs of locations are neighbors.
geostatistics A set of statistical tools used to exploit spatial autocorrelation contained in georeferenced data usually for spatial prediction purposes.
Moran scatterplot A scatterplot of standardized versus summed nearby standardized values whose associated bivariate regression slope coefficient is the unstandardized Moran coefficient.
semivariogram plot A scatterplot of second-order spatial dependence exhibited in georeferenced data.
spatial autoregression A set of statistical tools used to accommodate spatial dependency effects in conventional linear statistical models.
spatial statistics A recent addition to the statistics literature that includes geostatistics, spatial autoregression, point pattern analysis, centrographic measures, and image analysis.

Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional surface, introducing a deviation from the independent observations assumption of classical statistics.

Introduction

Social scientists often study the form, direction, and strength of the relationship exhibited by two quantitative variables measured for a single set of n observations. A scatterplot visualizes this relationship, with a conventional correlation coefficient describing the direction and strength of a straight-line relationship of the overall pattern. A variant of conventional correlation is serial correlation, which pertains to the correlation between values for observations of a single variable according to some ordering of these values. Its geographic version is spatial autocorrelation (*auto* meaning self), the relationship between a value of some variable at one location in space and nearby values of the same variable. These neighboring values can be identified by an n-by-n binary geographic connectivity/weights matrix, such as $\mathbf{C}$: If two locations are neighbors, then $c_{ij} = 1$, and if not, then $c_{ij} = 0$ (see Fig. 1, in which two areal units are deemed neighbors if they share a common nonzero length boundary).

Positive spatial autocorrelation means that geographically nearby values of a variable tend to be similar on a map: high values tend to be located near high values, medium values near medium values, and low values near low values. Most social science variables tend to be moderately positively spatially autocorrelated because of the way phenomena are geographically organized. Demographic and socioeconomic characteristics such as population density and house prices are examples of variables exhibiting positive spatial autocorrelation. Neighborhoods tend to be clusters of households with similar preferences. Families tend to organize themselves in a way that concentrates similar household attributes on a map—creating positive spatial autocorrelation among many variables—with government policies and activities, such as city planning and zoning, reinforcing such patterns.

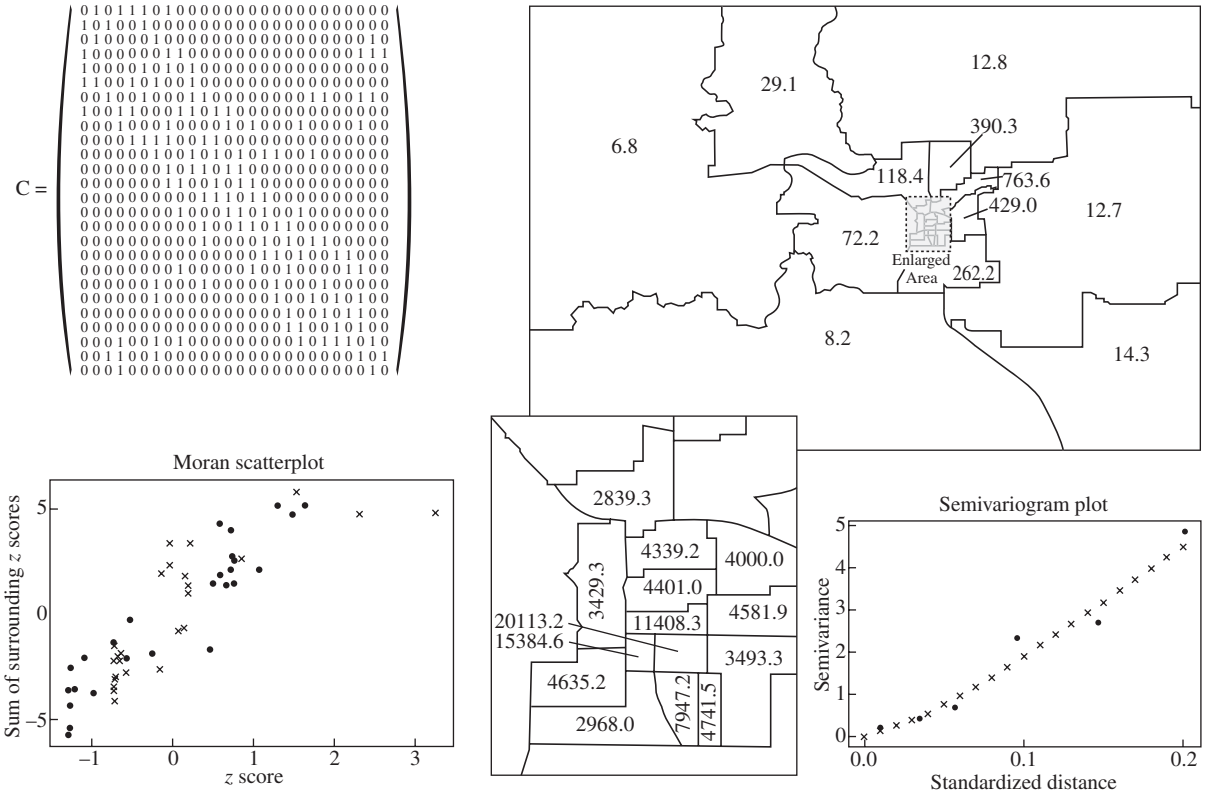


Figure 1 Adair County, Missouri, 1990 population density and census block areal units. (Top left) Binary geographic connectivity matrix C. (Top right) Geographic distribution of population density (with Kirkville as an inset). (Bottom left) Moran scatterplot for population density (x) and LN (population density + 164) (•). (Bottom right) Semivariogram plot for LN (population density + 164) (•) and its Bessel function predicted values (x).

Why Measure and Account for Spatial Autocorrelation?

Spatial analysis frequently employs model-based statistical inference, the dependability of which is based on the correctness of posited assumptions about a model's error term. One principal assumption states that individual error terms come from a population whose entries are thoroughly mixed through randomness. Moreover, the probability of a value taken on by one of a model's error term entries does not affect the probability of a value taken on by any of the remaining error term entries (i.e., the independent observations assumed in classical statistics). Nonzero spatial autocorrelation in georeferenced data violates this assumption and is partly responsible for geography existing as a discipline. Without it, few variables would exhibit a geographic expression when mapped; with it, most variables exhibit some type of spatial organization across space. Zero spatial autocorrelation means geographically random phenomena and chaotic landscapes.

Therefore, there are two primary reasons to measure spatial autocorrelation. First, it indexes the nature and degree to which a fundamental statistical assumption is violated and, in turn, indicates the extent to which conventional statistical inferences are compromised when nonzero spatial autocorrelation is overlooked. Autocorrelation complicates statistical analysis by altering the variance of variables, changing the probabilities that statisticians commonly attach to making incorrect statistical decisions (e.g., positive spatial autocorrelation results in an increased tendency to reject the null hypothesis when it is true). It signifies the presence of and quantifies the extent of redundant information in georeferenced data, which in turn affects the information contribution each georeferenced observation makes to statistics calculated with a database. Accordingly, more spatially autocorrelated than independent observations are needed in calculations to attain an equally informative statistic.

Second, the measurement of spatial autocorrelation describes the overall pattern across a geographic landscape, supporting spatial prediction and allowing detection of striking deviations. In many situations, spatial prediction is as important as temporal prediction/forecasting. Explicitly accounting for it tends to increase the percentage of variance explained for the dependent variable of a predictive model and does a surprisingly

good job of compensating for unknown variables missing from a model specification. Exploiting it tends to increase the R^2 value by approximately 5%, and obtaining 5% additional explanatory power in this way is much easier and more reliably available than getting it from collecting and cleaning additional data or from using different statistical methods.

Graphical Portrayals of Spatial Autocorrelation

By graphically portraying the relationship between two quantitative variables measured for the same observation, a scatterplot relates to the numerical value rendered by a correlation coefficient formula. Not surprisingly, specialized versions of this scatterplot are closely associated with measures of spatial autocorrelation.

The Moran scatterplot is one such specialized version. To construct it, first values of the georeferenced variable under study, Y, are converted to z scores, z_Y. Next, the adjacent or nearby z score values of Y are summed; this can be achieved with the matrix product $\mathbf{CZ}_Y$, where $\mathbf{Z}_Y$ is the vector concatenation of the individual z_Y values. Finally, the coordinate pairs $(z_{Y,i}, \sum_{j=1}^{n} c_{ij} z_{Y,1})$, $i = 1, 2, \ldots, n$, are plotted on the graph whose vertical axis is $\mathbf{CZ}_Y$ and whose horizontal axis is $\mathbf{Z}_Y$. This construction differs from that proposed by Anselin in 1995, who uses matrix $\mathbf{W}$, the row-standardized stochastic version of matrix $\mathbf{C}$, to define the vertical axis. An example of this graphic illustrates a case of positive spatial autocorrelation (Fig. 1).

Another specialized scatterplot is the semivariogram plot. To construct it, first, for each pair of georeferenced observations both the distance separating them and the squared difference between their respective attribute values are calculated. Next, distances are grouped into G compact ranges preferably having at least 30 paired differences, and then group averages of the distances and of the squared attribute differences are computed. Semivariance values equal these squared attribute differences divided by 2. Finally, on a graph whose vertical axis is average semivariance and whose horizontal axis is average distance, the following coordinate pairs are plotted:

$$\left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} (y_i - y_j)^2}{2K_g}, \quad \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} d_{ij}}{K_g} \right),$$

where K_g is the number of i–j pairs in group g, $\sum_{g=1}^{G} K_g = n(n-1)$, d_{ij} is the distance separating locations i and j, and δ_{ij} is a binary 0/1 variable denoting whether or not both locations i and j belong to group g. The steep slope in Fig. 1 indicates very strong positive autocorrelation that is due, in part, to a geographic trend in the data.

Autoregressive and Geostatistical Perspectives on Spatial Autocorrelation

Treatments of georeferenced data focus on either spatial autocorrelation (addressed in geostatistics) or partial spatial autocorrelation (addressed in spatial autoregression). The two classic works reviewing and extending spatial statistical theory are by Cliff and Ord, who have motivated research involving spatial autoregression, and Cressie, who has summarized research involving geostatistics. These two subfields have been evolving autonomously and in parallel, but they are closely linked through issues of spatial interpolation (e.g., the missing data problem) and of spatial autocorrelation. They differ in that geostatistics operates on the variance–covariance matrix, whereas spatial autocorrelation operates on the inverse of this matrix. More superficial differences include foci on more or less continuously occurring attributes (geostatistics) versus aggregations of phenomena into discrete regions (i.e., areal units) (spatial autoregression) and on spatial prediction (geostatistics) versus enhancement of statistical description and improvement of the inferential basis for statistical decision making (i.e., increasing precision) (spatial autoregression). Griffith and Layne, among others, present graphical, numerical, and empirical findings that help to articulate links between geostatistics and spatial autoregression.

Definition of Notation

One convention employed here denotes matrices with bold letters; another denotes names by subscripts. Definitions of notation used throughout appear in Table I.

Conceptual Meanings of Spatial Autocorrelation

Spatial autocorrelation can be interpreted in different ways. As a nuisance parameter, spatial autocorrelation is inserted into a model specification because its presence is necessary for a good description, but it is not of interest and only "gets in the way" of estimating other model parameters. In fact, if the value of this parameter were known, resulting statistical analyses would be much simpler and more powerful. Nonspatial analysts especially view spatial autocorrelation as an interference. They study the relationship between two quantitative variables that happen to be georeferenced, with spatial autocorrelation lurking in the background. Mean response and standard errors improve when spatial autocorrelation is accounted for, whereas conventional statistical theory could be utilized if the value of this parameter were known or set to 0. Ignoring latent spatial autocorrelation

Table I Symbols Used

AR	Abbreviation for the autoregressive response model
BB, BW, WW	Join count statistics: respectively, the number of neighboring ones, ones with zeroes, and zeroes
CAR	Abbreviation for the conditional autoregressive model
d	Distance separating two locations
det	Determinant of a matrix
$E(X)$	Expectation of the random variable X
exp	Inverse of the natural logarithm
Γ	Semivariance
G	The number of location pair groups in a semivariogram plot
GR	Abbreviation for the Geary ratio index
K_g	Number of distance pairs in group g for a semivariance plot
LN	Natural logarithm
MC	Abbreviation for the Moran coefficient index
MCMC	Abbreviation for Markov chain Monte Carlo
n	Number of locations in a georeferenced sample
n^*	Equivalent number of independent locations in a georeferenced sample
OLS	Abbreviation for ordinary least squares
s^2	Conventional sample variance
SAR	Abbreviation for the simultaneous autoregressive model
σ^2	Population variance
VIF	Abbreviation for the variance inflation factor
α_i	Population mean response
c_{ij}	Row i and column j entry of matrix $\mathbf{C}$
δ_{ij}	Binary 0–1 variable indicating membership of distance between locations i and j in semivariance grouping
$\lambda_{\max}$	Maximum eigenvalue of a matrix
$\lambda_{\min}$	Minimum eigenvalue of a matrix
$MC_{\max}$	Maximum possible Moran coefficient value
μ_y	Population mean of variable Y
$n_k!$	Factorial calculation for the number of entries in the kth group
ρ_j	Spatial autoregressive parameter for model j
Y_i	Value of variable Y for the ith observation
$z_{Y,i}$	z score of variable Y for the ith observation
$\mathbf{1}$	N-by-1 vector of ones
$\mathbf{ß}$	P-by-1 vector of regression parameters
$\mathbf{b}$	P-by-1 vector of regression parameter estimates
$\mathbf{C}$	N-by-n geographic weights matrix
$\mathbf{CZ}_Y$	Matrix summation of neighboring z scores of variable Y
ε	N-by-1 vector of random error terms
$\mathbf{E}_j^*$	jth eigenvector of matrix $(\mathbf{I} - \mathbf{11}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{11}^T/n)$
$\mathbf{I}$	N-by-n identity matrix
$\mathbf{V}\sigma^{-2}$	N-by-n inverse-covariance matrix
$\mathbf{W}$	Row-standardized version of the n-by-n geographic weights matrix
$\mathbf{X}$	N-by-p matrix of predictor variables
$\mathbf{Z}_Y$	N-by-1 vector of z scores for variable Y
$(y_i - \bar{y})/s_Y$	z score for the ith value of variable Y
$\langle \mathbf{1}^T\mathbf{C} \rangle_{\text{diagonal}}$	N-by-n diagonal matrix with diagonal entries of $\sum_{j=1}^{n} c_{ij}$
$TR(\mathbf{V}^{-1})/n$	Equation for the variance inflation factor
$n \times TR(\mathbf{V}^{-1})/\mathbf{1V}^{-1}\mathbf{1}$	Equation for the equivalent number of independent locations
$n^2/[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{11}^T\mathbf{V1}]$	Equation for measuring the efficiency of OLS estimators in the presense of nonzero spatial autocorrelation
$(\mathbf{I} - \mathbf{11}^T/n)$	Projection matrix that centers vector $\mathbf{Y}$
$(\mathbf{I} - \mathbf{11}^T/n)\mathbf{C}(\mathbf{I} - \mathbf{11}^T/n)$	Modified connectivity matrix appearing in the Moran coefficient numerator

continues

Table I *continued*

$\sum_{j=1}^{n} c_{ij}$	Sum of the ith row entries of matrix **C**
$\sum_{j=1}^{n} c_{ij} z_{Y,j}$	Sum of neighboring z score values, $z_{Y,i}$
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n$	Conventional sample covariation between variables X and Y
$\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}$	Conventional sample standard deviation of variable Y
$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}$	Sum of the cell entries of matrix **C**
$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - \bar{y})(y_j - \bar{y})$	Covariation of neighboring Y values
$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - y_j)^2$	Sum of squared differences of neighboring Y values
$\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij} d_{ij}/K_g$	Average distance for a given semivariance distance grouping
$\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_{ij}(y_i - y_j)^2/(2K_g)$	Semivariance estimate for a given distance grouping

results in increased uncertainty about whether findings are attributable to assuming zero spatial autocorrelation (i.e., misspecification).

As self-correlation, spatial autocorrelation is interpreted literally: Correlation arises from the geographic context within which attribute values occur. As such, it can be expressed in terms of the Pearson product moment correlation coefficient formula, but with neighboring values of variable Y replacing those of X:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}}$$

becomes

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - \bar{y})(y_j - \bar{y})/\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n}}. \qquad (1)$$

The first expression converts to Eq. (1) by substituting y's for x's in the right-hand side, by computing the numerator term only when a 1 appears in matrix **C**, and by averaging the numerator cross-product terms over the total number of pairs denoted by a 1 in matrix **C**. The denominator of the revised expression (Eq. 1) is the sample variance of Y, s_Y^2. Coupling this with part of the accompanying numerator term renders

$$\frac{(y_i - \bar{y})}{s_Y}\sum_{j=1}^{n} c_{ij}\frac{(y_j - \bar{y})}{s_Y},$$

where this summation term is the quantity measured along the vertical axis of the modified Moran scatterplot; Eq. (1) is known as the Moran coefficient (MC). Accordingly, positive spatial autocorrelation occurs when the scatter of points on the associated Moran scatterplot reflects a straight line sloping from the lower left-hand to the upper right-hand corner: high values on the vertical axis tend to correspond with high values on the horizontal axis, medium values with medium values, and low values with low values (Fig. 1). Negligible spatial autocorrelation occurs when the scatter of points suggests no pattern: high values on the vertical axis correspond with high, medium, and low values on the horizontal axis, as would medium and low values on the vertical axis. Negative spatial autocorrelation occurs when the scatter of points reflects a straight line sloping from the upper left-hand to the lower right-hand corner: high values on the vertical axis tend to correspond with low values on the horizontal axis, medium values with medium values, and low values with high values. These patterns are analogous to those for two different quantitative attribute variables—X and Y—rendering, respectively, a positive, zero, and negative Pearson product moment correlation coefficient value.

The semivariogram is based on squared paired comparisons of georeferenced data values. Emphasizing variation with distance rather than only with nearby values, the numerator of Eq. (1) may be replaced by

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(y_i - y_j)^2}{\left[2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right]},$$

converting this measure to the Geary ratio (GR) when the unbiased sample variance is substituted in the denominator.

As map pattern, spatial autocorrelation is viewed in terms of trends, gradients, or mosaics across a map. This more general meaning can be obtained by studying the matrix form of the MC, specifically the term $\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)\mathbf{Y}$ corresponding to the first summation in Eq. (1), where **I** is an n-by-n identity matrix, **1** is an n-by-1 vector of ones, T is the matrix transpose operation, and $(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)$ is the projection matrix commonly found in conventional multivariate and regression analysis that centers the vector **Y**. The extreme eigenvalues of matrix expression $(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n)$ determine the range of the modified correlation coefficient, MC; therefore, MC is not restricted to the range $[-1, 1]$. Furthermore, the full set of n eigenvalues of this

expression establishes the set of distinct MC values associated with a map, regardless of attribute values. The accompanying n eigenvectors represent a kaleidoscope of orthogonal and uncorrelated map patterns of possible spatial autocorrelation.

The first eigenvector, $\mathbf{E}_1^*$, is the set of numerical values that has the largest MC achievable by any set, for the spatial arrangement defined by the geographic connectivity matrix $\mathbf{C}$. The second eigenvector is the set of values that has the largest achievable MC by any set that is uncorrelated with $\mathbf{E}_1^*$. This sequential construction of eigenvectors continues through $\mathbf{E}_n^*$, which is the set of values that has the largest negative MC achievable by any set that is uncorrelated with the preceding $(n-1)$ eigenvectors. As such, these eigenvectors furnish distinct map pattern descriptions of latent spatial autocorrelation in georeferenced variables.

As a diagnostic tool, spatial autocorrelation plays a crucial role in model-based inference, which is built on valid assumptions rather than on the outcome of a proper sampling design. Sometimes, spatial autocorrelation is used as a diagnostic tool for model misspecification, being viewed as an artifact of overlooked nonlinear relationships, nonconstant variance, or outliers. Cliff and Ord provide an excellent example with their empirical analysis of the relationship between percentage change in population and arterial road accessibility across the counties of Eire. MC = 0.1908 for residuals obtained from a bivariate regression using these two variables; MC = 0.1301 for residuals obtained from a bivariate regression applying a logarithmic transformation to each of these two variables. However, Griffith and Layne determine and then employ an optimal Box–Tidwell linearization transformation for these data, which renders MC = −0.0554 for the residuals, a value that suggests the absence of spatial autocorrelation. In other words, the weak positive spatial autocorrelation detected by Cliff and Ord is due solely to a nonlinear model misspecification.

As redundant information, spatial autocorrelation represents duplicate information contained in georeferenced data, linking it to missing values estimation as well as to notions of effective sample size and degrees of freedom. For normally distributed variables, these latter two quantities establish a correspondence between n spatially autocorrelated and n^* zero spatial autocorrelation (i.e., independent) observations. Richardson and Hémon promote this view for correlation coefficients computed for pairs of geographically distributed variables. Haining demonstrates an equivalency between their findings and the results obtained by removing spatial dependency effects with filters analogous to those used in constructing time series impulse-response functions.

Inference about a geographic variable mean when nonzero spatial autocorrelation is present is impacted by a variance inflation factor (VIF), and has $n^* \leq n$. The respective matrix formulae, where TR denotes the matrix trace operation and $\mathbf{V}\sigma^{-2}$ denotes the n-by-n inverse variance–covariance matrix capturing latent spatial autocorrelation effects, are $\text{VIF} = \text{TR}(\mathbf{V}^{-1})/n$ and $n^* = n \times \text{TR}(\mathbf{V}^{-1})/\mathbf{1}\mathbf{V}^{-1}\mathbf{1}$. Selected results for these two formulae appear in Table II, where $\hat{\rho}_{\text{SAR}}$ denotes estimated spatial autocorrelation using a simultaneous autoregressive (SAR) model specification, and suggest that on average approximately two-thirds of the information content is redundant; spatial autocorrelation is at least doubling the variance; and, for example, a cluster of approximately 10 additional census tracts needs to be acquired for Houston before as much new information is obtained as is contained in a single, completely isolated census tract in this metropolitan region.

As a missing variables indicator/surrogate, spatial autocorrelation accounts for variation otherwise unaccounted for because of variables missing from a regression equation. This perspective is particularly popular among spatial econometricians. In essence, autocorrelation effects latent in predictor variables match autocorrelation effects in Y. For instance, one well-known covariate of population density is distance from the central business district (CBD). For Adair (Table II), a bivariate regression analysis reveals that this variable accounts for approximately 91% of the variation in population density across the county, whereas $\hat{\rho}_{\text{SAR}}$ decreases to 0.33686 and the total percentage of variance accounted for increases slightly to approximately 92%. (The trend contributes considerably to the nature of the semivariogram plot curve in Fig. 1.) For Chicago, a bivariate

Table II Redundant Information Measures for Selected Georeferenced Population Density Datasets

Dataset	*MC*	*GR*	$\hat{\rho}_{\text{SAR}}$	n	*VIF*	$\hat{n}^*$	*% of variance accounted for*
Adair County, Missouri, block groups	0.62035	0.30765	0.95298	26	20.77	1.1	88.8
Syracuse census tracts	0.68869	0.28128	0.82722	208	3.08	17.9	71.9
Houston census tracts	0.55780	0.40129	0.77804	690	2.16	70.5	59.6
Chicago census tracts	0.68267	0.30973	0.87440	1754	3.24	85.5	68.6
Coterminous U.S. counties	0.62887	0.28247	0.84764	3111	2.68	186.6	68.7

regression analysis reveals that this variable accounts for approximately 42% of the variation in population density across the city, whereas $\hat{\rho}_{SAR}$ decreases to 0.76994 and the total percentage of variance accounted for remains at approximately 68%.

As a spatial spillover effect, spatial autocorrelation results from effects of some phenomenon at one location "spilling over" to nearby locations, much like a flooding river overflowing its banks. Pace and Barry provide an empirical example of house price spillover: the value of a house is a function of both its dwelling attributes and the value of surrounding houses. They studied 20,640 California block groups having houses and reported $\hat{\rho}_{SAR} = 0.8536$, indicating the presence of strong positive spatial autocorrelation, with inclusion of the autoregressive term increasing the percentage of variance explained by 25%.

As a spatial process mechanism, spatial autocorrelation is viewed as the outcome of some course of action operating over a geographic landscape. The contagious spread of disease, the dissemination of information or ideas, and spatial competition illustrate this viewpoint, and an autologistic model describes it: 1 denotes the presence and 0 denotes the absence of some phenomenon at different locations in a geographic landscape.

As an outcome of areal unit demarcation, spatial autocorrelation relates to the modifiable areal unit problem, whereby results from statistical analyses of georeferenced data can be varied at will simply by changing the surface partitioning to demarcate areal units. In an analysis of variance framework, devising areal units in a way that manipulates attribute differences within and between them impacts on the nature and degree of measured spatial autocorrelation. If this practice is executed in a gerrymandering fashion, a range of possible spatial autocorrelation, from positive to negative, materializes. In part, accounting for detected spatial autocorrelation in statistical analyses attempts to neutralize such outcomes.

Estimators of Spatial Autocorrelation

Spatial autocorrelation may be indexed, quantified by including an autoregressive parameter in a regression model, or filtered from variables. Spatial autocorrelation can be quantified with indices. Equation (1) provides the MC index, which can also be rewritten in terms of the regression coefficient affiliated with a Moran scatterplot. Its range is approximately ± 1; more precisely, it is

$$\left[\frac{n\lambda_{\min}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}, \quad \frac{n\lambda_{\max}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}\right],$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the extreme eigenvalues of matrix $(\mathbf{I} - \mathbf{1}\mathbf{1}^{T}/n)\mathbf{C}(\mathbf{I} - \mathbf{1}\mathbf{1}^{T}/n)$. As MC approaches the upper limit, the paired values $(\mathbf{Z}_Y, \mathbf{C}\mathbf{Z}_Y)$ in a Moran scatterplot increasingly align with a straight line having a positive slope. As MC approaches the lower limit, the alignment is with a straight line having a negative slope. As MC approaches $-1/(n-1)$ (its expected value indicating zero spatial autocorrelation), the paired values should resemble a random scatter of points. The standard error of this statistic is approximately $\sqrt{(2/\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij})}$; in terms of practical importance, modest levels of positive spatial autocorrelation begin at $0.5\text{MC}/\text{MC}_{\max}$, whereas moderate levels begin at $0.7\text{MC}/\text{MC}_{\max}$ and substantial levels begin at $0.9\text{MC}/\text{MC}_{\max}$.

One variation of this index is the GR, which replaces the numerator of Eq. (1) with a squared paired comparison and the denominator with the unbiased sample variance estimate. This index roughly ranges from 0 (i.e., $y_i = y_j$), indicating perfect positive spatial autocorrelation, to 2, indicating strong negative spatial autocorrelation; 1 indicates zero spatial autocorrelation. GR is inversely related to MC. The extreme values are more precisely given by

$$\left[\frac{(n-1)\lambda_{n-1}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}, \quad \frac{(n-1)\lambda_{\max}}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}}\right],$$

where λ_{n-1} and $\lambda_{\max}$ are the second smallest and the largest eigenvalues of matrix $(\langle \mathbf{1}^{T}\mathbf{C}\rangle_{\text{diagonal}} - \mathbf{C})$, where $\langle \mathbf{1}^{T}\mathbf{C}\rangle_{\text{diagonal}}$ is a diagonal matrix whose diagonal entries are the row sums of $\mathbf{C}$. This feature indicates that GR emphasizes edges and numbers of neighbors far more than does MC.

Another variation is the triplet of join count statistics used to analyze 0/1 binary georeferenced data, conveniently coded as 1 denoting black (B) and 0 denoting white (W). The number of neighboring pairs of 1's on a map equals 2BB, the number of neighboring pairs of 0's equals 2WW, and the number of 1's with neighboring 0's equals 2BW:

$$\text{BB} + \text{BW} + \text{WW} = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}/2,$$

where n_1 is the number of 1's. These quantities can be interpreted in terms of binomial random variables. The numerator of the MC reduces to

$$n_1^2 \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}/n^2 + 2(1 - n_1/n)\text{BB} - 2(n_1/n)\text{BW}.$$

Spatial autocorrelation can be quantified by including an autoregressive parameter in a model specification. Regression becomes autoregression, with the most

common specifications being the SAR ($\mathbf{Y} = \rho_{SAR}\mathbf{WY} + (\mathbf{I} - \rho_{SAR}\mathbf{WY})\mathbf{X}\beta + \boldsymbol{\varepsilon}$), the autoregressive response (AR) ($\mathbf{Y} = \rho_{AR}\mathbf{WY} + \mathbf{X}\beta + \boldsymbol{\varepsilon}$), and the conditional autoregressive (CAR) ($\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I} - \hat{\rho}_{CAR}\mathbf{C})^{-1/2}\boldsymbol{\varepsilon}$) models, where β is a vector of regression coefficients and $\boldsymbol{\varepsilon}$ is a vector of independent and identically distributed random error terms. These are popular among spatial statisticians, spatial econometricians, and image analysts, respectively. The SAR model specifies spatial autocorrelation as being in the error term, with the attribute error at location i being a function of the average of nearby error values: $E(\mathbf{Y}) = \mathbf{X}\beta$, $\mathrm{VAR}(\mathbf{Y}) = \sigma^2[(\mathbf{I} - \rho_{SAR}\mathbf{W}^T)\times(\mathbf{I} - \rho_{SAR}\mathbf{W})]^{-1}$, where E denotes the calculus of expectation operator. The AR model specifies spatial autocorrelation as being a direct dependency among the Y values: $E(\mathbf{Y}) = (\mathbf{I} - \rho_{AR}\mathbf{W})^{-1}\mathbf{X}\beta$,$\mathrm{VAR}(\mathbf{Y}) = \sigma^2[(\mathbf{I} - \rho_{AR}\mathbf{W}^T)(\mathbf{I} - \rho_{AR}\mathbf{W})]^{-1}$. The CAR model specifies spatial autocorrelation as being in the error term, with a weaker degree and smaller spatial field than the SAR model, and with the attribute error at location i being a function of the sum of nearby error values: $E(\mathbf{Y}) = \mathbf{X}\beta$, $\mathrm{VAR}(\mathbf{Y}) = \sigma^2(\mathbf{I} - \rho_{CAR}\mathbf{C})^{-1}$. Parameters of these models must be estimated using maximum likelihood techniques.

Parameterizing spatial autocorrelation through the semivariogram plot involves modeling the relationship between semivariance, γ, and distance, d. Dozens of specifications may be employed, all describing spatial autocorrelation as a nonlinear decreasing function of distance. The most popular ones are the spherical, the exponential, and the Gaussian; one that should increase in popularity is the Bessel function. The empirical semivariogram in Fig. 1 is best described by a Bessel function, K_1, both before and after adjusting for the underlying distance decay trend. Each semivariogram model describes the decline in spatial autocorrelation with increasing distance in terms of an intercept (nugget), a slope, and an implicit/explicit range of spatial dependency. For Adair County population density (Fig. 1), $\hat{\gamma} = 0.13 + 14.07[1 - (d/0.25)K_1(d/0.25)]$. Summarizing the variance–covariance matrix with this type of equation allows new variance–covariance matrices to be constructed for locations whose attribute values are unknown, permitting sensible predictions of their values.

Spatial filters remove spatial autocorrelation from variables by casting it as redundant information or as an outcome of map pattern. The former interpretation is employed by Haining and renders a spatial linear operator filter ($\mathbf{I} - \hat{\rho}_{SAR}\mathbf{W}$). The latter interpretation renders predictor variables, such as the battery of eigenvectors $\mathbf{E}_j^*$, that capture locational information summarized by a spatial autocorrelation parameter such as $\hat{\rho}_{SAR}$. For Adair County, the correlation between log-distance and log-density is −0.936; the spatial linear operator filter result based on an SAR model ($\hat{\rho}_{SAR} = 0.95298$; Table II) is −0.985. Eigenfunction filtering identifies two eigenvectors (MC = 0.76 and 0.47) that account for the residual spatial autocorrelation and increase the variance accounted for in log-population density by approximately 4%.

Theoretical Statistical Properties of Spatial Autocorrelation

Classical statistics establishes the quality of parameter estimators with specific properties that discriminate between useful and useless ones. Four of these properties are described here.

An unbiased estimator's sampling distribution arithmetic mean equals its corresponding population parameter value. This property is evaluated with the calculus of expectations. In general, in the presence of nonzero spatial autocorrelation, conventional estimators for first-order moments are unbiased, whereas those for second-order moments are biased. For example, for linear regression, $E(\mathbf{Y}) = E(\mathbf{X}\beta + \mathbf{V}^{-1/2}\boldsymbol{\varepsilon}\sigma) = \mathbf{X}\beta$. Similarly, $E(\mathbf{b}) = E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{XY}] = \beta$, where $\mathbf{b}$ is the ordinary least squares (OLS) regression coefficients estimator. However, $E[(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)/n] = \sigma^2\mathrm{TR}(\mathbf{V})/n$, which reduces to σ^2 only when $\mathbf{V} = \mathbf{I}$. Griffith and Lagona report that misspecifying matrix $\mathbf{C}$, of which matrix $\mathbf{V}$ is a function, results in $\mathbf{b}$ remaining unbiased and s^2 remaining biased when autocorrelation is accounted for.

An efficient estimator is an unbiased estimator whose sampling distribution has the smallest possible variance, maximizing its reliability. Cordy and Griffith find that in the presence of nonzero spatial autocorrelation, the biased OLS variance estimator negates much of its computational simplicity advantage. Consequently, the OLS standard error estimator tends to underestimate the true standard error when positive spatial autocorrelation prevails. By accounting for latent spatial autocorrelation, gains in efficiency increase as both its magnitude and n increase. With regard to the aforementioned VIF and n^*, $E[(\bar{y} - \mu_Y)^2] = (\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}/n)(\sigma^2/n)$, and $E[(\mathbf{Y} - \mu_Y\mathbf{1})^T\times(\mathbf{Y} - \mu_Y\mathbf{1})/n] = [\mathrm{TR}(\mathbf{V}^{-1})](\sigma^2/n)$. These results reduce to their respective standard results of σ^2/n and σ^2 only when $\mathbf{V} = \mathbf{I}$. A measure of OLS efficiency when spatial autocorrelation is nonzero is given by $n^2/[\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}\mathbf{1}^T\mathbf{V}\mathbf{1}]$; when spatial autocorrelation is zero, this quantity equals 1, whereas when perfect positive spatial autocorrelation exists, it is 0. Griffith and Lagona report that misspecifying matrix $\mathbf{C}$ results in $\mathbf{b}$ being asymptotically efficient and s^2 remaining inefficient when autocorrelation is accounted for.

A consistent estimator's sampling distribution concentrates at the corresponding parameter value as n increases. Considering situations in which the maximum number of neighbors for any given location is finite, when sample size increases by increasing the size of a region to

an unbounded surface (i.e., increasing domain asymptotics), consistency of the mean and variance estimators is attained. When sample size increases by partitioning a region into an increasingly finer tessellation (i.e., infill asymptotics), consistency is lost. Griffith and Lagona report that misspecifying matrix **C** in a typical situation results in **b** and s^2 being consistent, with the autocorrelation parameter failing to converge to its true value.

A sufficient estimator utilizes all of the pertinent information content of a sample needed to estimate a particular parameter. This property is established using the factorization criterion for a likelihood function. A likelihood can be rewritten as the product of a term that depends on the sample only through the value of the parameter estimator and of a term independent of the corresponding parameter. The importance of this property is twofold: (i) Estimating missing georeferenced data requires imputation of the complete-data sufficient statistics, which in the case of an auto-normal probability model involves a spatial autocorrelation term, and (ii) the Markov chain Monte Carlo (MCMC) procedure used to estimate the auto-logistic and auto-binomial models requires the sufficient statistics, again one being a spatial autocorrelation term.

Common Auto-Probability Model Specifications

The normal distribution is the base of much statistical analysis of continuous data. The binomial distribution plays the same role for binary variables and percentages, whereas the Poisson distribution plays the same role for counts. Auto-specifications of these models almost always have pairwise-only spatial dependence.

The auto-Gaussian model specification yields the likelihood function

$$L = \text{constant} - (n/2)\text{LN}(\sigma^2) + (1/2)\text{LN}[\det(\mathbf{V}^{-1})] - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\text{T}}\mathbf{V}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})/(2\sigma^2),$$

where LN denotes the natural logarithm, and det denotes the matrix determinant operation. The normalizing constant, $(1/2)\text{LN}[\det(\mathbf{V}^{-1})]$, complicates calculation of the maximum likelihood estimates of parameters because it involves an n-by-n matrix. For the CAR model, $\mathbf{V}^{-1} = (\mathbf{I} - \rho_{\text{CAR}}\mathbf{C})$, whereas for the SAR model $\mathbf{V}^{-1} = (\mathbf{I} - \rho_{\text{CAR}}\mathbf{W})^{\text{T}}(\mathbf{I} - \rho_{\text{SAR}}\mathbf{W})$.

The auto-logistic model for a pure spatial autoregressive situation involves estimating the parameters of the following probability function:

$$E[Y_i = 1 \,|\, \mathbf{C}_i\mathbf{Y}] = \frac{\exp(\alpha_i + \rho \sum_{j=1}^{n} c_{ij}y_j)}{1 + \exp(\alpha_i + \rho \sum_{j=1}^{n} c_{ij}y_j)},$$

where α_i is the parameter capturing large-scale variation (and hence could be specified in terms of vector $\mathbf{X}_i$), ρ is the spatial autocorrelation parameter, and $\mathbf{C}_i$ is the ith row vector of matrix **C**. Spatial autocorrelation may be measured with the join count statistics. Parameters can be estimated with MCMC techniques: $\sum_{i=1}^{n} y_i \sum_{j=1}^{n} c_{ij}y_j/2 = \text{BB}$ (the join count statistic) is the sufficient statistic for spatial autocorrelation. The model for percentages is very similar.

The auto-Poisson model for a pure spatial autoregressive situation involves evaluating the following log-probability mass function term:

$$\sum_{i=1}^{n} \alpha_i n_i - \sum_{i=1}^{n} \text{LN}(n_i!) + \rho \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}n_i n_j,$$

where n_i are the counts for areal unit i. Parameters, which in this specification restrict ρ to being negative, can be estimated with MCMC techniques here, too. Also, the estimate of ρ could be driven to zero by introducing judiciously selected spatial autocorrelation filtering variables, such as the aforementioned eigenvectors.

What Should an Applied Spatial Scientist Do?

To ignore spatial autocorrelation effects when analyzing georeferenced data may be tempting, but it is ill-advised. Assuming independent observations—the atypical in a geographic world—is merely for the convenience of mathematical statistical theory. Introducing a spatial autocorrelation parameter into a model specification, which indeed results in more complicated statistical analyses, produces better statistical practice and results: Auto-models provide clearer data descriptions, parameter estimators exhibit better statistical properties and behavior, and data analysis results can be more easily interpreted. Modern statistics supplies the necessary estimation tools, and standard commercial statistical software packages supply the capability for executing this estimation.

When studying a georeferenced dataset, the applied spatial scientist should (i) compute a spatial autocorrelation index, (ii) estimate an auto-model keeping conventional regression analysis in mind, and (iii) inspect local spatial autocorrelation statistics as part of the battery of model diagnostics. Conventional regression model analysis protocol offers a guide because both the Moran scatterplot and the MC link directly to regression analysis. During this pursuit, computations such as effective sample size may help determine whether collecting supplemental data is worthwhile, more precise standard errors may help determine whether two variables are significantly correlated, and a sizeable quantity of variance explained by spatial autocorrelation may help determine

whether variables are missing from a model specification. Also, marked levels of spatial autocorrelation may be exploited for spatial interpolations and small geographic area estimations, providing the scientist with a window into the unknown. Moreover, accounting for spatial autocorrelation latent in georeferenced data has the capacity to increase our understanding of the social world.

See Also the Following Articles

Correlations • Ordinary Least Squares (OLS) • Spatial Databases • Spatial Pattern Analysis • Spatial Sampling

Further Reading

Anselin, L. (1988). *Spatial Econometrics.* Kluwer, Boston.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geogr. Anal.* **27,** 93–115.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Statistical Soc. Ser. B* **36,** 192–225.

Cliff, A., and Ord, J. (1981). *Spatial Processes.* Pion, London.

Cordy, C., and Griffith, D. (1993). Efficiency of least squares estimators in the presence of spatial autocorrelation. *Commun. Stat. Ser. B* **22,** 1161–1179.

Cressie, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

Getis, A. (1995). Spatial filtering in a regression framework, experiments on regional inequality, government expenditures, and urban crime. In *New Directions in Spatial Econometrics* (L. Anselin and R. Florax, eds.), pp. 172–188. Springer-Verlag, Berlin.

Getis, A., and Ord, J. (1992). The analysis of spatial association by use of distance statistics. *Geogr. Anal.* **24,** 189–206.

Griffith, D. (2000). A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Systems* **2,** 141–156.

Griffith, D., and Lagona, F. (1998). On the quality of likelihood-based estimators in spatial autoregressive models when the data dependence structure is misspecified. *J. Statistical Planning Inference* **69,** 153–174.

Griffith, D., and Layne, L. (1999). *A Casebook for Spatial Statistical Data Analysis.* Oxford University Press, New York.

Haining, R. (1991). Bivariate correlation and spatial data. *Geogr. Anal.* **23,** 210–227.

Pace, R., and Barry, R. (1997). Sparse spatial autoregressions. *Stat. Probability Lett.* **33,** 291–297.

Richardson, S., and Hémon, D. (1981). On the variance of the sample correlation between two independent lattice processes. *J. Appl. Probability* **18,** 943–948.

Tiefelsdorf, M., and Boots, B. (1995). The exact distribution of Moran's I. *Environ. Planning A* **27,** 985–999.

Spatial Cognition and Perception

Marina Vasilyeva
University of Chicago, Chicago, Illinois, USA

Glossary

allocentric coding Coding location in relation to external features of the environment.
egocentric coding Coding location in relation to self.
landmarks Elements of the environment that stand out from their surroundings.
large-scale space Space that cannot be seen from a single viewpoint and thus requires movement in order to be perceived in its entirety.
small-scale space Space that can be seen in its entirety from a single viewpoint.

Spatial cognition refers to a diverse set of abilities, all of which require perceiving and reasoning about spatial properties, such as size, shape, and location. These abilities are critically important for the functioning of all mobile organisms. Animals rely on spatial skills to find food, return to shelter and avoid danger. Likewise, humans are constantly confronted with a variety of problems that require reasoning about space. Some of these problems involve judging and transforming spatial relations among objects, as in assembling an item of furniture from its component parts. Other problems concern navigating through and locating objects in familiar or unfamiliar environments, such as finding the best route through a neighborhood or locating a car in a large parking lot. Spatial thinking is thus crucial to everyday life. Furthermore, it provides a foundation for successful performance in the sciences, in technological fields, and in engineering.

Regardless of the complexity of the task, spatial understanding depends on the basic ability to code metric information, namely distance and directional cues. This information can be used to determine the size and shape of objects and to specify object locations. Researchers distinguish two fundamental ways in which spatial cues can be used to determine location. The first is known as *egocentric coding*, that is, coding location with respect to oneself. The second is known as *allocentric (or geocentric) coding*, that is, coding location with respect to external features of the environment. Since the egocentric strategy specifies a target location relative to the observer, it can be used either when the observer remains stationary or when the observer moves but is able to keep track of the movement and update the changing relation between him or herself and the target location. In some situations, however, the observer may become sufficiently disoriented so that it is impossible to reconstruct the changes that led to the new position. In this case, one must rely on the allocentric strategy that provides information about the target location relative to the environment itself, independent of the observer. Research in the spatial domain indicates that both egocentric and allocentric types of coding are available to humans as well as nonhuman animals.

Spatial Cognition in Animals

Most animals must travel away from their shelters to get food and then find their way back. The path traversed in search of food may be quite long and unpredictable, and finding one's way home often involves navigating through unfamiliar environments. Yet animals ranging from mollusks to primates are impressively accurate in solving such spatial tasks. In doing so, they rely on a number of spatial strategies, the most common of which is dead reckoning that involves keeping track of changes in one's position. This tracking mechanism allows for constant updating of the egocentrically encoded location. To measure the animal's reliance on dead reckoning, researchers use a displacement technique in which they remove the animal from its natural path connecting the nest and the food source and examine how the animal attempts to return to the nest. A set of studies with desert ants conducted

by Wehner illustrates this technique. A desert ant travels along a circuitous path searching for food; once the food is found, the ant returns home following the most efficient straight path. If the ant is displaced by the experimenter right before it is about to start its journey back to the nest, it moves the same distance and in the same direction as if it were not displaced (thereby failing to arrive back at the nest). Thus, the desert ant seems to rely on the representation of the egocentric relation between itself and the location of the nest, rather than on the external features of the environment. Similar behavior has been observed in a variety of other animal species, including bees, spiders, gerbils, and hamsters.

An important question is how animals manage to keep track of continuous changes in their position. It has been proposed that an animal may represent its position as a vector that indicates distance and direction from the nest to the current position. As the animal moves to a new position, a vector specifying the distance and direction traveled during the move is added to the vector specifying the previous position, thereby updating the relation between the animal and the nest. A critical aspect of this mechanism is the ability to code distance and directional cues. Experimental evidence indicates that animals use the patterns of visual flow to obtain information about changes in distance traveled and rely on their vestibular system to represent changes in directions.

In addition to relying on dead reckoning, animals use landmarks to find locations of important objects and places. Although landmarks are features of the external environment, they can be integrated into an egocentric strategy. For example, the work of Collett and Rees shows that bees tend to approach landmarks (e.g., trees) marking the location of a goal (e.g., a hive) from the same direction in which they originally saw those landmarks. It has been argued that they match the image of landmarks as seen from their current viewpoint with the original view. However, many animals can approach landmarks from different directions and still locate the goal, leading researchers to posit that these animals store multiple images representing landmarks from different viewpoints. It should be pointed out, though, that the use of landmarks can also be integrated into an allocentric strategy in which location is coded independently of the viewer's perspective. To measure the ability to code location independent of one's own perspective, investigators use disorientation tasks in which the relation between the viewer's original position and the goal is disrupted completely. Performance on these tasks shows that many animal species use landmarks to locate a goal after being disoriented.

Studies of animal behavior following disorientation reveal another impressive spatial skill, namely the ability to use information about the shape of enclosed spaces in locating a goal. In one of the earliest studies by Cheng, rats were placed in a rectangular box with food hidden in a particular location. The rats searched for the food after they had been disoriented. On most trials, they looked for the food at either the correct location or at a geometrically equivalent location diagonally opposite from it. Since disorientation prevents the animals from tracking the change in their relation to the hiding corner, their search behavior must have been based on the use of features of the space itself. The observed behavior can be explained by positing that the animals coded the relative lengths of the walls and represented the target location as the corner with a particular relation between adjacent walls (e.g., the longer wall to the right and the shorter wall to the left or vice versa). Thus, when animals can use tracking during their movement, they rely on viewer-centered egocentric spatial strategies, but when tracking becomes unavailable they use environment-centered allocentric spatial strategies.

Development of Spatial Cognition in Humans

The systematic study of spatial development in human beings began in the 1940s with the work of Jean Piaget. He believed that children initially have a very impoverished understanding of space and that they progress through several developmental stages gradually acquiring a more advanced understanding of spatial relations. During the first, Topological stage, children are able to encode only a limited set of spatial relations, such as proximity (adjacent vs nonadjacent), order (in front vs behind), and enclosure (inside vs outside). Distances and lengths are not coded as features of stimuli themselves, but rather are specified in terms of action, e.g., as the amount of reach or movement. A major limitation of this stage is the egocentric nature of spatial coding—information about location is only preserved from an initial viewing position.

The topological stage is followed by the Projective and Euclidean stages, which have been described sometimes as sequential and sometimes as concurrent. The main line of development during the Projective stage is from egocentric to allocentric coding. Piaget demonstrated this development with a series of perspective-taking tasks, in which children had to indicate how a spatial array would look to a person who viewed it from a different position. The predominant response in children younger than 9 years of age was to choose the representation of the array that showed it from their own point of view, but older children were able to select the correct representation. Piaget concluded that around the age of 9 children start using actual spatial relations between objects in the layout rather than relying on egocentric information. Another major line of development, associated with

the Euclidean stage, is from nonmetric to metric spatial coding. Piaget argued that around 9–10 years of age children become accurate in coding locations of objects that are not immediately adjacent to landmarks, which requires using information about distance.

More recent findings, however, indicate that children represent metric properties of space at earlier ages than previously thought. For example, toddlers can find an object hidden in an enclosed space by remembering the distance between that object and the edge of the space. Even infants show sensitivity to distance information, which was demonstrated in a series of habituation studies. In the habituation paradigm, infants are repeatedly shown the same display until their looking time decreases substantially, after which they are shown a novel display. An increase in looking time indicates that children have noticed the difference between the two displays. Five-month-olds who have seen a toy being repeatedly hidden and then retrieved at a particular location within a space, look longer if a toy hidden in that location emerges from a new location.

An accumulating body of evidence also indicates that spatial coding in young children is not necessarily tied to their initial viewing position. As discussed above, one way to identify a location from a new position is by keeping track of changes in the relation between oneself and the location. Toddlers demonstrate this ability in tasks where they observe an object being hidden in a particular location and are then required to locate the object after walking to a new position. When the task is administered in a homogeneous environment where no landmarks are visible, children must rely on dead reckoning to perform successfully. Another way to locate an object after movement is by using the relation between the object's location and landmark(s). In the first two years of life, children mostly use landmarks to locate a target object when the landmarks are immediately adjacent to the object. Starting around the age of two, however, children also begin to benefit from the presence of distal landmarks. While these findings indicate that the ability to code object location relative to distal landmarks emerges earlier than proposed by Piaget, other evidence suggests that landmark use nonetheless undergoes substantial development through the elementary school age. During that period children become progressively more accurate and flexible in the use of landmark information; they begin integrating their knowledge of individual landmarks to represent relations among multiple locations and to form routes that connect ordered sequences of landmarks.

In addition to landmark use, children demonstrate the ability to rely on geometric properties of space (shape) to code object location. Adopting Cheng's paradigm, Hermer and Spelke showed toddlers an object being hidden in a corner of a rectangular room and then asked them to find the object following disorientation. Similar to the results of animal studies, children searched for the object either in the correct corner or in the geometrically equivalent corner. It has been shown that under certain conditions, namely in small enclosed spaces, children use geometric cues to the exclusion of landmark cues. For example, when a rectangular room contains a landmark that distinguishes the two geometrically equivalent corners (e.g., a wall of a particular color), children seem to ignore that landmark and search solely on the basis of geometry. These findings have led Hermer and Spelke to propose that toddlers' ability to locate an object following disorientation is modular, restricted to the use of geometric information. However, in larger spaces, children successfully combine landmark and geometric cues, suggesting that space size may be critical in determining the types of cues used to reason about space.

The research on spatial development reveals impressive abilities in young children, indicating that the skills they possess at the starting points of development are stronger than those posited by Piaget. At the same time, this research shows certain limitations of early spatial behaviors, suggesting a long developmental progression between the starting points and a mature spatial competence. Many of the spatial abilities revealed by children can be observed only in limited contexts. While a number of studies have shown that children code metric information to identify object location, other studies reveal that children do not always realize the need to consider distance and direction and may revert to a more primitive strategy of coding location only in terms of adjacent landmarks. Moreover, while children clearly have an ability to solve spatial tasks by using the relations between spatial features of the environment, they still have difficulty with Piagetian perspective-taking tasks (until about the age of 9 years), which require the use of allocentric rather than egocentric strategies.

Spatial Perception and Reasoning in Adults

When researchers discuss a mature form of spatial cognition, they often imply by the term "mature form" the ability to encode metric relations and to represent these relations independently of one's own viewpoint. Adults, indeed, are more precise in metric coding than children and more flexible in the ability to take different perspectives. Nevertheless, a closer examination of strategies underlying performance of adults on some spatial tasks reveals the influence of nonmetric information on their judgments of distance as well as viewpoint specificity in reasoning about space. Thus, the mental representation of space constructed by adults is not the exact replica of the actual physical space.

One factor underlying this phenomenon is the influence of categorical information. For example, when people are asked about the positions of two cities relative to each other, they rely on their knowledge of geographic categories. As a result, Seattle (USA) is often thought to be to the south of Montreal (Canada) because USA lies largely to the south of Canada even though for these particular cities the relation is reversed. In addition to geographic categories, people use other kinds of categorical information in reasoning about spatial relations. To examine the use of categories in spatial reasoning, researchers present people with sets of stimuli that vary in terms of size or location. For example, people are shown lines on the computer screen that vary in length. The lines are presented one at a time; after the line disappears, the subject is asked to reproduce it by pressing designated keys on the keyboard. In another type of reproduction task, people are shown a rectangular enclosure with a dot inside it. The location of the dot varies from trial to trial and the task is to reproduce the dot's location as accurately as possible.

The examination of performance on such reproduction tasks has led to a proposal that people combine fine-grained information about location or size of a particular stimulus with category information. In remembering the location of an object within an enclosed space, people seem to subdivide that space into categories (e.g., quadrants); their estimates are misplaced toward the prototypes (centers) of these imposed spatial categories. In remembering the size of an object, people combine the information about that particular object with the categorical information about objects of that kind; their estimates are biased toward a prototypical size of objects belonging to this category. The use of categorical information introduces some degree of bias in estimating individual stimuli, however, on average, it yields more accurate judgments.

Another important factor that affects spatial reasoning in adults is viewpoint dependence. While adults are capable of coding spatial relations in terms of external features of the environment (i.e., independent of their own perspective), they do not necessarily rely on this type of coding in solving spatial tasks. This phenomenon has been documented, for example, in object recognition studies. In identifying individual objects, people often rely on their spatial characteristics. To investigate factors affecting object recognition, researchers have used matching tasks, in which participants are shown a pair of objects, one after another, and are asked to determine whether the objects are the same. On some trials the objects are identical whereas on other trials they differ. The critical manipulation is whether the pair of objects is presented from the same or different perspectives. Using this design, Hayward, Tarr, and colleagues demonstrated an *alignment effect*—people find it easier to match the objects if they are aligned (i.e., presented from the same viewpoint) than if they are rotated relative to each other. At the same time, Biederman and colleagues argued that the perception of certain features of objects is virtually unaffected by rotation and that matching objects that contain these features is equally easy whether the objects are aligned or not. Some examples of spatial features that can uniquely specify the structural description of an object include whether a particular contour is straight or curved or whether pairs of contours are parallel or not. Biederman proposed that based on this kind of information, people represent objects in terms of relations between simple parts (called *geons*) that have a uniquely identifiable shape.

A parallel set of findings concerning viewpoint dependence versus viewpoint independence of spatial coding is reported in studies of how people represent object locations. These studies commonly use a perspective-change paradigm, in which people learn the positions of objects in a spatial layout from a particular perspective and then are asked to imagine themselves facing the layout from the same or different perspective; in either case the task is to point to various objects in the layout. Most studies report alignment effects—performance suffers, both in terms of pointing error and reaction time, when the viewing perspective at the time of testing is not the same as at the time of learning. Furthermore, the results show a roughly linear increase in reaction time as a function of angular difference between the two perspectives. These findings suggest that people may solve the task by transforming the view of the layout at testing into the view of the layout formed during learning, possibly by mental rotation. Larger transformations take longer and result in greater errors.

It should be noted that under certain conditions, particularly in tasks involving larger spatial layouts, people show a lack of alignment effects in reasoning about spatial relations. The finding that spatial behavior varies depending on the size of the space thus appears to be common to children and adults. Researchers distinguish between small-scale space that can be seen from a single viewpoint and large-scale space that requires movement to be perceived in its entirety. It has been proposed that by navigating in a large-scale space people construct a survey-like mental representation that includes landmarks as well as routes through the space. This type of representation is often referred to as a *cognitive map*, a term introduced by Tolman in his 1948 paper, "Cognitive maps in rats and men." Cognitive maps can be used to reason about relations among different locations, to create a shortcut or a detour through a familiar space. The degree to which such mental representations are complete or accurate depends on the amount of experience with the represented space. It is often argued that cognitive maps have no particular perspective, allowing people to analyze relations between

objects from different viewpoints. Thus, the research on cognitive maps combined with perspective-taking studies that demonstrate alignment effects suggests that human adults can form both viewpoint-dependent and viewpoint-independent representations of layouts, and that the nature of spatial representations seems to depend at least to some extent on the size of the layout.

The Use of Symbolic Spatial Tools

While the concept of cognitive map refers to the mental representation of space constructed by an individual, spatial behavior may also involve the use of maps that are external symbolic representations of space. In fact, the uniquely human ability to use symbolic representations significantly augments spatial capabilities by allowing humans to acquire and communicate information about space beyond that available from direct experience. Maps commonly provide a bird's-eye view of the space covering a large area and making explicit the relations among multiple locations that cannot be seen all at once (or may never have been seen at all). While maps and models provide important tools for spatial reasoning, using these tools requires additional skills that can be acquired though implicit or explicit learning.

In order to be able to use a map, one must establish the correspondence between symbols on the map and objects in the real world (representational correspondence) and also between the spatial relations on the map and those in the real world (geometric correspondence). The work of DeLoache has shown that the ability to establish correspondence between individual symbols and their real-world spatial referents is acquired early in life, around 3 years of age. Dealing with the correspondence of spatial relations that involve distances and angles is a more complicated task. Interpreting metric information presented on the map requires translating distances between spaces of different sizes (i.e., scaling). Recent studies indicate that the ability to carry out a scaling transformation of distance develops during the preschool years. However, this ability is initially quite limited and in many map tasks children ignore the difference in scale between the two spaces. Using a map to find locations in the actual space is especially challenging when the map is not aligned with that space. Children often fail to correct for the lack of the alignment, which leads them to particular patterns of errors in map-reading tasks. When adults are presented with a small-size map that is not aligned with the environment, they often rotate the map until it is oriented in the same way. If they are not allowed to reorient the map, they attempt to correct mentally for differences in orientation, which takes more time and often results in greater errors compared to dealing with aligned maps.

Experience with maps may have important implications for spatial reasoning, particularly in thinking about large-scale spaces that are typically represented on geographic maps. As argued by Uttal, the influence of maps on spatial cognition is both specific and general. Using a map provides people with an opportunity to think about a particular space in terms of multiple relations between different locations, as they are represented on the map. However, maps may also transform our understanding of the environment in a more general sense, by providing a structure that can be mentally imposed on a space in thinking about its overall shape, its constituent elements and the relations among them.

Individual Differences in Spatial Cognition

It is commonly known that individuals differ substantially in the level of their spatial skills. This is true both for tasks involving navigating a large-scale space and for tasks involving manipulating objects and small-scale spatial arrays. Some people get lost easily, even after some degree of familiarization with the space, whereas others learn about spatial environments rapidly. People also vary in the ability to assemble objects, such as items of furniture or equipment, out of the component parts. These individual differences in everyday functioning parallel scientific findings that show a wide variation in performance on tasks that require spatial reasoning. The experimental paradigms used to compare levels of performance across individuals range from psychometric paper-and-pencil tasks (e.g., solving mazes, completing patterns, and matching shapes) to tasks that require problem solving in real-world large-scale spaces (e.g., map reading and wayfinding). It has turned out to be rather difficult to identify a small number of basic underlying spatial factors that could account for performance on such diverse tasks. Whereas some individuals are successful in a wide range of spatial tests, others show more within-subject variability. For example, a person who is good at solving mazes may not be as good at object assembly or map reading.

Individual differences in spatial processing are often discussed under the rubric of sex differences. Indeed, research indicates that males and females differ on both measures of accuracy and reaction time on many spatial tasks. Before addressing these differences, however, it is important to emphasize that there is also large variation within each gender group and that, in fact, there is a substantial overlap between the two groups. That is, particular women may outperform particular men on tasks where males, on average, have an advantage.

One type of sex difference reported in the literature concerns the kinds of cues (e.g., landmarks versus metric information) that are used in representing a spatial environment. This difference is revealed, for example, in direction-giving tasks in which subjects are asked to provide directions specifying how to get from one place to another. Males are more likely to use cardinal directions (e.g., North) and to indicate the distance between points (e.g., number of miles), whereas females are more likely to use landmark information (e.g., buildings). Note though that females increase the use of metric and cardinal information when explicitly asked to do so. Other tasks show that females and males may actually differ in how accurate they are in using different kinds of spatial cues obtained in a large-scale environment. For example, when people are walked through a particular area and then are asked to draw a map of their route, males tend to make smaller errors for distance and direction, whereas females tend to make fewer errors in reporting landmarks that lie along the route. If people are familiarized with the area not by walking through it but rather by studying a picture of that area, gender differences in sensitivity to particular cues disappear. Namely, when the picture is taken away and people are asked to draw a map of a route that they have studied, females and males show comparable levels of accuracy in reporting distances, directions, and landmarks. Thus, the differential sensitivity to landmark versus metric cues may only apply to acquiring information in the large-scale environments.

Another ability that has been associated with sex-related differences is mental rotation. Some mental rotation tasks use a matching paradigm similar to the object recognition studies—participants judge which one of the choice figures represents the target form in a different orientation and which one represents a totally different form. In other tasks, participants are required to rotate mentally and/or fit together component pieces to form a target shape. The results across studies show that males outperform females, as measured by accuracy and reaction time. It has been suggested that the observed differences in performance can be partly due to the difference in strategies used to solve mental rotation tasks. In particular, on matching task, males may favor the strategy of mentally rotating the whole object and comparing the result of this rotation to the target while females may approach the task by comparing the individual parts comprising the target and choice figures. Both strategies can lead one to the right solution, but the latter appears to be less efficient.

It has long been believed that sex-related differences in spatial cognition emerge around the age of puberty. However, more recent evidence indicates the existence of such differences even in preschool children. A number of factors have been proposed as a possible explanation for the observed variation in performance. Biological theories emphasize the relation between hormone levels and spatial ability. It has been shown that males who have androgen deficiency early in life have low spatial ability compared to males with normal hormonal levels. At the same time, females with high androgen levels early in life have higher spatial ability compared to normal controls. Furthermore, hormonal changes within subjects also appear to influence spatial abilities. Thus, females perform best on spatially demanding tasks when their estrogen levels are lowest (in contrast to verbal skills that correlate positively with estrogen levels). The evidence of biological effects does not exclude, however, a possibility that spatial skills are affected by environmental factors. It has been proposed that early play experience, for example playing with construction sets and other spatially relevant toys, may be an important factor in the growth of spatial abilities. Convincing evidence concerning the malleability of spatial skills comes from studies showing that training can significantly enhance performance on tasks involving mental manipulation of spatial stimuli. Thus, both biological and environmental factors appear to be integral to the development of differentiation in spatial skills. It remains to be determined how these factors interact to produce particular skill levels and particular patterns of performance.

Cultural Differences in Spatial Cognition

Culture may influence spatial thinking and behavior in a variety of ways. Symbolic spatial representations, such as maps and models, provide one example of a culturally mediated spatial cognition. The use of measurement tools exemplifies another way in which human cognition is aided by cultural devices. Recent evidence shows that the growth of spatial skills is greater during periods when children are in school compared to vacation periods. This effect of schooling may be carried by a number of factors—engaging children in activities (for example, in math and science classes) that require mental transformation, object assembly, and other spatial skills; teaching them how to use symbolic tools in spatial tasks; and providing them with spatial terminology.

The way in which language codes spatial relations may be related to the way people think about space. There is large variation across languages in how features of space are captured. As a result, what counts as the same spatial relation in one language can be counted as different relations in another language. For example, English uses the preposition "in" to represent the relation of containment, such as "putting an apple in a bowl" or "putting a videotape in its case." In Korean, however, these two examples would be expressed with different prepositions

because the Korean language makes a critical distinction between the relation of a loose fit (apple in a bowl) and a tight fit (videotape in a case). Cross-linguistic studies suggest that people speaking different languages have different criteria for carving up space and for forming spatial categories. The work of Bowerman and colleagues shows that even very young children are sensitive to particular spatial distinctions made in their language.

Some researchers (e.g., Levinson) raise the possibility that people from different cultures may vary not only in how they speak about space, but also in how they reason about space in nonverbal tasks. Levinson and colleagues compared the spatial performance of people whose languages vary quite radically in their approach to coding spatial relations. In particular, when coding the positions of objects in small arrays, speakers of Dutch as well as speakers of English rely on relative terms, such as left/right/front/back (e.g., the fork is on the left side of the plate). In contrast, speakers of the Mayan language Tzeltal lack relative terms and, instead rely on absolute coding by using terms equivalent to the English cardinal directions (e.g., the fork is to the North of the plate).

In a series of studies in which participants were shown an array of objects and were asked to reconstruct the array after been rotated 180°, the speakers of Tzeltal performed differently than their Western counterparts. That is, the Dutch participants preserved the egocentric relations so that the object originally shown in the leftmost position was also placed in the leftmost position following the viewer rotation. In contrast, the Mayan participants preserved the fixed bearings of the array which meant reversing the egocentric relations; e.g., the object originally shown in the leftmost position relative to the viewer was placed in the rightmost position relative to the viewer after rotation. Parallel findings have been obtained on a wide range of tasks involving spatial perception, memory, and inference. To be sure, it is possible that the observed differences in spatial cognition may not be due to language per se, they may reflect cultural differences more generally. However, these findings highlight the importance of considering the mechanisms underlying variation in spatial skills.

Summary

In sum, spatial cognition covers a wide range of skills that enable organisms to navigate their environment and to transform spatial relations among objects or their constituent parts. Some aspects of spatial cognition are shared by humans and animals. Others are uniquely human, such as the use of spatial symbols. Across species, the development of spatial skills is shaped, to a large extent, by characteristics of the physical environment. In humans, spatial cognition is also influenced by cultural tools that augment spatial abilities by providing a way to represent and think about space beyond that which is available through direct experience.

See Also the Following Articles

Cognitive Maps • Cognitive Neuroscience

Further Reading

Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. F. (eds.) (1996). *Language and Space*. MIT Press, Cambridge, MA.

Burgess, N., Jeffery, K. J., and O'Keefe, J. (eds.) (1999). *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford University Press, London.

Cheng, K. (1986). A purely geometric module in the rat's spatial representation. *Cognition* **23,** 149–178.

Choi, S., and Bowerman, M. (1991). Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition* **41,** 83–121.

Collett, T. S., and Rees, J. A. (1997). View-based navigation in Hymenoptera: Multiple strategies of landmark guidance in the approach to a feeder. *J. Comparative Physiol.* **181,** 47–58.

Diwadkar, V., and McNamara, T. (1997). Viewpoint dependence in scene recognition. *Psychol. Sci.* **8,** 302–307.

Foreman, N., and Gillet, R. (eds.) (1997). *A Handbook of Spatial Research Paradigms and Methodologies, Volume 1: Spatial Cognition in the Child and Adult*. Psychology Press, Hove, England.

Hermer, L., and Spelke, E. (1996). Modularity and development: A case of spatial reorientation. *Cognition* **61,** 195–232.

Huttenlocher, J., Hedges, L. V., and Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychol. Rev.* **98,** 352–376.

Learmonth, A. E., Newcombe, N. S., and Huttenlocher, J. (2001). Toddlers' use of metric information and landmarks to reorient. *J. Exp. Child Psychol.* **80,** 225–244.

Levine, S. C., Huttenlocher, J., Taylor, A., and Langrock, A. (1999). Early sex differences in spatial skill. *Develop. Psychol.* **35,** 940–949.

Montello, D., Lovelace, K., Golledge, R., and Self, C. (1999). Sex-related differences and similarities in geographic and environmental spatial abilities. *Ann. Assoc. Am. Geograph.* **89,** 515–534.

Newcombe, N. S., and Huttenlocher, J. (2000). *Making Space: The Development of Spatial Representation and Reasoning*. MIT Press, Cambridge, MA.

Shettleworth, S. J. (1998). *Cognition, Evolution, and Behavior*. Oxford University Press, New York.

Uttal, D. (2000). Seeing the big picture: Map use and the development of spatial cognition. *Develop. Sci.* **3,** 247–286.

Wehner, R. (1992). Arthropods. In *Animal Homing* (F. Papi, ed.), pp. 45–144. Chapman and Hall, London.

Spatial Databases

Shashi Shekhar
University of Minnesota, Minneapolis, Minnesota, USA

Pusheng Zhang
University of Minnesota, Minneapolis, Minnesota, USA

Sanjay Chawla
University of Sydney, New South Wales, Australia

Glossary

database management system (DBMS) A collection of computer programs designed to enable creation and maintenance of a large database.
geographic information system (GIS) A popular family of software designed to enable visualization, manipulation, and analysis of small spatial data sets.
object-relational database management system (OR-DBMS) A modern database management system that allows the effective modeling of spatial data using user-defined data types and operations.
query A question posed to a database.
query language A medium to express interesting questions within a computer program; a structured query language (SQL) is used for commercial database management systems.
spatial data model A type of data abstraction that hides the details of data storage.
spatial database A collection of large, interrelated spatial data items stored within a computer environment.
spatial database management system (SDBMS) A software module that manages the database structure and controls access to a large spatial database.
spatial query A question involving spatial concepts.

Spatial databases address the growing data management and analysis needs of spatial applications such as geographic information systems, remote sensing, urban planning, and natural resource management. Spatial database systems have been the focus of an active area of data management research and application for more than two decades. The research has produced major accomplishments including a taxonomy of models for space, spatial query languages, spatial storage and indexing, and query-processing and optimization strategies.

Introduction

Spatial Database and Spatial Database Management Systems

The world is undergoing an information revolution. The raw material powering this controlled upheaval—data—is being gathered constantly via sensors and other data-gathering devices. For example, the Earth Observing System developed by the United States National Aeronautics and Space Administration generates approximately 1 terabyte of data every day. Spatial data (with the term "spatial" referring to any type of space, not only the space on and above Earth's surface) are abundant in many application domains. Satellite images are one prominent example of spatial data. Information extracted from a satellite image must be processed with respect to a spatial frame of reference, possibly Earth's surface. But satellites are not the only source of spatial data, and Earth's surface is not the only frame of reference. A silicon chip can be, and often is, a frame of reference. In medical imaging, the human body acts as a spatial

frame of reference. In fact, even a supermarket transaction is an example of spatial data, if, for example, a zip code is included. A collection of interrelated spatial data is called a spatial database. Examples of publicly accessible spatial databases include the U.S. Census 2000 database created by the U.S. Census Bureau, the United States Crimes Database of the Federal Bureau of Investigation, and the County Boundary Database of the U.S. Geological Survey.

Data are housed in and managed via a database management system (DBMS). Databases and the software (computer programs and documentation) that manage them are the silent success story of the information age. They have slowly permeated all aspects of daily living, and modern society would come to a halt without them. Despite their spectacular success, the prevalent view is that a majority of the DBMSs in existence today are either incapable of managing spatial data or are not user friendly when doing so. Why is that true? The traditional role of a DBMS has been that of a simple but effective warehouse of business and accounting data. Information about employees, suppliers, customers, and products can be safely stored and efficiently retrieved through a DBMS. The set of likely queries is limited, and the database is organized to answer these queries efficiently.

Compared to the data in traditional DBMSs, spatial data are more complex because they include extended objects, such as points, lines, and polygons. The relationships among spatial objects are often implicit, i.e., relationships such as "overlap," "intersect," and "behind" are not explicitly modeled in a DBMS. A traditional database management system must be specialized to meet the requirements of spatial data. By doing so, spatial domain knowledge can be extrapolated to improve the overall efficiency of the system. Thus, simply speaking, a spatial database management systems (SDBMS) may be defined as a software module that manages the database structure and controls access to data stored in a spatial database. More specifically, a spatial database management system is a software module that can work with an underlying database management system, e.g., an object-relational database management system (OR-DBMS). It supports multiple spatial data models, commensurate with spatial data types and operations. It also supports spatial indexing, efficient algorithms for spatial operations, and domain-specific rules for query optimization.

Commercial examples of spatial databases management include IBM/Informix's DataBlade module (e.g., 2D, 3D, and Geodetic); Oracle's Universal server, with either the Spatial Data Option or the Spatial Data Cartridge; and ESRI's Spatial Data Engine. PostGIS/PostgreSQL is a research prototype example of spatial database management systems. The functionalities provided by these systems include a set of spatial data types, such as point, line segment, and polygon, and a set of spatial operations, such as inside, intersection, and distance. The spatial types and operations may be made part of an extensible query language such as structured query language (SQL), which allows spatial querying when combined with an OR-DBMS. The performance enhancement provided by these systems includes a multidimensional spatial index and algorithms for spatial access methods, spatial range query, and spatial joins.

Geographic Information Systems and Spatial Database Management Systems

A geographic information system (GIS) provides a rich set of operations over few objects and layers, whereas an SDBMS provides simpler operations on sets of objects and sets of layers. For example, a GIS can list neighboring countries of a given country (e.g., France), given the political boundaries of all countries. However, it will be fairly tedious for a GIS to answer set queries (e.g., "list the countries with the highest number of neighboring countries" or "list countries that are completely surrounded by another country"). Set-based queries can be answered easily in an SDBMS. SDBMSs are also designed to handle very large amounts of spatial data stored on secondary devices (magnetic disks, compact disk—read-only memory, jukeboxes, etc.) using specialized indices and query-processing techniques. SDBMSs inherit the traditional DBMS functionality of providing a concurrency-control mechanism to allow multiple users to access shared spatial data simultaneously, while preserving the consistency of that data. A GIS can be built as the front end of an SDBMS. Before a GIS can carry out any analysis of spatial data, it accesses that data from an SDBMS. Thus, an efficient SDBMS can greatly improve the efficiency and productivity of a GIS.

Space Taxonomy and Data Models

Space Taxonomy

"Space" is indeed the final frontier, not only in terms of travel on and beyond Earth's surface, but also in the difficulty of capturing the meaning of the word and the concept with a simple, concise description. Consider the following refrain echoed by hapless drivers all over the world: "I don't remember how far Mike's house is: once I am nearby, I might recall on which side of the street it lies, but I am certain that it is adjacent to a park." This sentence gives a glimpse of how the human brain (mind) structures geographic space. We perform poorly in estimating distances, maybe only slightly better in retaining direction and orientation, but fairly capably when it comes to remembering topological relationships such as "adjacent," "connected," and "inside." Topology,

a branch of mathematics, is exclusively devoted to the study of relationships that do not change due to elastic deformation of underlying space. For example, if there are two rectangles, one inside the other, or both adjacent to each other, drawn on a rubber sheet, and if the sheet is stretched, twisted, or shrunk, the named relationships between the two rectangles will not change!

Another clue about how the human mind organizes space is revealed by examining how languages communicate concepts. For example, the shapes of objects are major determinants in how objects are described. Is that the reason why we have trouble accepting a whale as a mammal and a sea horse as a fish? Objects are described by nouns, and languages have as many nouns as there are different shapes. On the other hand, the spatial relationships among objects are described by prepositions, which encode very weak descriptions of shapes. For example, in saying that "Coffman Student Union is to the southeast of Vincent Hall," the shapes of the buildings play almost no role in the relationship "southeast." We could easily replace the buildings with coarse rectangles without affecting their relationship.

Space taxonomy refers to the multitude of descriptions that are available to organize space: topological, network, directional, and Euclidean. Depending on why we are interested in modeling space in the first place, we can choose an appropriate spatial description. Table I provides an example of a spatial operation associated with a different model of space. It is important to realize that no universal description (model) of space can answer all queries.

Spatial Data Model

A data model is a type of data abstraction that hides the details of data storage. For example, Minnesota is the "land of ten thousand lakes." How can these lakes be represented? An intuitive, direct way is to represent each lake as a two-dimensional region. Similarly, a stream, depending on the scale, can be represented as a one-dimensional curved line, and a well site can be represented by a zero-dimensional point. This is the object model. The object model is ideal for representing discrete spatial entities such as lakes, road networks, and cities. The object model is conceptual; it is mapped into the computer using a vector data structure. A vector data structure maps regions into polygons, curves into polylines, and points into points.

Table I Different Types of Space Descriptions

Family of descriptions	*Example concept*
Topological	Adjacent
Network	Shortest path
Directional	North of
Euclidean	Distance

The field model is often used to represent amorphous or continuous entities (e.g., clouds or temperature maps). A field is a function, which maps the underlying spatial reference frame into an attribute domain. For temperature, popular attribute domains are Celsius and Fahrenheit. The raster data structure implements the field model on computers. A raster data structure is a uniform grid imposed on the underlying space. Because field values are spatially autocorrelated, i.e., smoothly varying over space, the value of each cell is typically the average of all of the field points that lie within the cell. Other popular data structures for fields are the triangulated irregular network (TIN), contour lines, and point grids.

Spatial Query Languages

Spatial Query

Once the design of a database is complete and a DBMS is chosen to implement the database, we can carry out queries to the database. It is worthwhile to view a query as a question posed to a database. A query is expressed in a high-level declarative manner, and the algorithms needed to answer the query are not specified in the query. For example, typing a keyword in a search engine (e.g., Google or Yahoo Search) forms a query corresponding to the question "Which documents on the web contain the given keyword?" Queries involving spatial concepts are called spatial queries. For example, the query "List the counties with populations over 500,000 in the United States" is an example of a nonspatial query. On the other hand, the query "List crime hot spots within 10 miles of downtown Minneapolis" is an example of a spatial query because it uses the spatial concept of distance.

Spatial Query Languages

A query language is a medium used to express interesting questions about data, and queries expressed in query languages retrieve answers from databases. Query languages include natural languages, such as English, which can express almost all queries, and computer languages, such as Java, which can express computable queries. Many query languages often restrict the set of possible queries. For example, an interactive map may allow users to drag a mouse pointer over or to point and click on locations to ask questions about different spatial properties of those locations. Although drag and point-and-click methods are user-friendly features, users are limited in the kinds of queries they can make. Structured query

language (SQL) was thus designed to express set-oriented queries.

SQL is now the standard query language for commercial relational databases. SQL is a comprehensive database language, and it has statements for data definitions, manipulations, and controls. Queries are expressed using SQL constructs for data manipulations. The queries are expressed in a high-level declarative manner, so users specify only what the results are to be, leaving how to execute the queries to the DBMS. SQL is text based and is mostly geared toward set-based operations. For example, the query "List the counties with populations over 500,000 in the United States" finds a set of counties with populations of more than 500,000.

In 1999, the Open GIS Consortium (OGC), led by major GIS and database vendors, established a specification for incorporating spatial data types (e.g., polygons) and operations (e.g., distance) in SQL. The OGC extensions are based on the object model and support multiple spatial data types, such as the point, curve, line string, polygon, and surface. The operations supported in the OGC extensions fall into three categories:

1. Basic operations applicable to all geometry data types. For example, a spatial reference returns the underlying coordinate system where the geometry of the object was defined. Examples of common reference systems include the well-known latitude and longitude system and the Universal Traversal Mercator.
2. Operations that test for topological relationship among spatial objects. For example, an intersect operation tests whether the interior of two objects has a nonempty set intersection.
3. General operations for spatial analysis. For example, a distance operation returns the shortest distance between two spatial objects.

Others efforts to standardize spatial extensions to SQL include SQL/MM and ISO/TC 211 Geographic Information/Geomatics.

Examples can be used to illustrate query operations. Say there are three data sets in a spatial database, Country, City, and River. The Country data include the name, continent, population, and spatial footprint for each country in the world. The City data consist of the name, country, population, and spatial footprint for each city in the world. The River data are made up of the name, origin, length, and spatial footprint for each river in the world. The following example queries can be expressed easily in SQL using the OGC spatial data types and operations:

Query 1: List the name, population, and area of each country.
Query 2: Find the names of all countries that are neighbors of the United States.
Query 3: For each river, find the countries through which they pass.
Query 4: For each river, identify the closest city.
Query 5: Find the shortest path from Minneapolis to Duluth.

Spatial Storage and Indexing

Database management systems have been designed to handle very large amounts of data. This translates into a fundamental difference in how algorithms are designed in a GIS data analysis vs. a database environment. In the former, the main focus is to minimize the computation time of an algorithm, assuming that the entire data set resides in the main memory of a computer. In the latter, emphasis is placed on minimizing the sum of the computations using data in the main memory and the input/output (I/O) to fetch data from secondary storage data. This is because, despite falling prices, the main memory is not large enough to accommodate all of the data for many applications.

It is worthwhile to view the secondary storage device as a book. The smallest unit of transfer between the secondary storage and main memory is a page, and the records of tables are like structured lines of text on the page. A query issued by a user essentially causes a search within a few selected lines embedded in pages throughout the book. Some pages can reside in the main memory, and pages can be fetched only one at a time. To accelerate the search, the database uses an index. Thus, in order to search for a line on a page, the DBMS can fetch all of the pages spanned by a table and can scan them line by line until the desired record is found. The other option is to search in the index for a desired keyword and then go directly to the page specified in the index. The index entries in a book are sorted in alphabetical order. Similarly, if the index is built on numbers, such as social security numbers, the numbers can be numerically ordered.

The R-tree data structure was one of the first indexes specifically designed to handle multidimensional spatial objects on secondary storages, such as disks. The R-tree groups together objects in spatial proximity on the same leaf nodes of a tree-structure index. Figure 1 shows an example of how the R-tree organizes extended objects. Because a leaf node can point only to a certain number of objects, minimum bounding rectangles are applied to divide the space. The internal nodes of R-trees are associated with rectangles, the areas of which cover all of the rectangles in the subtree. Hence, R-trees can easily answer queries; they can find all objects in a given area by limiting the tree search to those subtrees with rectangles that intersect with the given query area. Note that the minimum bounding boxes may have

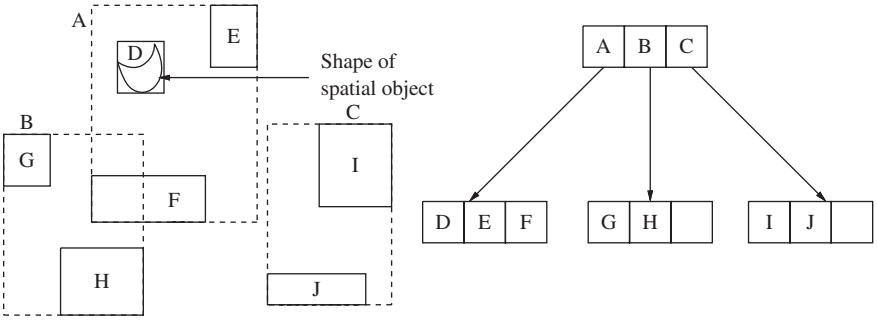


Figure 1 Use of an R-tree to organize spatial objects.

overlapping spatial areas (e.g., rectangles A and B in Fig. 1); this type of problem is handled in different ways by the many different variations of R-trees. Other spatial storage structures include the grid file, quad-trees, and their variations.

Query Processing

As noted earlier, a database user interacts with the database using a declarative query language such as SQL. The user specifies only the result desired, and not the algorithm to retrieve the result. The DBMS must automatically implement a plan to execute the query efficiently. Query processing refers to the sequence of steps that the DBMS will initiate to process the query.

Queries can be broadly divided into two categories: single-scan queries and multiscan queries. In a single-scan query, a record (tuple) in the table (relation) being queried has to be accessed at most once. Thus, the worst-case scenario, in terms of time, is that each record in the table will be accessed and processed to verify whether it meets the query criterion. The spatial query introduced earlier, "List crime hot spots within 10 miles of downtown Minneapolis," is an example of a single-scan query. The result of this query will be all crime hot spots that intersect a circle of 10 miles from downtown Minneapolis. This query is also an example of a spatial range query, whereby the "range" refers to the query region. Here the query region is the circle of radius 10 miles. If the query region is a rectangle, the spatial range query is often referred to as a window query.

A join query is a prototype example of a multiscan query. To answer a join query, the DBMS has to retrieve and combine two tables in the databases. If more than two tables are required to process the query, then the tables may be processed in pairs. The two tables are combined, or "joined", on a common attribute. Because a record in one table can be associated with more than one record in the second table, records may have to access more than once to complete the join. In the context of spatial databases, when the joining attributes are spatial in nature, the query is referred to as a spatial-join query.

The SDBMS processes range queries using the filter–refine paradigm. This is a two-step process. In the first step, the objects to be queried are represented by their minimum bounding rectangles. The rationale is that it is easier (computationally cheaper) to compute the intersection between a query region and a rectangle rather than between the query region and an arbitrary, irregularly shaped spatial object. If the query region is a rectangle, at most four computations are needed to determine whether the two rectangles intersect. This is called the filter step, because many candidates are eliminated by this step. The result of the filter step contains the candidates that satisfy the original query. The second step is to process the result of the filter step using exact geometries. This is a computationally expensive process, but the input set for this step, thanks to the filter step, often has low cardinality, i.e., few candidates are to be checked in the second step.

Summary

A SDBMS manages the database structure and controls access to data stored in a spatial database. The SDBMS plays a prominent role in the management of query of spatial data. Spatial data management is of use in many disciplines, including geography, remote sensing, urban planning, and natural resource management. The Open GIS Consortium has established a specification for incorporating two-dimensional spatial data types (e.g., point, curve, and polygon) and operations. Using the OGC specification, common spatial queries can be posed in structured query language. The spatial indexes (e.g., R-trees) were designed to facilitate efficient accesses to spatial databases, and the filter–refine strategy can be applied to process spatial queries efficiently.

See Also the Following Articles

Geographic Information Systems • Spatial Autocorrelation • Spatial Pattern Analysis

Further Reading

Cressie, N. (1990). *Statistics for Spatial Data.* Wiley, New York.

Eisenberg, A., and Melton, J. (2001). SQL multimedia and application packages (SQL/MM). *SIGMOD Rec.* **30**(4).

Elmasri, R., and Navathe, S. (2004). *Fundamentals of Database Systems.* 4th Ed., Addison Wesley, Boston, MA.

Guttman, R. (1984). R-Tree: A dynamic index structure for spatial searching. In *Proceedings of ACM SIGMOD Conference (Boston)*, pp. 47–57. Assoc. for Computing Machinery, Washington, D.C.

International Organization for Standardization. (2004). *ISO/TC 211 Geographic Information/Geomatics*. Available on the Internet at www.isotc211.org

Kanth, K., Ravada, S., and Abugov, D. (2002). Quadtree and R-tree indexes in Oracle spatial: A comparison using GIS data. In *Proceedings of ACM SIGMOD Conference* (Madison, Wisconsin), pp. 546–557. Assoc. for Computing Machinery, Washington, D.C.

Laurini, R., and Thompson, D. (1992). *Fundamentals of Spatial Information Systems.* Academic Press, London.

Longley, P., Goodchild, M., Maguire, D., and Rhind, D. (2002). *Geographic Information Systems and Science.* John Wiley & Sons, New York.

Open GIS Consortium. (1999). *OGIS Simple Features Specification for SQL*. Available on the Internet at www.opengis.org

PostGIS. (2003). *PostGIS Online Manual*. Available on the Internet at http://postgis.refractions.net

Rigaux, P., Scholl, M., and Voisard, A. (2000). *Spatial Databases with Application to GIS.* Morgan Kaufmann Publ., Burlington, MA.

Samet, H. (1990). *The Design and Analysis of Spatial Data Structures.* Addison Wesley, Boston, MA.

Shekhar, S., and Chawla, S. (2003). *Spatial Databases: A Tour.* Prentice Hall, Englewood Cliffs, NJ.

Shekhar, S., Chawla, S., Ravada, S., Fetter, A., Liu, X., and Lu, C. (1999). Spatial databases: Accomplishments and research needs. *IEEE Transact. Knowl. Data Eng.* **11**(1), 45–55.

Worboys, M. (1995). *GIS: A Computing Perspective.* Taylor & Francis, Bristol, UK.

Spatial Discounting

Bruce Hannon
University of Illinois, Champaign-Urbana, Champaign, Illinois, USA

Glossary

biomagnification The increased concentration of a chemical found in species as one moves up the food chain.
discounting To reduce value (positive or negative) in proportion to a given metric.
intragenerational Between members of the present generation.
space In this article, restricted to the surface of the earth.
utility A general measure of usefulness or desirability as used in economics.

Spatial discounting is a parallel process to time discounting. Just as people devalue actions in their future, they devalue present actions at a distance. People are both socially myopic and parochial.

Introduction

There exists in people a desire to be near things that they consider "good" and to be far from things they consider "bad." People prefer to live near schools, churches, and grocery stores and far from sewage treatment and power plants, landfills, and prisons; they prefer to live near the ocean and far from harsh climates. In the workplace, people prefer higher offices and shun basement offices. At home, people prefer living on top of a hill to the bottom, upstream to downstream. People prefer to live away from a national boundary rather than close to it. People tend to discount their fear of (desire for) an object the further they are away from it—a negative (positive) geographic discounting. Animals and plants exhibit spatial discounting as well. Spatial discounting is one of five distinguishable forms of discounting; the others are the discounting of time, uncertainty, the insensible, and the interpersonal.

Basic Requirements for the Concept

For simplicity, assume that the geographic discount rate is constant with respect to time and distance, and that the geographic discounting function is one of exponential decline with respect to distance. The exponential function is mathematically convenient, but its use implies a rather specific kind of feedback relationship with the quantity being discounted.

This is not as constraining a requirement as it might first seem. Consider the temporal accumulation process in finance. The amount of interest money that accrues this year on an amount of invested money depends directly on the amount of invested money itself; at the end of the first year, the base amount has grown by the amount of interest, so the second year's interest money is larger than it was in the first year, and so on. The initial investment grows at an accelerated rate as interest accumulates. The investment amount provides a feedback signal that controls the size of the next period's accumulation. The reverse of this process is called time discounting: a stream of future payments to a person is said to have a unique value at this moment because that person is assumed to have a known discount rate. Therefore, he or she values a promised payment in the future in proportion to its size and how far into the future the payment is supposed to be. The farther a particular payment is seen in the future, the less he or she values it now; that is, the more the person discounts it. The interesting part of the devaluation of future payments into today's value is that the devaluation is proportional to time.

When a separate factor is developed to represent spatial discounting, it is an admission that it is convenient to do so, it is appropriate to do so, or both. It can be argued

that any function that represents the benefits and costs (the utility) of a living organism should contain all geographical discounting effects. But this may not be possible or practical, even though the benefits and costs are functions of distance. The geographical discounting function being proposed should multiply the net benefits to represent the psychological state of the organism and any of its distance-related perceived risks. In such a use, the geographical discounting function exactly parallels the time discounting function so common in economic optimization applications.

The basic question then becomes, "Exactly what is being discounted with respect to distance and why should that discounting process be exponential?" As examples of geographic discounting in humans, animals and plants were evaluated and no units were found that were commensurate with those economists use. That is, not everything that is geographically discounted in living systems can be cast into the same units of measure. In humans, that which is discounted with distance seems to be fear or desire: the fear of something dangerous, or the desire to be in some place or to have some desired sensory input from some particular place. The further the object of threat or desire is from the individual, the more that object is discounted. Clearly, in this sort of discounting, one can find positive and negative expressions, both of which decline in importance as the distance from the object of attention grows.

This sort of positive and negative expression of desire and fear is also demonstrable over time. Such thoughts as "I can't wait until I see her again" or "I dread the thought of tomorrow's dental appointment" are good examples. People usually do not quantify any aspect of time value except the positive financial one, but these alternatives are possible.

Finally, the difference between geographical discounting by an individual and by a group must be clarified. The rate of geographical discounting will vary for an individual over time, with increasing information and with changes in personal outlook. While such changes may be useful to individuals and help them better understand their perceptions of the landscape, the policy implications of geographical discounting are most interesting. Consequently, the focus is on the rates as they might be revealed by typical groups of people. These rates should represent a meaningful average for such groups, and they are likely to change more slowly and be more comparable between different groups in similar situations.

The Nature of Spatial Discounting

Early studies of willingness to pay for the avoidance of aesthetic or environmental threat did not include distance between the observer and the object of concern in the criteria. More recent literature must be consulted for references to the quantified connection between valuation and distance.

People who object to the proposed nearness of generally objectionable activity are seen as being fixed to geographic location and wishing to force the activity away from them. This is one view of a sense of place, from the home place. But suppose that someone is aware of a desired home place but is presently located away from it. That person then has a desire to be nearer the home place, and his or her desire varies with the separation distance. It seems reasonable to assume that those who move frequently will not develop a sense of place, and consequently, that their distance discounting is very significant. Without a sense of place, residents tend to allow deterioration of that place without resistance. Sense of place and its attribute, geographical discounting, appears to be a necessary condition for environmental concern, at least in the novice environmentalist. Genuinely nomadic peoples, however, have a low rate of geographical discounting. Although they move frequently, they repeatedly circulate through a series of spaces, the whole of which can be seen as their territory or home.

Part of the reason people discount distance may be derived from their perception of risk: the closer an undesirable object is, the greater the risk of exposure to its undesirable aspects. If this were the case, then, would it be easier to locate an undesirable object in an area where the perceived risks are already high than in an area where risks are seen as low? The percentage increase in risk is less in the first instance. This view of risk perception is confounded by the possibility that the areas where general risk of personal harm is higher are urban areas, in contrast with rural areas. While the percentage increase in risk individually perceived from the proposed location of a certain object is lower in the city than in the rural areas, the collective perception of risk in the urban area may be higher. The higher the collective perceived risk, the greater the likelihood of the emergence of an opposition leader.

Early interest in geographical discounting comes from the often-indicted "not in my backyard" (NIMBY) behavior on the part of those who publicly object to the proposed location of a generally undesirable object near them. The condemnation of geographic selfishness is reminiscent of the biblical excoriation of usury. But who would not object to the proposed location of a hazardous waste incinerator near their home? Aren't most people likely to exhibit NIMBY behavior under the right circumstances? Aren't most, if not all, zoning regulations just a formalization of this distancing desire of the local population? If people have sufficient income, they will not live near factories, landfills, sewage treatment

plants, etc. Conversely, commercial and industrial firms will often pay premiums for property near facilities similar to their own rather than locate in a cheaper, perhaps more remote, but allowable location.

The NIMBY response is an expression of territoriality and has been noted since the beginning of the human record. In China in the 6th century BC, the Imperial Palace was situated at the city center, with declining social prestige assigned to those living farther from the palace. Those at the very edge of the city were judged the most uncivilized, the least human, living in the "zone of cultureless savagery." The round city of Mansur, in the 8th century, placed the palace and the mosque at its center and the prison in the outer wall. The Yurok Indians of northern California lived at "the center of the world" with the unknowable land and ocean "beyond the world." In the late Middle Ages, the city was viewed as a sacred place, and the surrounding wilderness was seen as "profane." By the 19th century, the landscape of the yeoman was perched between the "profane" city and the "profane" wilderness. Today, the wilderness is considered a place of innocence and purity, while the city is still profane and the suburb is the hoped-for combination of utility and sacredness.

NIMBY behavior is a natural reflection of people's insistence to be near objects of desire and far from objects of fear. It is no more a dispensable part of human behavior than is eating when hungry or sleeping when tired. It is a behavior very much like the tendency to devalue useful items when that use is delayed.

Evidence from Opinion Surveys

One of the first studies that attempted to measure the discounting of distance was reported by Mitchell and Carson in 1986. They surveyed the degree of acceptance of nuclear and electric power plants, a hazardous waste disposal facility, a large office building, and a factory within various distances from home. The data on power plants was translated into the fraction of the rejecting population (i.e., a measure of the level of concern) and its variation with distance from home (see Fig. 1). About 9% of the population indicated that they did not want any coal plants built anywhere; the comparable figure for nuclear plants was 29%. Of the remaining population, about 60% did not want either type of plant within a mile of their home. By statistically fitting the response data, it can be shown that the rate at which rejection declined with distance was about 4.3% per mile for the coal plant and about 2% per mile for the nuclear plant. The goodness of fit was very high: $r^2 = 0.97$ and 0.98.

These geographic discount rates clearly reflect the declining level of concern as distance from the proposed object of concern increases. They are not the same number, however, and neither would the time discount rates regarding these proposals be the same. For a given distance, the time discount rate would probably reveal a greater level of time discounting of a coal plant compared to a nuclear plant. Individuals with similar incomes do not necessarily express the same level of time discounting about potential financial gain even when they

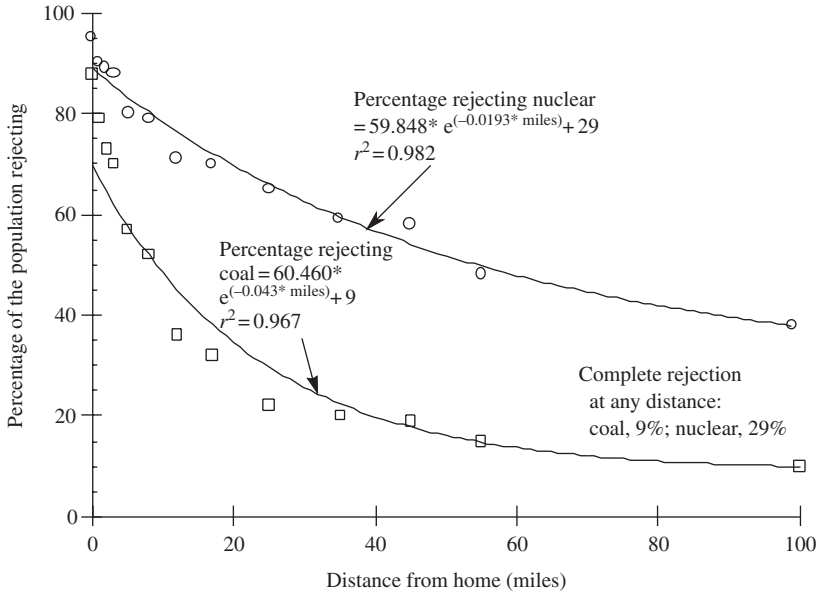


Figure 1 Percentage of the population rejecting a hypothetically proposed electrical generating plant with various distances from home. Data adapted from Mitchell and Carson (1986).

see the same risk of investment. People discount the delay of food consumption much more highly than they do delay in financial return. Therefore, the appearance of two different geographic discount rates regarding different facilities with the same output does not seem unreasonable.

The geographic discount rate has use in economic analysis. For example, the utility of the individual electricity user contains functions of time, price, and distance from the generator. To properly assess the social welfare of electricity, an analyst must therefore include a geographical discounting factor, such that the utility decreases as the distance to the generator decreases.

Mitchell and Carson also found that people discounted distance from a proposed hypothetical hazardous waste plant at about the same rate as they did a nuclear power plant. They geographically discounted a large factory at about the same rate as they did a coal plant. They also discounted a 10-story office building less than any of these facilities. The authors did not elaborate on these sorts of differences, but they did distinguish between the hazardous waste plant and the other facilities, in that the waste plant siters do not have the complete freedom to locate where they wish. Although the authors saw this problem as a lack of property rights, it is possible to view the issue somewhat differently.

Waste facilities are difficult to site for psychological reasons as well: people do not want to be reminded regularly that their consumption activities generate a socially negative signal. They conform to the old maxim "out of sight, out of mind." This is an aspect of what is called sensual discounting: "If I can't see or smell the facility, I tend to sensually discount it." But the change in the rejection rate as distance increases between home and a waste plant, for example, is pure spatial discounting.

The survey data in Fig. 1 are hypothetical. If levels of concern about such facilities after they are installed are measured, a slightly different picture emerges. Those living 1.4 km from a nuclear plant expressed nearly the same level of concern as those living 10 km from the plant. In another study of toxic waste sites, the perceived rather than the actual distance was inversely correlated with level of concern.

People who want to be away from such facilities as power plants can also be seen as expressing territorial behavior, in much the way that it can be imagined some animals do. If such territorial behavior is manifest, it should be found that the value of private property can be influenced by the geographic distance from certain objects, with values rising with increasing proximity to objects of desire (e.g., elementary schools, shopping centers) and falling with increasing proximity to objects of negative concern (e.g., high voltage transmission lines, noisy business districts). In fact, the actual distribution of property value changes may reflect the combination of negative and positive geographic discounting.

Evidence from Real Estate Markets

If people strongly dislike or prefer certain objects located at specific places, then variations in their property values should reflect this dislike or preference. All other real estate variables being unchanged, the property values should increase with distance away from the object of popular scorn, or decrease with distance away from an object of popular desire. Some objects have both positive and negative values to residential property owners. People may like living near a shopping center, for example, but not next to it. Certain attributes of these sorts of objects are positive (e.g., minimal travel time and distance) and some are negative (e.g., noise, nighttime lighting, lack of trees). So, variations in property values should reveal an optimal distance between home and certain objects. Other objects, such as nuclear power plants, seem to have no apparent positive features regarding the nearness to people. In the complex urban environment, it may be the nearest neighbors to generally undesirable objects who are selected for their unique indifference to the object. For example, people who live near busy shopping centers may seldom use their yards or may seldom open their windows, i.e., they may be focused on the interior of their place, disregarding the physical environment outside their home or office. Such people reveal very high geographic discount rates. Thus, people's geographic discount rates, on average, may vary with time (i.e., people may be able to adapt to nearly anything, given enough time).

Furthermore, the complex urban environment could present such an array of positive and negative objects that it could be nearly impossible to sort out the geographical discount rate because of the overlapping effects. Some financial experts believe this problem is best approached by mapping the unexplained residual of the hedonic equation (object price = function of the attributes) and looking for geographic and other patterns. In this way, the explanatory power of the equation can be improved, perhaps reducing the number of variables required in the equation.

Consider the impact of a recently completed shopping center on housing values. The statistical model indicates that there is an apparent negative effect up to about 1500 feet from the shopping center. Beyond that distance, up to about 4000 feet, housing sale prices were higher than historical values. This result indicates that people negatively discount the increased noise of autos and trucks and the nighttime lighting at the shopping center. But people also positively discount the convenience of the shopping center. These effects are difficult to separate in a statistical model, although one might be able to base the negative effects on straight-line distances and the positive effects on street-based distances. These discount effects and their net impact on the property values are illustrated in Figs. 2 and 3.

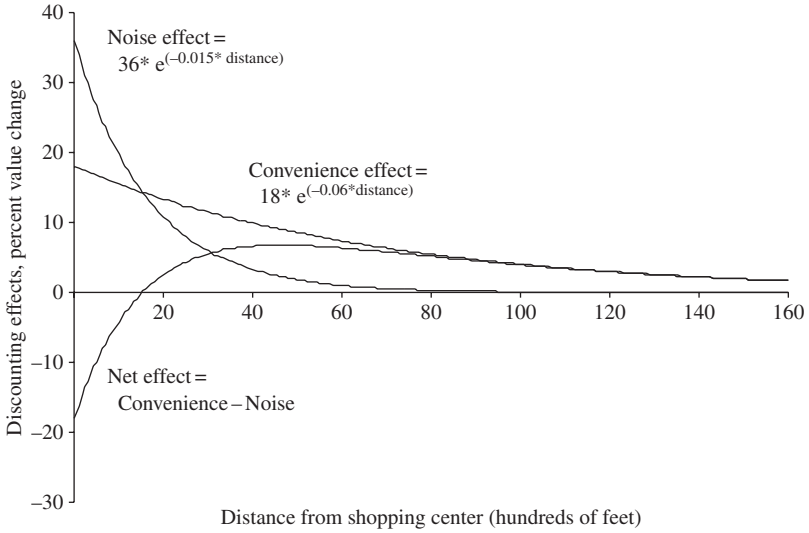


Figure 2 The effects of positive and negative geographical discounting on the real value of housing prices. An approximate fit to the results of Colwell *et al.* (1985).

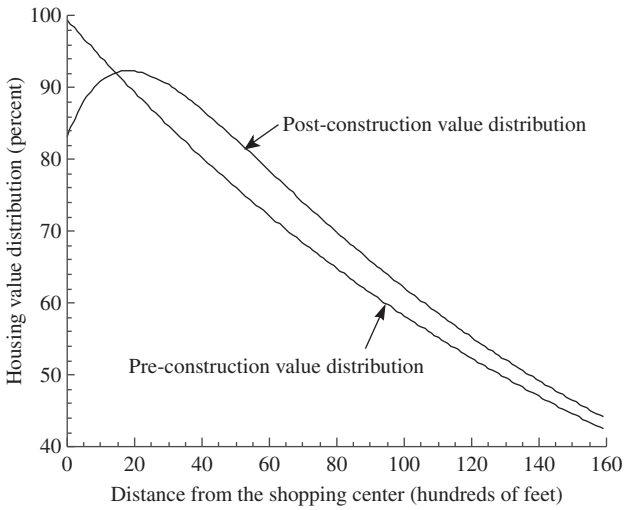


Figure 3 The combined effects of positive and negative geographical discounting on the real value of housing prices, comparing the values before and after the construction of a shopping center. An approximate fit to the results of Colwell *et al.* (1985).

Figure 2 shows a rough estimate of the discounting functions and their net effects on property values for the shopping center issue. Noise is discounted at a rate of 6% per 100 feet and convenience is discounted at a rate of 1.5% per 100 feet, approximately. The array of net value change should be added to the distribution of property values that existed before the shopping center proposal was announced to get an estimate of the percentage change in the post-shopping center property values (see Fig. 3). The agreement with Colwell's results is only approximate, but these graphs are intended to show a possible decomposition into two controlling effects. The curves in both Figs. 2 and 3 show that the neutral distance is about 1500 feet from the shopping center. At this distance, the negative effects of the shopping center on housing valuation are just offset by its positive effects.

In another study, the change in residential property values with distance from the central business district of a community (CBD) is enumerated. Here, the distance effect was $\exp(-0.044 \times \text{distance from the CBD})$, a straightforward exponential decline in value at the rate of 4.4% per mile (lot area and appreciation held constant).

The use of property value variation with distance forms the basis of some novel arguments. If people feel so strongly about nearness to elementary schools that their residential property values rise as one nears the school, then desegregating busing programs may result in an uncompensated property loss.

It is possible to generate a theory for the optimal size of a group's territory with the concept of spatial discounting. Suppose that the target population has no time preference and that the territory is to be established within a circular field of uniform resources, desired by the people in the group. What is the optimal territory size? Let $B(D)$ be the known benefit (e.g., food, shelter, reproductive potential) derived from this territory; D is distance from the center of the circular territory to the edge. The benefit is a known composite measure of food, shelter, or other desired property of the territory. Let $C(D)$ be the known cost of harvesting and defending the territory. Assume that the

distance preference is $e^{-\delta D}$, where δ is the known spatial discount rate. Then the optimality problem becomes:

$$\max \int_0^D [B(D) - C(D)]e^{-\delta D} dD$$

the solution of which will yield the optimum (circular) territory size. $B(D)$ is a monotonically rising function of D, in the same units as $C(D)$, whose derivative is positive but declining with distance. This means simply that the benefit for the animal grows as D increases, but gains in benefit per unit D decrease as D increases. The net gain must be spatially discounted ($e^{-\delta D}$) because the animals act as though they fear their boundaries, an effect not included in the benefit or cost functions. Given B, C, and δ, the optimal territorial size can be found. This procedure is a relatively simple example of the way in which the concept of geographic or spatial discounting might be used. Certainly, if the resource is not uniform and a set of physical barriers block the travel in particular directions, the shape of the potential territory becomes very complex, and the optimal problem would have to be cast in two dimensions and numerically solved.

An Economic Basis for Spatial Discounting

The debates of the early part of the last century on the ethics of time discounting focused on the problem of utility discounting and its moral implications. In particular, they concerned the notion that preferences over time reflect not only individual or collective impatience to consume, but also a judgment about the responsibility that today's decision-makers have for the future consequences of their actions. Positive rates of time preference were taken to imply that members of the current generation choose to bear strictly limited responsibility for any harm they might inflict on future generations. This is certainly a motivation for ethical objections to discounting. By analogy, positive rates of spatial preference, the spatial equivalent of time preference, reflect not only a preference for consumption at home, but also a judgment about the responsibility that local decision-makers have for the distant consequences of their actions. Because positive rates of spatial preference imply a preference for local consumption relative to distant consumption, they indicate that the relative deprivation of distant members of the present generation may be a locally desirable outcome. The reference point for evaluating consumption flows that are separated in space and time is consumption "here and now."

The utility rate of time discount defines the rate at which present consumption is preferred to future consumption. By analogy, the utility spatial discount rate should capture people's preferences for consumption at a given location relative to consumption at some distance from that location. This may refer either to their own consumption or to consumption by other members of the present generation. As is the case for time discounting, positive rates of spatial preference imply an ethical judgment about the responsibility of the decision-maker for the welfare of others. More particularly, positive rates of spatial preference imply that people care more about actions that are close to them than about those that are distant. The picture is obviously complicated by familial, tribal, ethnic, linguistic, or political affiliation (as it is with time discounting). People tend to care more about their kith and kin than about others. They may also be very tightly constrained in their ability to implement the concern they have for the well-being of those outside of their own location. A positive rate of spatial preference implies that consumers place greater weight on home interests than on outside interests. Home in this case may be the household, village, city, region, or nation.

In 2001, Perrings and Hannon gave a detailed technical explanation of the warranted or socially desirable rate of spatial discounting. It is the spatial decay rate of the effects of local actions, and it is also the spatial rate at which consumption recovers from the effects of local actions. They also pointed out that the effects of spatial discounting are not necessarily reflected in local market prices.

A rate of time or spatial preference will be said to be warranted if distant effects discounted at that rate may be fully compensated, i.e., those effects are valued at their marginal social cost. Utility rates of time discount have been argued to be ethically neutral or sustainable only if they do not exceed the net growth rate of the capital. The warranted rate of spatial preference is analogous to that time-based discount rate. Positive utility rates of spatial preference may be said to be ethically neutral only if they do not exceed the rate at which the spatially dispersed effects of local activity decay with distance. In the case of environmental effects, this can be thought of as the rate at which they are diffused or filtered out by the environment.

It is assumed for simplicity in technical analysis that spatial effects decay smoothly, continuously, and monotonically. This is a strong assumption, though no stronger than the analogous assumptions made about time-based rates of growth or regeneration, depreciation, or decay. In reality, many spatial and temporal effects are far from smooth. There are certainly cases where the effects of current activities are not diffused over time or space. Biomagnification, for example, can lead to the concentration of effects at points distant in both time and space. Emissions to rivers can lead to the concentration of pollutants in downstream sinks. Mixing can mean that emissions to air can affect all locations equally. But for

the most part, the adverse effects of local activity tend to be a decreasing function of distance from the source. Environmental effects are transmitted through the flows of mass and energy in biogeochemical cycles. All such flows involve some energy or material cost; that is, they involve losses in transmission. Not all of the SO_2 that is emitted from Ohio, for example, lands in New York and Boston. Some drops out over the Great Lakes. Not all the toxins entering a food chain are consumed by species at the top of the chain; some are transformed and excreted by those at an intermediate point in the chain. There are various determinants of the filtering effect of ecosystems on flows of mass and energy through the system, and not all effects are filtered at the same rate. In reality, the decay of effects is neither smooth and continuous nor monotonic. Nonetheless, it is useful to think in terms of a diffusion rate that is uniform at least within the boundaries of a given ecosystem.

Economists are interested in the warranted spatial rate of discount associated with this form of diffusion. The higher the actual rate of discount, the more the distant costs of local emissions will be ignored. Intuitively, effects that are discounted at the rate at which the pollutant diffuses will be neutral in their effect. Discount rates that exceed the rate of diffusion imply that the distant costs of local emissions will not be fully taken into account.

Summary and Conclusions

Spatial discounting derives from a sense of place. How large that place is and how well it is occupied depends on the relative strength of the individuals involved and the harshness of their environment. Spatial discounting can be seen as a reaction to threats to those in the home place. The manifestation of the discounting process is seen in opinion polls and property value distribution.

There is physical evidence of both positive and negative spatial discounting. The paucity of such data is an indication that the spatial discounting process is still thought of as a character defect rather than a basic attribute of human nature. The rates of discounting are assumed to be constant, and the discounting functions are assumed to be declining exponentials. These assumptions were made to retain simple displays of the idea, but the fit to actual data of such functions is remarkably good. The net effect of the geographical discounting process is the interplay of both negative (e.g., noise) and positive (e.g., convenience) effects.

It should not be surprising that high spatial discount rates have the potential to prejudice the well-being of distant members of the present generation in the same way that high time discount rates prejudice the well-being of members of future generations. Time discount rates above the warranted rate of regeneration/assimilation imply a myopic approach to the management of environmental resources that is potentially dangerous and is certainly inequitable (in intergenerational terms). In the same way, spatial discount rates above the natural rate of diffusion imply a parochial approach to the management of environmental resources that is equally inequitable (in intragenerational terms). Nevertheless, high spatial discount rates may be warranted by high rates of diffusion (or decay) of environmental effects. Where the environmental consequences of emissions are localized and decay quickly, high discount rates may still be ethically neutral in the sense that they appropriately weight the damage to distant members of the present generation resulting from local decisions.

See Also the Following Articles

Land Use Mapping • Location Analysis • Mathematical Demography • Urban Studies

Further Reading

Colwell, P., and Sirmans, C. (1978). Area, time, centrality and the value of urban land. *Land Econ.* **54,** 514–519.

Colwell, P., and Guntermann, K. (1984). The value of neighborhood schools. *Econ. Ed. Rev.* **3,** 177–182.

Colwell, P., Gujral, S., and Coley, C. (1985). The impact of a shopping center on the value of surrounding properties. *Real Estate Issues* Spring–Summer, 35–39.

Farber, S. (1998). Undesirable facilities and property values: A summary of empirical studies. *Ecol. Econ.* **24,** 1–14.

Hallman, W., and Wandersman, A. (1989). Perception of risk and toxic hazards. In *Psychosocial Effects of Hazardous Toxic Waste Disposal on Communities* (D. Peck, ed.), pp. 31–56. C. Thomas, Springfield, IL.

Hannon, B. (1987). The discounting of concern. In *Environmental Economics* (Pillet and Mirowata, eds.), pp. 227–247. R. Leimgruber Press, Geneva.

Hannon, B. (1994). Sense of place: Geographic discounting by people, animals and plants. *Ecol. Econ.* **10,** 157–174.

Howe, H. (1988). A comparison of actual and perceived residential proximity to toxic waste sites. *Arch. Environ. Health* **43,** 415.

Laundré, J., and Keller, B. (1984). Home-range size of coyotes: A critical review. *J. Wildl. Manage.* **48,** 127.

Maderthaner, R., Guttmann, G., Swaton, E., and Otway, H. (1978). Effect of distance on risk perception. *J. Appl. Psychol.* **63,** 380.

Mitchell, R., and Carson, R. (1986). *Property rights, protest, and the siting of hazardous waste facilities.* Report No. 230. Resources for the Future, Washington, DC.

Norton, B., and Hannon, B. (1997). Environmental values: A place-based approach. *Envir. Ethics* **19,** 227–246.

Norton, B., and Hannon, B. (1998). Democracy and sense of place values in environmental policy. In *Philosophy and Geography, Vol. 3: Philosophies of Place* (A. Light and J. Smith, eds.), pp. 119–146. Rowman and Littlefield, Lanham, MD.

Pate, J., and Loomis, J. (1997). The effect of distance on willingness to pay values: A case study of salmon in California. *Ecol. Econ.* **20,** 199–208.

Perrings, C., and Hannon, B. (2001). Spatial discounting: endogenous preferences and the valuation of geographically distributed environmental externalities. *J. Reg. Sci.* **41,** 23–38.

Steininger, K. (1997). Spatial discounting and the environment: An empirical investigation into human preferences. Paper presented at the Annual Meeting of the European Association of Environmental and Resource Economists, Tilburg.

Tuan, Y. (1971). Man and nature. Resource Paper 10. *Commission on College Geography.* Association of American Geographers, Washington, D.C.

U.S. Department of Commerce. (1984). Franchising in the economy, 1982–1984. Washington, D.C.

Spatial Econometrics

James P. LeSage
University of Toledo, Toledo, Ohio, USA

Glossary

generalized moments estimation A method of estimation that determines parameter estimates by matching moments of the observed sample data to the data-generating process.
limited dependent variable models Models involving a dependent variable that takes on discrete values rather than the typical continuous values.
Markov Chain Monte Carlo estimation A method of estimation that samples sequentially from the complete sequence of conditional distributions for all parameters in the model.
maximum likelihood estimation A method for determining parameter estimates that maximize the likelihood of the observed sample data.
non-parametric locally linear models Models that construct estimates of complicated phenomena from a series of many linear models fitted to sub-samples of observations.
spatial dependence When observations collected from points or regions in space at one location are a function of the value of observations from nearby locations.
tobit models Models in which the dependent variable is subject to censoring for values greater than or less than a particular magnitude.

Spatial econometrics provides models for situations in which sample data observations are taken with reference to points or regions on a map. These sample data often exhibit spatial dependence, ruling out use of conventional econometric estimation methods.

Regression Analysis with Spatial Data

Spatial data samples involve observations collected with reference to points or regions in space, for example, a sample of observations based on all counties in a state, or a sample of census tracts in a particular city. Each observation reflects values of the observed variable associated with a particular geographical area such as a country, state, county, or census division. Another example is a sample of sales prices for real estate parcels, where the location (perhaps in the form of map coordinates such as latitude–longitude coordinates) of each parcel is known. Cross-sectional observations collected with reference to spatial locations often exhibit spatial dependence, violating the independence assumption of traditional least-squares regression.

Spatial Dependence

Spatial dependence in a collection of sample data implies that observations at location i depend on other observations at locations $j \neq i$. This can be stated formally as

$$y_i = f(y_j), \quad i = 1, \ldots, n \quad j \neq i. \tag{1}$$

Note that the dependence can be among several observations, as the index i can take on any value from 1 to n.

Spatial dependence can arise from theoretical as well as statistical considerations. From a theoretical viewpoint, for example, maintenance of a neighbor's house may directly impact the value of one's property. Local governments might engage in competition that leads to local uniformity in taxes and services. Pollution can create systematic patterns over space; clusters of consumers who travel to a more distant store to avoid a nearby store located in a high-crime zone would also generate these patterns. In addition, spatial dependence can arise from unobservable latent variables that are spatially correlated. Consumer expenditures collected at spatial locations such as census tracts exhibit spatial dependence, as do other variables such as housing prices. It seems plausible that difficult-to-quantify or unobservable characteristics

such as the quality of life may also exhibit spatial dependence.

Estimation Consequences of Spatial Dependence

In some applications, the spatial structure of the dependence may be a subject of interest or provide a key insight. In other cases, it may be a nuisance similar to serial correlation in time-series regression. In either case, inappropriate treatment of sample data with spatial dependence can lead to inefficient and/or biased and inconsistent estimates.

For models of the type $y_i = f(y_j) + X_i\beta + \varepsilon_i$, least-squares estimates for β are biased and inconsistent, similar to the simultaneity problem in econometrics. The right-hand-side function of variables y_j (where j indexes spatially dependent observations) cannot be treated as fixed in repeated sampling. This leads to a situation similar to that encountered with relationships involving simultaneously determined variables.

For a model of the type $y_i = X_i\beta + u_i$, $u_i = f(u_j) + \varepsilon_i$, least-squares estimates for β are inefficient but consistent, similar to the serial correlation problem. As in serial correlation, variance-covariance estimates for the vector of parameter β constructed using least-squares formulas are biased.

A Family of Spatial Econometric Models

A family of regression-based models that incorporate spatial dependence was introduced by Ord in 1975 and popularized by Anselin in 1988. Two members of this model family are most frequently employed in applied practice, the spatial autoregressive model (SAR) shown in Eq. (2) and the spatial error model (SEM) in Eq. (3). The model in Eq. (2) is sometimes referred to as a "mixed regressive spatial autoregressive model" to distinguish it from the model $y = \rho Wy + \varepsilon$, which is labeled a spatial autoregressive model. This simpler model is referred to as a first-order spatial autoregressive (FAR) model in this article, so the less awkward acronym SAR can be used for the model in Eq. (2).

$$y = \rho Wy + X\beta + \varepsilon \tag{2}$$

$$y = X\beta + u, \quad u = \lambda W + \varepsilon. \tag{3}$$

In Eqs. (2) and (3), y represents an $n \times 1$ vector of cross-sectional observations associated with points in space, X represents an $n \times k$ matrix of explanatory variables, ε is an $n \times 1$ vector of normally distributed disturbances, u is a vector of disturbances that follows a spatial autoregressive process, and W denotes an $n \times n$ spatial weight matrix, which is discussed later. The parameters to be estimated in the SAR and SEM models are the usual vector β and noise variance σ_ε^2, along with the scalar parameters ρ in the case of the SAR model and λ in the SEM model. The SAR and SEM models subsume least-squares as a special case that arises when $\rho = 0$ in the SAR model and $\lambda = 0$ in the SEM model.

The spatial weight matrix defines the connectivity structure between spatial observations, and various approaches to specifying W have appeared in the literature. Most specifications set $w_{ij} > 0$ for observations $j = 1, \ldots, n$ sufficiently close (as measured by some metric) to observation i, where w_{ij} denotes the (i, j)th element of the matrix W. Two of the more popular definitions used to construct W are first-order contiguity relationships and nearest neighbors. A matrix W based on first-order contiguity relationships would set $w_{ij} = 1$ for observations j that have borders touching region i, and $w_{ij} = 0$ for observations j that do not border on i. The nearest neighbors approach to specifying W would involve finding the m nearest neighbors to each observation i and using these to set $w_{ij} = 1, j = 1, \ldots, m$.

Two other points concerning the matrix W are that diagonal elements $w_{ij} = 0$, $i = j$, and the matrix W is usually row-standardized to have row-sums of unity. This is motivated by the use of the matrix W in the SAR and SEM models to create $n \times 1$ variable vectors called "spatial lags" formed by the products W_y and W_u. Given row-standardization of W, these vectors represent an average of spatially neighboring values. For example, if observations 1 and 3 represent the only regions with borders touching region 2, the second row of the SAR model will take the form

$$y_2 = \rho(0.5y_1 + 0.5y_3) + \sum_{j=1}^{k} x_{2j}\beta_j + \varepsilon_2 \tag{4}$$

This should make it clear why $w_{ii} = 0$, as this precludes an observation y_i from directly predicting itself. It also motivates row-standardization of W, which makes each observation y_i a function of the spatial lag Wy, an explanatory variable representing an average of spatially neighboring values to each observation i. Similarly, the spatial autoregressive disturbance process in the SEM model allows dependence between the disturbance from observation i and an average of disturbances from nearby observations specified by the spatial lag Wu.

A relatively simple extension of the SAR model called the spatial Durbin model (SDM) adds spatial lags of the explanatory variables in the matrix X, created by WX^* as

shown in Eq. (5), where the matrix X^* equals X with the constant term excluded.

$$y = \rho Wy + X\beta + WX*\gamma + \varepsilon. \quad (5)$$

The $(k-1)x_1$ parameter vector γ measures the marginal impact of the explanatory variables from neighboring observations on the dependent variable y.

Maximum Likelihood Estimation

Maximum likelihood estimation of the SAR, SDM, and SEM models involves maximizing a concentrated log likelihood function with respect to the parameter ρ or λ.

For the case of the SAR model

$$\begin{aligned} \ln L &= C + \ln |I_n - \rho W| - (n/2)\ln(e'e) \\ e &= e_o - \rho e_d \\ e_o &= y - X\beta_o \\ e_d &= W_y - X\beta_d \\ \beta_o &= (X'X)^{-1}X'y \\ \beta_d &= (X'X)^{-1}X'Wy, \end{aligned} \quad (6)$$

where C represents a constant not involving the parameters. The computationally troublesome aspect of this is the need to compute the log-determinant of the $n \times n$ matrix $(I_n - \rho W)$. The operation counts for computing this determinant grow with the cube of n for dense matrices. This same approach can be applied to the SDM model by simply defining $X = [X\ WX^*]$ in Eq. (6). A concentrated log-likelihood can also be devised for the SEM model.

Computational Issues

One of the earlier computationally efficient approaches to solving for estimates in a model involving a sample of 3107 observations was proposed by Pace and Barry. They suggested using direct sparse matrix algorithms such as the Cholesky or LU decompositions to compute the log-determinant. A sparse matrix is one that contains a large proportion of zeros. As a concrete example, consider the spatial weight matrix for the sample of 3107 U.S. counties used by Pace and Barry. This matrix is sparse because the largest number of neighbors to any county is 8 and the average number of neighbors is 4. To understand how sparse matrix algorithms conserve on storage space and computer memory, consider that only the non-zero elements need be recorded along with an indication of their row and column position. This requires a 1×3 vector for each non-zero element consisting of a row index, a column index, and the element value. Because non-zero elements represent a small fraction of the total $3107 \times 3107 = 9{,}653{,}449$ elements in the weight matrix, computer memory is saved. For the example of the 3107 U.S. counties, only 12,429 non-zero elements were found in the weight matrix, representing a very small amount (about 0.4%) of the total elements. Storing the matrix in sparse form requires only $3 \times 12{,}429$ elements, or more that 250 times less computer memory than would be needed to store 9,653,449 elements.

In addition to storage savings, sparse matrices result in lower operation counts, which speeds computations. In the case of non-sparse (dense) matrices, matrix multiplication and common matrix decompositions such as the Cholesky require $O(n^3)$ operations, whereas for sparse W these operation counts can fall as low as $O(n_{\neq 0})$, where $n_{\neq 0}$ denotes the number of non-zero elements.

In addition to proposing the use of sparse matrix algorithms, Pace and Barry proposed a vector evaluation of the SAR or SDM log-likelihood functions over a grid of q values of ρ to find maximum likelihood estimates. It is well known that for row-stochastic W, $\lambda_{\min} < 0$, $\lambda_{\max} > 0$, where $\lambda_{\min}$, $\lambda_{\max}$ represent the minimum and maximum eigenvalues of the spatial weight matrix. In this case, ρ must lie in the interval $[\lambda_{\min}^{-1}, \lambda_{\max}^{-1}]$, but typical applied work simply relies on a restriction of ρ to the $(-1, 1)$ or $[0, 1)$ interval to avoid the need to compute eigenvalues. The computationally intense part of this approach still is calculating the log-determinant, which takes approximately 201 s for a sample of 57,647 U.S. census tracts. This is based on a grid of 100 values from $\rho = 0$ to 1 using sparse matrix algorithms in MATLAB Version 6.0 on a 600 Mhz Pentium III computer. The SEM model log-likelihood is not as amenable to the vectorized form, but can be solved using more conventional optimization, such as a simplex algorithm. Nonetheless, a grid of values for the log-determinant over the feasible range for λ can be used to speed evaluation of the log-likelihood function during optimization with respect to λ.

Recent Advances in Computation

An improvement based on a Monte Carlo estimator for the log determinant suggested by Barry and Pace allows larger problems to be tackled without the memory requirements or sensitivity to orderings associated with the direct sparse matrix approach. Their method not only provides an approximation to the log-determinant term but also produces an asymptotic confidence interval around the approximation. As an illustration of these computational advantages, the time required to compute a grid of log-determinant values for $\rho = 0, \ldots, 1$, based on 0.001 increments for the sample of 57,647 census tracts, was 3.6 s, which compares quite favorably to 201 s for the direct sparse matrix computations cited earlier. LeSage and Pace reported experimental results indicating that the Monte Carlo log-determinant estimator produces nearly the same estimates as the direct method.

Estimates of Dispersion

An implementation issue is constructing estimates of dispersion for the maximum likelihood parameter estimates for the purpose of inference. In problems involving a small number of observations, an asymptotic variance matrix based on the Fisher information matrix shown below for the parameters $\theta = (\rho, \beta, \sigma^2)$ can be used to provide measures of dispersion for the estimates of ρ, β, and σ^2. Anselin provided the analytical expressions needed to construct this information matrix.

$$[I(\theta)]^{-1} = -E\left[\frac{\partial^2 L}{\partial\theta\partial\theta'}\right]^{-1} \tag{7}$$

This approach is computationally impossible when dealing with large scale problems involving thousands of observations. The expressions used to calculate terms in the information matrix involve operations on very large matrices that would take a great deal of computer memory and computing time. In these cases, the numerical Hessian matrix can be evaluated using the maximum likelihood estimates of ρ, β, and σ^2 and the sparse matrix representation of the likelihood. Given the ability to evaluate the likelihood function rapidly, numerical methods can be used to compute approximations to the gradients shown in Eq. (7).

Alternatives to Maximum Likelihood Estimation

A number of alternative approaches to dealing with spatial samples that exhibit spatial dependence have been proposed in the literature.

Generalized Moments Estimation

Kelejian and Prucha suggested a generalized moments estimation approach for the family of spatial models that they labeled generalized two-stage least squares (G2SLS). This is purported to be computationally simpler than maximum likelihood estimation but is unlikely to be efficient relative to maximum likelihood. Note that the asymptotic distribution of the G2SLS estimator is not established, but Bell and Bockstael provided Monte Carlo evidence that these estimates are only slightly inefficient relative to maximum likelihood estimates. The G2SLS procedure involves three steps. The first step generates residuals using 2SLS with an $n \times m$ instrumental variables (IV) matrix $Q = (X, WX, W^2X, \ldots)$ formed using functions of X and W. Residuals based on these estimates are used to construct a consistent generalized moments estimate of ρ or λ in the second step. The third step involves generalized 2SLS estimation on a set of quasi-differenced variables constructed using the consistent estimates of ρ or λ, from step two. It should be noted that the Kelejian-Prucha GMM approach does not impose a restriction on the parameter ρ or λ, so that estimates having values greater than unity may arise in practice.

Bayesian Methods

LeSage proposed the use of Bayesian estimation methods for the SAR, SDM, and SEM models based on Markov Chain Monte Carlo (MCMC) estimation methods described in Gelfand and Smith. This approach allows an extended version of the SAR, SDM, and SEM models that introduces nonconstant variance scalars as parameters to accommodate spatial heterogeneity and outliers that often arise in applied practice. Maximum likelihood estimation methods rely on the assumption $\varepsilon \sim N(0, \sigma^2)$, where the noise variance is assumed constant over space. The Bayesian method introduces a set of variance scalars $(v_1, v_2, \ldots, v_n)$, as unknown parameters that need to be estimated. This allows the constant variance assumption to be replaced with $\varepsilon \sim N(0, \sigma^2 V)$, where $V = \text{diag}(v_1, v_2, \ldots, v_n)$.

Application of Bayesian estimation methods to SAR, SDM, and SEM spatial regression models should result in estimates nearly identical to those from maximum likelihood methods when the number of observations is large. This is a typical result when prior information is dominated by a large amount of sample information. However, the heteroscedastic Bayesian variants of the SAR and SDM models represent a situation in which prior information regarding the variance scalars exerts an impact even in very large samples. The sparse matrix approaches to computing the log-determinant term discussed in the context of maximum likelihood estimation can be applied to MCMC estimation of the Bayesian variant of these models, making them relatively fast.

Nonparametric Locally Linear Models

One branch of spatial econometrics uses distance-weighted subsamples of the data in conjunction with ordinary least-squares to produce parameter estimates at various points in space. McMillen introduced this form of nonparametric locally linear weighted regression (LWR), which Brunsdon, Fotheringham, and Charlton termed geographically weighted regressions (GWR). By estimating separate models using data near each observation, these intuitively appealing methods attempt to overcome the problem of spatial heterogeneity. If spatial dependence arises due to inadequately modeled spatial heterogeneity, LWR can potentially eliminate this problem. These models often rely on the estimated parameters to detect systematic patterns over space.

Using the previously introduced notation and letting $W(i)$ represent an $n \times n$ diagonal matrix containing distance-based weights for observation i that reflect the distance between observation i and all other observations, the LWR model can be written as

$$W(i)^{1/2}y = W(i)^{1/2}X\beta_i + W(i)^{1/2}\varepsilon_i. \tag{8}$$

The subscript i on β_i indicates that this $k \times 1$ parameter vector is associated with observation i. The LWR model produces n such vectors of parameter estimates, one for each observation. These estimates are calculated using

$$\hat{\beta}_i = [X'W(i)X]^{-1}[X'W(i)y]. \tag{9}$$

A number of alternative approaches have been proposed for constructing the distance-based weights for each observation i contained in the vector on the diagonal of $W(i)$. As an example, McMillen suggested a tri-cube weighting function,

$$\text{diag}[W(i)] = \left[1 - \left(\frac{d_i^j}{d_i^m}\right)^3\right]^3 \quad \text{I}\left(d_i^j < d_i^m\right), \tag{10}$$

where d_i^j represents the distance between observation j and observation i, d_i^m represents the distance between the mth nearest neighbor and observation i, and I( ) is an indicator function that equals one when the condition is true, and zero otherwise. In practice, the number of nearest neighbors used (often referred to as the bandwidth) is determined with a cross-validation procedure, typically a prediction criterion based on excluding a single observation for each i.

LeSage pointed out that aberrant observations or outliers arising from spatial enclave effects or shifts in regime can exert a large impact on the locally linear estimates. Because LWR estimates are based on a small number of observations, and the sample data observations are re-used when estimates are produced for each point in space, a single outlier can contaminate estimates covering large regions of the spatial sample. LeSage proposed a Bayesian variant of the locally linear models that overcomes sensitivity to outliers by introducing explicit stochastic spatial parameter transition relationships as prior information in the model. Because prior uncertainty regarding parameter variability is under the control of the user, a continuum of estimates ranging from highly volatile to relatively constant over space can be produced.

Pace and LeSage argued that traditional LWR methods exhibit a trade-off: increasing the sample size produces less volatile estimates that contain increasing spatial dependence. Selecting a smaller sample size reduces the spatial dependence, but at the cost of increased parameter variability that impedes detection of systematic patterns of parameter variation over space. They introduced a spatial autoregressive locally linear estimation method that extends the LWR approach to include a spatial lag of the dependent variable, which they labeled spatial autoregressive local estimation (SALE). They accomplished this using a recursive approach for maximum likelihood estimation of spatial autoregressive models. This allows consideration of a series of estimates based on sub-samples of varying size in the spirit of the LWR methods. They argued that inclusion of the spatial autoregressive term in the model results in improves prediction and stability of the parameter estimates, decreasing the sensitivity of performance to the bandwidth that is typically observed.

Matrix Exponential Spatial Models

Pace and LeSage introduced a spatial model specification based on the matrix exponential. Use of the matrix exponential spatial specification (MESS) eliminates the log-determinant term from the log-likelihood function, and a closed-form solution exists for this model. The MESS is shown in Eq. (11), where W represents a spatial weight matrix, and the scalar parameter α plays the role of the spatial dependence parameter ρ in the SAR model.

$$\begin{aligned} Sy &= X\beta + \varepsilon \\ S &= e^{\alpha W} = \sum_{i=0}^{\infty} \frac{\alpha^i W^i}{i!}. \end{aligned} \tag{11}$$

If $W_{ij} > 0$ for the nearest neighbors of observation i, $W_{ij}^2 > 0$ contains neighbors to these nearest neighbors for observation i. Similar relations hold for higher powers of W, which identify higher order neighbors. Thus, the matrix exponential S, associated with matrix W, can be interpreted as assigning rapidly declining weights for observations involving higher order neighboring relationships. That is, observations reflecting higher order neighbors (neighbors of neighbors) receive less weight than lower order neighbors.

Maximizing the log-likelihood is equivalent to minimizing $y'S'MSy$ with respect to S, where $M = I - H$ and $H = X(X'X)^{-1}X'$ are idempotent matrices. This is essentially a closed-form problem in that it involves solving a constrained non-linear (polynomial) least-squares problem.

LeSage and Pace provided a Bayesian variant of this approach, estimated using MCMC methods. This approach involves a more flexible spatial weight matrix specification that allows posterior inferences regarding the magnitude and extent of spatial influence. In an economic context, where the spatial structure can arise from externalities or spillovers, the magnitude and extent of influence from one observation or spatial location on other observations at nearby locations may be a subject of interest.

The flexible specification for the spatial weights that allows hyperparameters to control the number of

neighboring entities as well as decay of influence over space is shown in Eq. (12), where m denotes the maximum number of neighbors considered.

$$W = \sum_{i=1}^{m} \left(\frac{\gamma^i N_i}{\sum_{i=1}^{m} \gamma^i} \right) \tag{12}$$

In Eq. (12), γ^i weights the relative effect of the ith individual neighbor matrix N_i, so that S depends on the parameters γ as well as m in both its construction and the metric used. By construction, each row in N sums to 1 and has zeros on the diagonal. To see the role of the spatial decay hyperparameter γ, consider that a value of $\gamma = 0.87$ implies a decay profile where the 6th nearest neighbor exerts less than half the influence of the nearest neighbor. This value of γ can be thought of as having a "half-life" of six neighbors. On the other hand, a value of $\gamma = 0.95$ implies a half-life between 14 and 15 neighbors.

Censored and Limited Dependent Variables

As in the case of non-spatial data samples, spatial data often involve binary or censored dependent variables. Binary data might arise when the sample indicates the presence or absence of some phenomena at various points in space. Censored samples often arise from census reporting methods. As in the case of continuous dependent variables, the presence of spatial dependence requires special estimation approaches for modelling these problems.

Autologistic Model

Besag introduced the autologistic estimator for binary dependent variable data problems exhibiting spatial dependence. In this case, the log-likelihood no longer has a closed form solution, but Besag proposed an approximate or pseudo-likelihood technique. He showed that one can use the typical logistic regression estimation algorithms with an additional explanatory variable constructed from the spatial average of the binary dependent variable, i.e., W_y. The pseudo-likelihood estimates that arise from proceeding in this fashion are consistent and asymptotically normal.

Spatial Tobit Models

These models accommodate situations in which the dependent variable y can be partitioned into one group of observations that are censored and a second set of observations that are continuous. This situation might arise because government agencies suppress information for confidentiality reasons. For example, the census might report housing values greater than a particular level, say C dollars using the value C. These observations are said to be censored. A partitioning of the data can be used, where y_u stands for the uncensored, or continuous, sample data and y_c, denotes observations that are subject to censoring.

For the case of independent data as in ordinary tobit regression, Chib introduced a latent variable z for the censored observations. He showed that the conditional distribution of z simplifies into a product of independent distributions, taking the form of a truncated univariate normal distribution. For the case of spatial dependence in y, a sequence of univariate conditional distributions is arrived at that embody the multivariate spatial dependence between observations. The censored observations can be sampled using a sequence of univariate conditional truncated normals arising from the multivariate normal distribution, taking the form

$$p(z_i \mid i \in y_c, \beta, \sigma^2, \rho) \sim TN_{(-\infty, C]}\{E(y_i \mid y_j, \forall i \neq j), \mathrm{var}(y_i \mid y_j)\}. \tag{13}$$

Spatial Probit Models

The difference between the expression for the likelihood function of the SAR or SEM model involving a continuous dependent variable and the likelihood for a binary probit model in the presence of a spatial dependence covariance structure is tremendous. McMillen noted that for the family of spatial models, the likelihood involves an n-dimensional integral, where n is the number of observations.

In contrast, for the Bayesian SAR/SEM models, a vector z of latent continuous dependent variables can be introduced, as in the case of the tobit model from the previous section. With binary dependent variables, the latent vector can be interpreted formally as utility differences associated with individuals making decisions at various points in space, or more generally as the expected value of the explanatory variables in a standard linear model. In either case, this continuous vector becomes part of an MCMC sampling scheme, so that depending on these continuous values, all remaining parameters of the model can be sampled, as in the case of a continuous dependent variable model. In the case of a spatially dependent covariance structure, this approach requires modification similar to that described for the case of spatial tobit models in the previous section.

McMillen made the point that heteroskedasticity in spatial probit models will lead to inconsistent estimates if ignored during estimation. This makes the Bayesian heteroskedastic variant of the SAR and SEM models quite useful here. The use of the variance scalars (v_1, $v_2, \ldots, v_n$) with the sequence of independent, identically distributed $\chi^2(r)$ priors produces a model that

is equivalent to one based on an assumed student t-distribution for the disturbances of the model. Albert and Chib pointed out that in the case of the probit regression model, use of these variance scalars can be viewed as a probability rule based on a family of t-distributions that represent a mixture of the underlying normal distribution used in the probit regression, since the normal distribution can be modeled as a mixture of t-distributions.

The most popular choice of probability rule to relate fitted probabilities with binary data is the logit function corresponding to a logistic distribution for the cumulative density function. The quantiles of the logistic distribution correspond to a t-distribution with 7 or 8 degrees of freedom, and the normal probability density is similar to a t-distribution when the degrees of freedom are large. Setting the prior hyperparameter for the $\chi^2(r)$ prior to a value of $r=7$ or $r=8$ will approximate a spatial logit model, whereas setting this parameter to a large value such as $r=50$ approximates a probit model.

An alternative approach to spatial probit was set forth by Beron and Vijverberg, who used maximum likelihood methods to estimate a spatial probit model. As already noted, maximum likelihood requires evaluating an n-dimensional normal probability integral, similar to the case that arises in non-spatial multinomial probit (MNP). The n-dimensional normal probability integration is implemented using a method known as GHK simulation.

Summary

Econometric methods exist for regression modeling of spatial data samples that exhibit spatial dependence. These estimation methods model spatial dependence using spatial weight matrices to construct spatial lags of the dependent variable or the disturbances. Methods have been devised for continuous and dichotomous as well as censored dependent variables. Maximum likelihood, Bayesian, and method of moments estimators are available, as well as nonparametric locally linear approaches.

Areas for future work are simultaneous equation systems and the case of multinomial probit estimation for relationships involving spatial dependence. Also, estimation methods that could be applied to space-time data samples would be of use.

See Also the Following Articles

Bayesian Statistics • Maximum Likelihood Estimation • Spatial Databases

Further Reading

Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assn.* **88,** 669–679.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic Publishers, Dordrecht.

Barry, R., and Pace, R. K. (1999). A Monte Carlo estimator of the log determinant of large sparse matrices. *Lin. Algeb. Appl.* **289,** 41–54.

Bell, K. P., and Bockstael, N. E. (2000). Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *Rev. Econ. Statist.* **87,** 72–82.

Beron, K. J., and Vijverberg, W. P. M. (2004). Probit in spatial context: A Monte Carlo analysis. In *Advances in Spatial Econometrics. Methodology, Tools and Applications.* (L. Anselin, R. J. G. M. Florax, and S. J. Rey, eds.), Springer Verlag, Berlin.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B* **36,** 192–225.

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial non-stationarity. *Geograph. Anal.* **28,** 281–298.

Chib, S. (1992). Bayes inference in the tobit censored regression model. *J. Econ.* **51,** 79–99.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Rev. Ed. John Wiley, New York.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assn.* **85,** 398–409.

Kelejian, H., and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *J. Real Estate Fin. Econ.* **17,** 99–121.

LeSage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *Int. Reg. Sci. Rev.* **20,** 113–129.

LeSage, J. P. (2004). A family of geographically weighted regression models. In *Advances in Spatial Econometrics. Methodology, Tools and Applications.* (L. Anselin, R. J. G. M. Florax, and S. J. Rey, eds.), Springer Verlag, Berlin.

LeSage, J. P., and Pace, R. K. (2001). Spatial dependence in data mining. In *Data Mining for Scientific and Engineering Applications,* (R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. R. Namburu, eds.), pp. 439–460. Kluwer Academic Publishing.

LeSage, J. P., and Pace, R. K. (2004). Models for spatially dependent missing data. *Real Estate Fin. Econ.* **29,** 233–254.

McMillen, D. P. (1992). Probit with spatial autocorrelation. *J. Reg. Sci.* **32,** 335–348.

McMillen, D. P. (1996). One hundred fifty years of land values in Chicago: A nonparameteric approach. *J. Urban Econ.* **40,** 100–124.

Ord, J. K. (1975). Estimation methods for models of spatial interaction. *J. Am. Stat. Assn.* **70,** 120–126.

Pace, R. K., and Barry, R. (1997). Quick computation of spatial autoregressive estimators. *Geograph. Anal.* **29,** 232–246.

Pace, R. K., and LeSage, J. P. (2004). Spatial autoregressive local estimation. In *Recent Advances in Spatial Econometrics* (Jesus Mur, Henri Zoller, and Arthur Getis, eds.), pp. 31–51. Palgrave Publishers, Hampshire, UK.

Smith, T. E., and LeSage, J. P. (2001). A bayesian probit model with spatial dependencies. http://www.spatialeconometrics.com

Spatial Externalities

Dean M. Hanink
University of Connecticut, Storrs, Connecticut, USA

Glossary

distancing The imposition of a negative externality at some distance from the source of its generating action.
externality The difference between the marginal social cost of an action and its marginal private cost.
geographical externality An externality that exists without variation across an entire geographical unit, such as a country or metropolitan area.
increasing returns A rate of increase in output that exceeds the rate of increase in tangible inputs.
net externality The difference between the negative and positive externalities imposed by a single source.
Pigouvian tax A tax that exacts the cost of a negative externality from its source agency.
spatial externality An externality that varies continuously with distance from its source.

Externalities exist when the marginal social cost of an action is different than its marginal private cost. Spatial externalities are differences in marginal private costs and marginal social costs that co-vary with distance from the place where the action occurs. A negative externality, or marginal social cost, occurs when marginal total cost is greater than marginal private cost. The externality is that part of an action's total marginal cost that is imposed on society. Negative spatial externalities exist when the difference between a higher marginal total cost and a lower marginal private cost changes with distance from the source of the action. A positive externality exists when social costs are negative, meaning that a social benefit is provided completely at private cost. Positive spatial externalities exist when the difference between lower marginal total cost and higher marginal private cost changes with distance from the source of the action.

Externality Effects

Externalities can be positive or negative, and they can result from production or consumption. Negative production externalities include, for example, costs of pollution abatement that are borne by the public at large rather than individual producers. Positive production externalities may result from an individual company's labor training that spills over to an industry at large. Roadside littering is an example of a negative consumption externality, while becoming educated (consuming education) often yields positive externalities in the broadest social sense. In their effects, externalities can often be compared to subsidies when they are positive—they increase consumption and production beyond expected levels. Negative externalities, on the other hand, can be treated like taxes or transaction costs that can decrease consumption and production below otherwise expected levels.

Externalities cause changes in behavior because they either reduce the cost of an action (positive externalities) or increase it (negative externalities). Such a change in behavior is illustrated in Fig. 1, which illustrates an externality's impact on consumption of a hypothetical product from the supply side. In that figure, S represents the typical upward sloping supply function of conventional economic analysis. Assuming a competitive market, S represents price of output as well. The line D represents the typical downward sloping demand function. In a typical example of this sort, S should represent marginal production costs in full, but that is not the case in Fig. 1. In this case, E represents production costs that are externalized by producers, which for purposes of illustration are indicated as marginally increasing with levels of production. If the external costs were internalized by producers, then the relevant supply function would be traced by S'—the total of S and E. The effect of externalized cost on consumption is readily apparent. When S traces costs, then

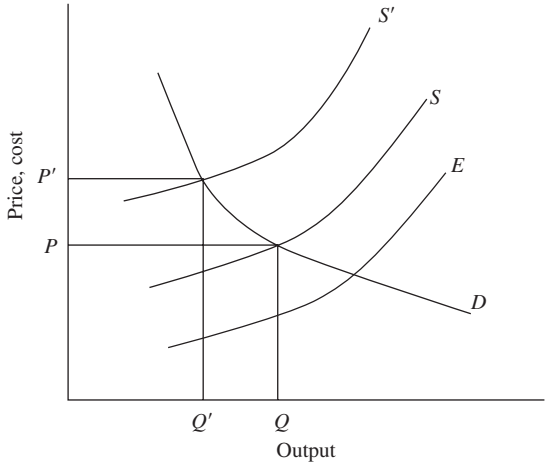


Figure 1 Externalities in production.

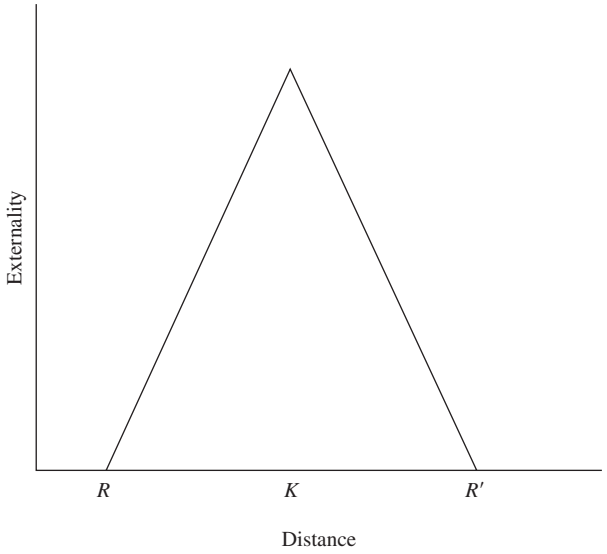


Figure 2 Spatial externality.

the equilibrium price is P with supply and demand at Q. If, however, external costs are internalized, there is a supply shift to the left (S'), and the equilibrium price rises to P' and supply and demand equilibrium shrinks to Q'.

Is the externality described in Fig. 1 positive or negative? In this case, the quality of the externality is relative. From the viewpoint of the producers, the external cost (E) is positive, because their internal marginal cost of production is lower than the full marginal cost of production. The externality is also indirectly positive to consumers, who are able to consume more of the product. On the other hand, E is negative from a social perspective because it represents costs that must be paid by parties beyond the set of producers and consumers represented in the stylized example. Again using Fig. 1, say that the externality arises due to consumption rather than as part of the production process. That makes the externality directly beneficial to consumers because they can consume without paying the full cost themselves, and indirectly beneficial to producers, who are able to sell more output. If the externality is internalized to consumers, it is effectively added to the cost of the product, and again there is a shift from S to S', a price increase from P to P', and a decrease in consumption from Q to Q'.

Spatial and Geographical Externalities

Externalities can have a temporal characteristic. For example, the motive for conservation of natural resources is often the reduction of externalities of current consumption being imposed on future generations. Such externalities could take the form of species extinction or exhaustion of a particular natural resource. Externalities also often have areal extents, meaning their incidence can be mapped. Typically, externalities that are mappable are broadly described as "spatial," but it is worthwhile to classify mappable externalities into two groups: spatial externalities *per se* and geographical externalities.

Spatial externalities often vary continuously with distance and/or direction from their sources. That means that a spatial externality effect often can be modeled as a diffusion process or as an autocorrelation function, either of which typically incorporates distance (d) decay (for example, d^{-k} or $\exp^{-d}$) as a characteristic. A simple spatial externality is graphed in Fig. 2. Its source is at K, and it decreases with distance until it vanishes at R and R'; the externality's spatial extent, or range, is defined as KR. Extended to a plane, KR would be the radius of a circle that defines the externality's field.

The illustration in Fig. 2 represents only a simple case. It does not indicate any directional bias to the externality field. Such a directional bias would lead to a non-circular externality field, for example, in the case of pollution that is carried by either prevailing water or air currents. Further, the continuity of externality fields may be broken or diverted in direction by barriers.

Spatial externalities are fairly common. Almost any point-sourced pollution provides an example. Water effluents, air effluents, and noise usually dissipate along with their externality effects as distance from their origin increases. Alternatively, the positive externality of convenience associated with public transport stations decreases with their distance from residences.

Geographical externalities have effects that are spatially trivial in that they do not vary across an entire geographical unit such as a country or metropolitan area. Such an externality is illustrated in Fig. 3. While the

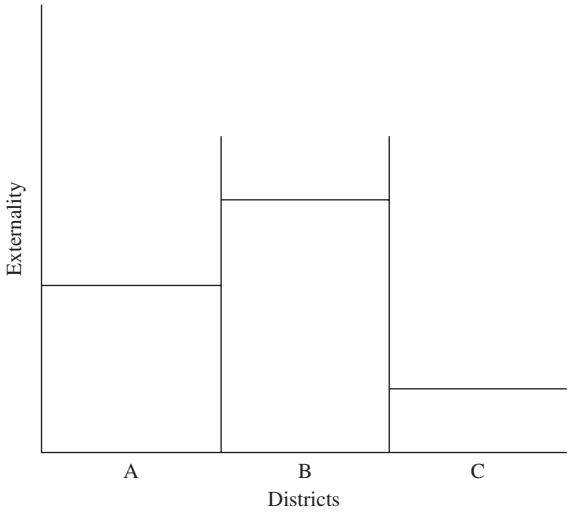


Figure 3 Geographical externality.

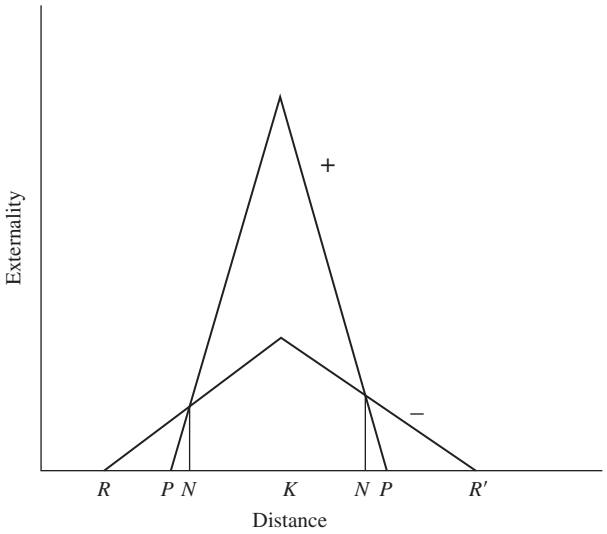


Figure 4 Net spatial externalities.

externality effect does vary in that figure from district to district, it is uniform within districts and does not vary with distance or direction from a particular source. Figure 3 could be describing the externality effects of variations in the quality of public education across school districts, for example, or different minimum regulation standards for water quality. Such geographical externalities may also have spatial properties. For example, the positive externality associated with higher air quality standards in one government jurisdiction can spill over into neighboring jurisdictions, with neighboring ones gaining more benefit than more distant ones.

A single place can be the origin of both negative and positive externalities. On a temporal basis, for example, a public park can provide positive externalities associated with recreation during the day but impose negative ones associated with criminal activity at night. On a spatial basis, airports may provide positive externalities associated with improved accessibility over an entire region but may considerably increase local costs associated with congestion and noise pollution.

Athletic stadiums are frequent sources of spatial externalities, which can be both positive and negative. For example, people who live near major college or professional football stadiums in the United States are affected periodically by the negative externalities of congestion before and after games. At the same time, some of those residents may be able to earn income from renting parking places on their own driveways and in their yards. Such income is a positive externality because it is obtained due to the action of the game. On days when games are not played, the parking places are worthless. Some local businesses may lose income on game days because regular customers may not want to incur traffic congestion costs, while other local businesses may experience a surge in income associated with the crowds attracted by the game.

Such a net spatial externality effect can be illustrated as in Fig. 4. An activity at *K* generates both positive and negative externalities. The positive externality field has range *KP*, while the negative externality field has range *KR*. Given that the maximum positive externality is greater in effect than the negative externality, but that the positive externality is subject to much greater distance decay than the negative externality, regions of net positive and net negative externality are defined by *KN* and *NR*, respectively.

Positive Externalities and Increasing Returns

Economies of scale are a decrease in the average cost of production with an increase in output. That is, increasing size leads to increasing efficiency, at least to a point. At an extreme, if economies of scale did not exist, then everyone would produce everything for her- or himself. Without economies of scale, there would be no need for firms. In a spatial sense, without externalities, there would be no need for cities; the landscape could consist of uniformly distributed households instead of the various levels of concentrated settlements that actually occur.

Economies of scale are often called increasing returns. Conventional economic analysis often characterizes production in a constant returns model that indicates that there are no economies of scale (and in relation, perfect competition). The Cobb-Douglas production function is a constant returns model that takes the following form:

$$Q = (K^{\beta}, L^{\alpha}), \qquad \alpha + \beta = 1, \tag{1}$$

where Q is output and K and L are capital and labor inputs, respectively. Increasing returns can be similarly modeled:

$$Q = (K^{\beta}, L^{\alpha}), \quad \alpha + \beta > 1, \tag{2}$$

so that output increases in observed, internal factors of production.

Externalities that contribute to output are often associated with concentrations of people and producers. For analytical purposes, whether the externalities are spatial or geographical is largely a matter of scale of interest (although that is not the case with respect to policy). Given that externality effects seem to vary with concentration, however, most of them that arise without policy inducement can be considered spatial in the sense that they will spill over from place of origin to neighboring places.

Concentrations of producers may generate so-called localization externalities that result, for example, from the ability to share a trained labor force. Costs of training are externalized, either to competitors or perhaps to public educational institutions. Other externality benefits arise from the availability of specialized suppliers (input–output relationships) that develop in industrial concentrations. Adam Smith noted that the division of labor "is limited by the extent of the market." The division of labor is efficiency-inducing specialization, which increases with the size of the localized market.

Other positive localization externalities arise from formal and informal producer networks. Formal networks with political goals are often strengthened when they represent geographical concentrations rather than dispersed ones, especially when it comes to guiding the development and implementation of local industrial regulations and policies. Other local networks facilitate knowledge spillovers so that production technology and managerial expertise are effectively reduced in cost for producers found within concentrations.

Agglomeration externalities are associated simply with concentrations of people as opposed to a specific set of producers. High levels of market potential, concentrations of general infrastructure, and a dynamic, knowledge-generating milieu have all been cited as sources of agglomeration-based efficiencies.

Linear regression analysis is often used to detect the externality effects in concentrations or agglomerations. For example, a cross-sectional production function could be estimated that takes the general operational form

$$\ln(Q_j) = \alpha + \beta \ln(K_j) + \gamma \ln(L_j) + \delta \ln(E_j) + \varepsilon, \tag{3}$$

where Q is output at the jth place, K and L are capital and labor quantities, and E is a measure of externality-inducing characteristics.

Negative Environmental Externalities

There are both positive and negative environmental externalities, but more attention has been given to the latter. Ironically, the concentrations of people or industries that yield positive externalities with respect to output may also yield negative externalities, such as those associated with traffic congestion. In fact, important negative environmental externalities often result from such concentrations. Air and water pollution, and the negative externalities they entail, can certainly arise from single sources, but certainly both are exacerbated by density of producers. Such negative externality effects of increased density or congestion are illustrated in the "tragedy of the commons." In that story, a village common sustainably supports pastoralists without constraint until a critical population is reached. Once that population is reached, however, each pastoralist's consumption of the common resource limits its supply for the others, and self-interest works against any individual effort at conservation.

So-called environmental distancing can reduce the problem of negative externalities by spatially separating them from their real source. Examples include international trade agreements that allow polluting components of a production process to be placed in one country while the benefits of that process are employed in another. Another example is the consumption of electricity from a nuclear power plant in one place, with storage of the power plant's waste in another. The effects of such distancing can be described in a cost–benefit framework, such as

$$NPV_i = \sum_{t=0}^{t} \frac{\left[B_{ti} - \left(C_t / d_{ij}\right)\right]}{(1+r)^t}, \tag{4}$$

in which the net present value at the ith place, NPV_i, of a project is the time discounted $(1+r)^t$ stream of benefits, B, minus the costs that accrue as a function of distance from the jth place, C_t/d_{ij}. Such distancing allows an alternative distribution of net spatial externality, as illustrated in Fig. 5. The maximum externality benefit of an action occurs at O; it is less but still positive at distance OK and negative at OM, and the negative externality is maximized at Z.

Environmental distancing often raises questions of environmental racism and environmental justice. In the United States, for example, waste disposal sites are often located in low-income or minority neighborhoods, far removed from the wealthier and whiter places in which the waste is generated. Negative externalities, including health effects, are therefore imposed on marginalized populations in their locations, while the positive externality of consumption below its social cost is enjoyed elsewhere.

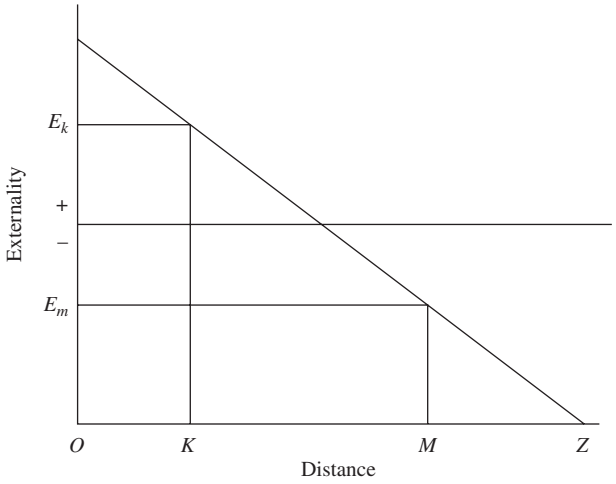


Figure 5 Distancing and spatial externality.

Externality Policy

Government policy toward externalities typically is focused upon the facilitation of positive externalities and the removal of negative externalities. The former typically requires some type of subsidy and the latter some form of tax or physical restriction on the activity that generates external costs. Much of government activity is involved with directly providing positive externalities. In the traditional sense, a public good was just that—something that could be used by the public with no direct charge or a charge lower than the actual value. Non-toll highways and other forms of infrastructure, for example, provide widespread benefits to many. Their cost is borne by the public at large, but individual users are not charged their marginal costs while their marginal benefits are significant.

Governments also have tried to specifically target externalities to particular areas. Enterprise zones (also called empowerment zones) are one example. They provide special tax benefits or labor subsidies to companies operating within specifically bounded districts, thereby directly providing geographical externalities, with little or no spatial effect. Recently, proponents of industrial policy have used the results of analyses incorporating increasing returns to argue for more active geographical policy toward inducing positive production externalities. Policies that facilitate locational clustering of firms within a broad industry are developed with the intent of nurturing such positive externalities as knowledge spillovers among firms and the development of a targeted labor force culture. Tax incentives for particular types of research and development activities or promotion of public university–private industry research or labor training consortiums are typically promoted as methods of inducing positive spatial externalities.

Much of land use planning is applied government policy toward externalities. Zoning, for example, is used to alleviate the negative spillover effects that certain land uses have on others, and green space requirements are designed to promote positive externalities associated with environmental amenities. Governments restrict certain production processes to certain places because of problems associated with water pollution or noise, and sometimes production processes are simply outlawed because of especially dangerous externality effects. The use of DDT as an agricultural insecticide is one example.

Taxation is another method of controlling negative externalities—the so-called "polluter pays" principle. Pigouvian taxes, named for their originator A. C. Pigou, are charged so as to ensure that a polluting firm pays its social cost of production rather than its marginal private cost. Returning to Fig. 1, a Pigouvian tax in the amount of $P'-P$ would reduce polluting activity to a "socially efficient" level as production is decreased from Q to Q'.

An argument against Pigouvian taxation specifically and government policy toward externalities in general is derived from the Coase theorem, which states that the initial distribution of property rights has no bearing on the use of the property because its use will ultimately be acquired by the person who values it the most. Given Coase's theorem, if the costs imposed by a negative externality become extreme, a market will be established so that the externality, as an unpaid cost or unpriced benefit, will be eliminated. Government intervention is unnecessary. For example, if water pollution becomes too severe, the polluter will be paid not to pollute. In effect, the right to pollute will be purchased from the polluter. Under Coase's theorem, negative externalities will be bought off and the benefits associated with positive externalities will be purchased outright. Land use zoning, for example, is unnecessary because private property markets will, theoretically, provide any compensation required for costs to one property owner raised by the actions of another property owner.

The only drag on the process is transaction cost. If transaction costs are high, externalities remain because the total purchase price (including the transaction cost) of an externality's source can exceed the social cost or benefit of the externality. If transaction costs are high, government intervention may indeed be necessary, even under Coase's theorem, to prevent a loss to one property owner from the actions of another.

See Also the Following Articles

Built Environment • Taxation

Further Reading

Breschi, S., and Lisoni, F. (2001). Localised knowledge spillovers vs. innovative milieux: Knowledge "tacitness" reconsidered. *Papers Reg. Sci.* **80,** 255–273.

Chase, J., and Healey, M. (1995). The Spatial externality effects of football matches and rock concerts. *Appl. Geog.* **15,** 18–34.

Coase, R. H. (1988). *The Firm, the Market, and the Law.* University of Chicago Press, Chicago, IL.

Freeman, A. M., III (1993). *The Measurement of Environmental and Resource Values.* Resources for the Future, Washington, D.C.

Hardin, G. (1968). The tragedy of the commons. *Science* **162,** 1243–1248.

Hanink, D. M. (1995). The economic geography in environmental issues: A spatial-analytic approach. *Prog. Hum. Geog.* **19,** 372–387.

Hanson, G. H. (2001). Scale economies and the geographic concentration of industry. *J. Econ. Geog.* **1,** 255–276.

Papageorgiou, G. J. (1978). Spatial externalities I: Theory. *Ann. Assn. Am. Geog.* **68,** 465–476.

Sunley, P. (2000). Urban and regional growth. In *A Companion to Economic Geography* (E. Sheppard and T. Barnes, eds.), pp. 187–201. Blackwell, Oxford, UK.

Spatial Pattern Analysis

Arthur Getis
San Diego State University, San Diego, California, USA

Glossary

georeferenced data Data for which each element of the data set can be identified by its exact geographic location by means of a coordinate system.
geographical information system A technology comprising a set of computer-based tools designed to store, process, manipulate, explore, analyze, and present spatially identified information.
modifiable areal units A set of measurements on georeferenced zones that can be subdivided or aggregated into smaller or larger zones, respectively. Any results of an analysis of these subdivisions or aggregations may be conditional upon the chosen set of zones.
spatial autocorrelation The correlation between values of the same variable at different spatial locations.
spatial heterogeneity The condition when the mean and/or variance of a spatial distribution of a georeferenced variable differs over the study region.
variogram A function that describes the differences in values between all pairs of georeferenced data over a range of distances.

Spatial pattern analysis includes procedures for (1) the identification of the characteristics of georeferenced data, especially as they are portrayed on maps, (2) tests on hypotheses about mapped patterns, and (3) construction of models that give meaning to relationships among georeferenced variables. Georeferenced variables are the set of observations (data) about a variable for which each observation can be located exactly on a map by means of the use of a specified coordinate system, such as the latitude-longitude system. The data can be represented on a map as points, lines, and areas. This article devotes attention to maps of points, either where each point represents the exact location of the occurrence of a particular phenomenon (a single event), or where a point is weighted to represent a realization of a random variable for a particular bounded region (multiple events).

Introduction

The identification of the characteristics of mapped data can use a wide variety of procedures, many of which can be carried out within a geographical information systems (GIS) environment. For example, GIS provides tools that make it possible to measure distances between mapped objects, find summary measures of the density of mapped data, and identify similarities and differences between spatial patterns. Identification procedures are often exploratory in nature. Practical examples of exploratory procedures are to find the area of a region that is covered by a particular type of land cover and to provide a measure of the spatial distribution of a georeferenced variable.

Those who theorize about the spatial configuration of georeferenced variables often create hypotheses in order to verify or reject notions about spatial patterns. For example, suppose it is assumed that a georeferenced variable will display a spatial trend over a particular region. Suppose that this assumption is based on a relationship between the variable in question and another georeferenced variable (e.g., rainfall affected by elevation). Special spatial analytic techniques are required to verify or reject hypotheses about the spatial relationship between the variables. These techniques are part of the tool kit of the spatial pattern analyst.

The construction of models usually requires that assumptions about relationships among variables hold to a specific degree. For example, suppose a model predicts that a number of georeferenced variables interact in specified ways. These variables can be studied in a mathematical model that expresses the relationships

between them. The spatial pattern analysis procedures must take into account the nature of the spatial patterns of the variables in question.

An important special concern of the analyst is the error that may arise in model development that cannot be explained by any variable. Appropriate spatial analytic techniques can be used to ensure that the error has no systematic manifestation on the map. Otherwise, the model may be deemed misspecified.

The Nature of Spatial Pattern Analysis

A Brief History

Knowing that the spatial perspective is an important aspect of knowledge, analysts have always sought to better ways to depict data on maps and test notions based on some expected pattern form or structure. In the academic field of geography, there is a long history of the development of cartographic devices that allow for particularly insightful views of spatial data. From the wind rose of Leon Lalanne in the 19th century to the map transformations of Waldo Tobler in the 1960s, the literature is filled with interesting ideas designed to allow the spatial data to "speak for themselves." With the widespread arrival of powerful computers in the 1970s and 1980s, it was just a short step from relatively uncomplicated explorations and tests to powerful research tools that have the ability to manipulate large amounts of georeferenced data on many variables.

Spatial pattern analysis was a field given little attention until the 1950s and 1960s, when biologists such as Clark and Evans, Pielou, and Skellam sought to test hypotheses about the spatial distribution of certain plants. Of note is the work of the geographer Michael Dacey, who, taking the lead from plant ecologists, tested various statistical distribution theories by using sets of georeferenced data that represented the location of towns in a settlement system. Geographers Garrison, Berry, Dacey, and King, among others, set out to test the ideas of the great German settlement theorists, Christaller and Lösch. The antecedents for the modern statistical analysis of spatial pattern data include this work on settlement theory and also a wide variety of work in other areas, especially in spatial interaction and spatial correlation theory.

Also, in the mid-20th century, the statisticians Moran and Geary developed distribution theories for spatial autocorrelation, the association of elements of a georeferenced variable to each other. Building on their work, Dacey addressed the issue of the possible association among contiguous spatial units. This led to the work of Cliff and Ord, whose monograph *Spatial Autocorrelation*, published in 1973, opened the door to new, more analytically sophisticated approaches to spatial pattern analysis.

Part of the spatial pattern analysis movement, concomitant with these developments but totally separate from them, has been the development of the field of geostatistics. Geostatistics was largely a response by geologists to the problem of predicting the location of yet-undiscovered valuable minerals such as gold and petroleum. The pioneer in this field, Matheron, in the 1960s used the term regional variable as this article uses the term georeferenced variable. The statistician Cressie has done much to organize the extensive literature on the subject and to show how specialized techniques such as variogram analysis and kriging can be used to study map patterns.

All of this work has been greatly affected by computers. In fact, the resurgence of interest in spatial pattern analysis in the 1980s and 1990s is directly associated with the ability of computers to process large amounts of spatial data and to map data and outcomes of experiments very quickly and cheaply. Today, most spatial pattern analysis techniques are parts of different, and sometimes competing, software routines.

Current Uses of Spatial Pattern Analysis

In recent years, concerns about the physical and cultural environment have stimulated the rapid development of spatial pattern analysis. There are a variety of spatial pattern analysis research topics that not only are of considerable concern to societies around the world, but also are the kinds of problems to which the latest technologies lend themselves. These include such issues as the study of

- the occurrence and transmission of disease,
- the location and abatement of crime,
- the development and testing of models concerning environmental variables,
- traffic management,
- the data and models related to social, cultural, and economic trends.

Limitations of Spatial Pattern Analysis

The value of spatial pattern analysis comes from its ability to yield insights about processes that occur in the real world. The spatial patterns observed by analysts, however, are only abstract depictions of the real world. What is learned from spatial pattern analysis is only as valid as the assumptions that are made about the depictions.

Problems Associated with Spatial Pattern Analysis

The following six problems faced by the spatial pattern analyst help to define this field of study. By taking each of

these problems into account, the analyst gives more meaning and authenticity to the subject.

The Modifiable Areal Unit Problem

The modifiable areal unit problem consists of two related parts; the scale problem and the zoning problem. The scale problem is the challenge of choosing an appropriate geographical scale for analysis: should it be state, county, town, neighborhood, household, or individual? Oftentimes, the answer to the question is not necessarily obvious. Sometimes available data provide the only option, but if the level of analysis is inappropriate for the subject being studied, the conclusions reached after study will be of little use. If data must be aggregated into larger units because, say, individual data represent too little of the earth's surface, the question arises whether too much is lost by the aggregation process. The so-called ecological fallacy results when conclusions based on data at an inappropriate spatial scale are used to interpret a process at another scale.

The zoning problem concerns the spatial configuration of the sample units. Study results can differ depending on the boundaries of the spatial units under study. For example, if the counties of a state were configured differently, the results based on data taken from those counties would be different.

The Spatial Association Problem

The association between spatial units affects the interpretation of georeferenced variables. When the boundaries of spatial units bisect a single entity, such as a metropolitan area or a cluster of animals, the effect that the bifurcation has on pattern analysis is referred to as spillover, and the result is spatial dependence between bisected entities. For most studies of patterns, the degree of dependence within the spatial configuration must be known. The study of spatial autocorrelation addresses the spatial association problem. Certain regression models, called spatial autoregressive models and spatial filtering models, include the spatial dependence effects as part of the model structure.

The Spatial Heterogeneity Problem

The spatial analyst must be aware of the degree to which a pattern differs from area to area within the study region. As an extreme example, a study region split between water and land will normally show wide differences in observations taken from variables representing the entire region. If the resulting heterogeneity is not taken into account, false conclusions about spatial processes might ensue. The analytical assumption that calls for a homogeneous spatial surface is called stationarity, a fundamental assumption used in geostatistics. When homogeneity cannot be assumed, analysts must find ways to understand the nature of the heterogeneity.

The Boundary Effects Problem

In spatial analytical work, there is the danger that research results will be biased by a boundary effect. It is very common when considering spatially depicted phenomena, that areas close to boundaries, such as near a nation's boundary or an ocean boundary, will be very much different than the remaining part of the region. For many studies, the boundaries of the study region are selected arbitrarily. Most statistical devices available for the analysis of mapped patterns do not take boundaries into account. Measurements in and around boundaries may alter conclusions that might be drawn from the remaining data.

The Sample Size Problem

Modern data collection methods, such as remote sensing, are capable of supplying information in amounts, detail, and combinations that can boggle the mind. The increased availability of large, georeferenced data sets and improved capabilities for visualization, rapid retrieval, and manipulation within a GIS all point to the need for ways to approach spatial pattern analysis. In recent years, exploratory spatial data analytic techniques have been augmented by data-mining routines. They are a good example of the possibilities and problems that a new technology generates. On the one hand, data mining is able to tease out of large data sets patterns of location and behavior that previously could not have been identified easily. On the other hand, without theoretical justification, it is not always clear if the mined patterns are legitimate abstract views of objective reality.

Associated with this problem is what is called the missing data problem. For example, a series of temperature readings in an area represent only a small sample of all of the possible readings. Thus, the analyst is forced to interpolate and thereby create values that are not readily available. The unavailable values are sometimes called missing data. The analyst must use valid interpolation techniques to avoid unexpected biases.

Spatial Pattern Statistics

This section briefly describes one example of a spatial pattern statistic taken from each of three areas of concern. Each concern is representative of a particular way that a point pattern may be viewed. Point pattern analysis (see Fig. 1) is designed mainly for the identification of clustering or dispersion of events or cases represented as points. K-function analysis is used as the point pattern example technique. Spatial dependence analysis is

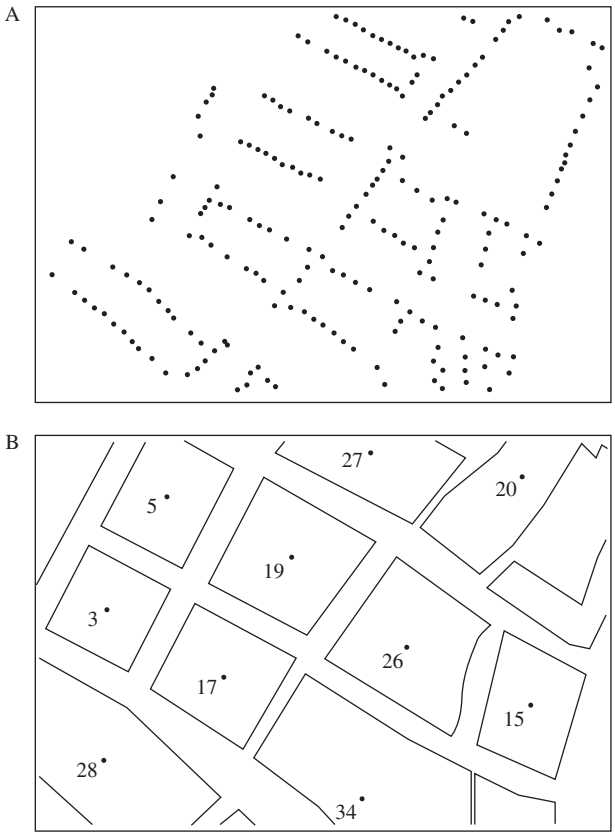


Figure 1 Types of point pattern representations. (A) Each point represents a single event; in this case, a point is a house site in a section of Iquitos, Peru. (B) Each point represents multiple events; in this case a weighted point is the sum total of houses on city blocks in a section of Iquitos, Peru.

carried out using one or more of a number of spatial autocorrelation statistics or geostatistics representations. The representative statistic in this case is the popular Moran's *I* test for spatial autocorrelation. Finally, the evaluation of patterns by focusing on individual observations is illustrated using what are called local statistics. The statistics demonstrated here are part of a family known as Getis and Ord's G statistics.

Point Pattern Analysis

There is a wide variety of point pattern statistics. Two helpful computer packages are ClusterSeer and Point Pattern Analysis. Often used is Ripley's k-function analysis, which is outlined below.

K-function is also called second-order analysis to indicate that the focus is on the variance, or second moment, of pairs of inter-event distances. All distances between all pairs of points are considered. The number of observed pairs within some specified distance (d) is compared to the expected number of pairs of points that would be obtained in a randomly created pattern of points [a test on the hypothesis of complete spatial randomness (CSR)]. The density of points, the boundaries, and the size of the sample are taken into consideration.

The k-function describes the number of pairs of points for each d. It is used in a formula [see Eq. (1)] that includes a correction for the boundary effect, stabilizes the variance, and allows the expected value [$L(d)$] to equal d. The confidence interval is generated by examining a specified number of permutations of randomly generated patterns of N points over the entire study area. If for any distance the observed $L(d)$ falls above or below the expected $L(d)$, the null hypothesis of CSR can be rejected at an appropriate level of significance. The level of significance is determined by the confidence envelope. An observed $L(d)$ value that falls below the envelope indicates that the points are dispersed at that distance, whereas an observed value above the envelope indicates that clustering is present.

The formula is

$$L(d) = \sqrt{\frac{\left(A\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}k(i,j)\right)}{(\pi N(N-1))}}, \qquad (1)$$

where A is the size of study area, N is the number of points, d is the distance, and $k(i, j)$ is the weighted number of pairs of points, which is estimated in one of threeways:

1. If there are no edge corrections, $k(i, j) = 1$, which is the case when $d(i, j) \leq d$; otherwise $k(i, j) = 0$.
2. If a point i is closer to one boundary than it is to a point j, the border correction is employed.

$$k(i,j) = \left[1 - \cos^{-1}\frac{e}{d(i,j)}\Big/\pi\right]^{-1},$$

where e is the distance to the nearest edge.

3. If a point i is closer to two right angle boundaries than it is to a point j, the weighting formula is

$$k(i,j) = \left\{1 - \left[\cos^{-1}\left(\frac{e_1}{d(i,j)}\right) + \cos^{-1}\left(\frac{e_2}{d(i,j)}\right) + \frac{\pi}{2}\right]\Big/(2\pi)\right\}^{-1},$$

where e_1 and e_2 are the distances to the nearest vertical and horizontal borders, respectively.

Spatial Dependence Analysis

It is imperative in any type of spatial analysis to recognize and account for the degree of spatial dependence found in the georeferenced data. Variables that have not been checked for spatial dependence or models that have not accounted for spatial dependence may contribute to unacceptable bias and misunderstanding.

Moran's *I* is the best known test for spatial autocorrelation. It is a cross-product statistic of the form characteristic

of Pearson's correlation coefficient. Because spatial association is a more complex phenomenon than the simple Pearson's correlation, several important modifications of the usual correlation coefficient formulation must be introduced. Moran's I is produced by standardizing the spatial autocovariance by the variance of the data. The statistic depends on a carefully chosen spatial structural specification, such as a spatial weights matrix [$\mathbf{W}$ with elements $w(i, j)$] or a distance-related decline function. The expected value of Moran's I is $-1/(N-1)$. Observed values of I that exceed the expected value indicate positive spatial autocorrelation, in which similar observations, either high numbers or low numbers, are spatially clustered. An I below the expectation indicates negative spatial autocorrelation, in which neighboring values are dissimilar. Moran's I is defined as:

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}, \quad i \neq j, \quad (2)$$

where $\bar{x}$ is the mean of x_i

$$\bar{x} = \sum_{i=1}^{N} x_i/N$$

and

$$S_0 = \sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j), \quad i \neq j.$$

The variance of I differs for different assumptions about the data. Under a randomization assumption, the variance of I is

$$Var(I) = \frac{N\left[S_1(N^2-3N+3) - NS_2 + 3S_0^2\right]}{(N-1)(N-2)(N-3)S_0^2} - \frac{K\left[S_1(N^2-N) - 2NS_2 + 6S_0^2\right]}{(N-1)(N-2)(N-3)S_0^2} - \left(\frac{1}{N-1}\right)^2,$$

where

$$S_1 = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[w(i,j)+w(j,i)\right]^2, \quad i \neq j;$$

$$S_2 = \sum_{i=1}^{N}\left[\sum_{j=1}^{N} w(i,j) + \sum_{j=1}^{N} w(j,i)\right]^2, \quad i \neq j; \text{ and}$$

$$K = \frac{N\sum_{i=1}^{N}(x_i-\bar{x})^4}{\left[\left(\sum_{i=1}^{N}(x_i-\bar{x})^2\right)\right]^2}.$$

Local Clustering Analysis

Local spatial autocorrelation statistics are observation-specific measures of spatial association. They focus on the location of individual points and allow for the decomposition of global or general statistics, such as Moran's I, into the contribution by each individual observation. Because the statistics can be used to detect local spatial clustering around an individual location, they are particularly well suited for finding "hot spots," or areas of elevated levels of the variable, or where a single measure of global association may contribute little meaning.

Local statistics that are used to find hot spots in an additive or multiplicative situation are G_i statistics. $G_i(d)$ and $G_i^*(d)$ were developed by Getis and Ord in 1992 and Ord and Getis in 1995. They indicate the extent to which a location is surrounded by a cluster of high or low values. The $G_i(d)$ statistic excludes the value at i from the summation while the $G_i^*(d)$ includes it. Positive $G_i(d)$ or $G_i^*(d)$ indicates spatial clustering of high values, whereas negative $G_i(d)$ or $G_i^*(d)$ indicate spatial clustering of low values.

The $G_i(d)$ statistic is written:

$$G_i(d) = \frac{\sum_{j=1,j\neq i}^{N} w_{ij}(d)x_j - \bar{x}_i\sum_{j=1,j\neq i}^{N} w_{ij}(d)}{S(i)\sqrt{\frac{\left\{(N-1)\sum_{j=1,j\neq i}^{N} w_{ij}^2(d) - \left[\sum_{j=1,j\neq i}^{N} w_{ij}(d)\right]^2\right\}}{(N-2)}}}, \quad i \neq j \quad (3)$$

where

$$\bar{x}_i = \frac{\sum_{j=1,j\neq i}^{N} x_j}{N-1} \quad \text{and} \quad S(i) = \sqrt{\frac{\sum_{j=1,j\neq i}^{N} x_j^2}{N-1} - (\bar{x}_i)^2}.$$

Both $G_i^*(d)$ and $G_i^*(d)$ are asymptotically normally distributed as d increases. Under the null hypothesis that there is no association between i and the j within d of i, the expectation is 0 and the variance is 1; thus, values of these statistics are interpreted as is the standard normal variate.

Spatial Pattern Modeling

Types of Models

A number of generic models are used to describe and study relationships between variables that are in some way affected by spatial location. These fall into a number of categories, briefly described below.

1. Spatial autoregressive models: These are regression models of the form

$$\begin{aligned} \boldsymbol{y} &= \rho\boldsymbol{W_1 y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &= \lambda\boldsymbol{W_2}\boldsymbol{\varepsilon} + \boldsymbol{\mu}, \end{aligned} \quad (4)$$

where it is assumed that the dependent variable, $\boldsymbol{y}$, is spatially autocorrelated and that the nature of the autocorrelation is subsumed by the stochastic variable $\boldsymbol{W_1 y}$. $\boldsymbol{W}$ is a spatial weights matrix, and the coefficient ρ is known as a spatial autocorrelation coefficient. Each independent variable, $\mathbf{X}$, can have a spatial weights matrix associated with it, and the error term, $\boldsymbol{\varepsilon}$, might be fashioned as a spatial variable as in Eq. (4).

2. Spatial filtering models: The models are designed to divide each independent variable into its spatial and nonspatial components. The degree of spatial autocorrelation embodied in each variable is extracted from the variable and then recast as a spatial variable in the model formulation.

3. Geographically weighted regression: The purpose of this type of model is to create a series of representative equations for a complex region when it cannot be assumed that regression coefficients will be stationary over the region being studied.

4. Variogram models: These models identify the nature of the decline in spatial autocorrelation as distance increases from each site within the study region. For example, it may be assumed that as distance between sites increases, the degree of spatial association decreases according to, say, a negative exponential function.

Tests on Pattern Models

Tests on pattern models is one of the most vexing problems in spatial pattern analysis. Often, spatial analysts require multiple tests of significance. An example is the use of local statistics to identify hot spots or clusters in a spatial data set. The search procedure might require that tests be carried out on a series of, or on all, georeferenced data points in a region. This gives rise to the problem of finding the appropriate bounds or cutoff values for multiple simultaneous tests. In addition, when the test sites are near each other, it is common that some of the data required for one test will be needed for another test. In those instances, the statistical tests are not independent of each other. In a GIS data set, one can easily imagine the need for, say, 100,000 tests, one for each data point. Of course, only the spatially close sites are likely to be correlated, but conducting so many tests raises the issue of the appropriateness of the well-known Bonferroni bounds. In a different sense, the problem is known in regression analysis, in which the estimation procedure induces some correlation among the standardized regression residuals, and the net effect is that the empirical distribution may be thinner tailed than in the normal distribution.

See Also the Following Articles

Computer-Based Mapping • Geographic Information Systems • Spatial Autocorrelation • Spatial Databases

Further Reading

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer, Dordrecht.

Boots, B., and Getis, A. (1988). *Point Pattern Analysis.* Sage, Newbury Park, CA.

Cliff, A. D., and Ord, J. K. (1973). *Spatial Autocorrelation.* Pion Press, London.

Cliff, A. D., and Ord, J. K. (1981). *Spatial Processes: Models and Applications.* Pion Press, London.

Cressie, N. (1991). *Statistics for Spatial Data.* Wiley, New York.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press, London.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000). *Quantitative Geography.* Sage Publications, London.

Getis, A. (1995). Spatial filtering in a regression frame work: Experiments on regional inequality, government expenditures, and Urban crime. In *New Directions in Spatial Econometrics* (L. Anselin and R. J. G. M. Florax, eds.), pp. 172–188. Springer Berlin.

Getis, A., and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geog. Anal.* **24,** 189–206.

Ord, J. K., and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geog. Anal.* **27,** 286–306.

Spatial Sampling

Peter A. Rogerson
State University of New York, Buffalo, New York, USA

Glossary

adaptive sampling Sampling of observations that is dependent upon the values of previous observations.
random spatial sampling Sampling of spatial units based upon random spatial coordinates.
spatial sampling Sampling in those instances when the sampling unit has a spatial dimension (point, line, or area).
stratified spatial sampling Selection of spatial units from spatial strata that are based upon subdividing the study area into mutually exclusive and collectively exhaustive strata.
systematic spatial sampling Systematic selection of spatial locations to achieve a spatial sample that has a uniform spatial distribution.

Spatial sampling is employed when the variable of interest has a set of spatial locations. The choice of spatial sampling locations needs to be made with particular prudence because spatial dependence is common, since sampled values from nearby locations will often lack independence.

Introduction

Like other forms of sampling, spatial sampling is concerned with selecting a sample from some larger population that is of interest. Spatial sampling is distinguished from many other types of sampling, however, by the fact that the variable of interest is distributed over geographic space. This leads to special considerations in the design of a sampling strategy because spatial variables exhibit spatial dependence—it would often be redundant to take two samples from locations in space that were very close to one another.

Different types of populations may be sampled; some are distributed over continuous space (such as the population of air pollution values in an urban area), and others consist of a set of discrete objects (such as the set of all households or the set of all census tracts).

Spatial sampling is also partially distinguished from other forms of sampling by the fact that the sampling units may themselves consist of geographic units (such as points or subareas) within the study area. There are other instances where the sampling units may not be geographic units (e.g., the sampling units may consist of families, households, or individuals), but because sampling those units yields a set of collected information on entities that have locations in space, the design and analysis of these studies also require a recognition of likely spatial dependencies in the data.

Spatial Sampling Problems, Sampling Designs, and Subsequent Inference

There are three distinct types of sampling problems:

1. Obtaining an independent, or almost independent, set of observations for use with classical statistical procedures. An example would be the use of a set of spatially sampled households in a regression analysis aimed at explaining household travel behavior as a function of socioeconomic variables.
2. Estimating a nonspatial characteristic of a spatial population. Examples include estimating a mean, a proportion, or a total.
3. Estimating a spatial quantity, such as a variogram or correlogram (which in turn summarizes spatial dependence), or an interpolated surface (e.g., a contour map).

Addressing these questions requires attention to the spatial design of the sample, which is in turn related to the subsequent use of the sample for statistical inference.

The approaches to inference following sample collection include design-based approaches and frequentist model-based approaches. In design-based approaches, the observed values are taken as fixed and nonstochastic. Inference is dependent upon the sample design and the associated probabilities of sample selection. Sampling strategies that are design unbiased (i.e., that have an expected value of an estimator that is equal to the population value) retain this property irrespective of the nature of the underlying population.

In frequentist model-based approaches to inference, the observations are taken to be random variables and are thus considered as realizations from some underlying stochastic process. For conventional designs (in which the method of sample selection does not depend upon the observations), inference can be based solely upon the stochastic model and is independent of the particular (conventional) sampling strategy. One example is model-unbiased estimators that remain unbiased regardless of the conventional sampling strategy.

Because a goal of sampling is to generalize from the sample to the population, sampling methods that lead to estimates with "good" statistical properties (for example, estimates that are unbiased, with minimum variance) are desired. Much work in spatial sampling has been devoted to assessing how well the quantities resulting from particular methods of sampling estimate their respective population values. For example, suppose that the mean commuting distance in a city is to be estimated. Figure 1 depicts two alternative sampling schemes that collect information from individuals at 10 spatial locations. The sampling plan in Fig. 1A would result in an estimate of the mean commuting distance that was highly uncertain, because so much of the city's area is not well represented in the sample. This is particularly true if commuting distance is positively spatially autocorrelated (that is, commuting distances by residents at one location are highly correlated with the commuting distances of residents at nearby locations). Any positive spatial dependence in commuting distance values implies that at least some of the information collected in a set of observations that are close together in space will be redundant.

In the commuting distance example, the population is a set of discrete objects—individuals. Similar considerations apply when sampling population variables are defined over continuous space, such as air pollution. Assuming that the variable of interest at each location is fixed (nonstochastic) at the time of sampling, there are some alternative spatial sampling strategies, described in the next section.

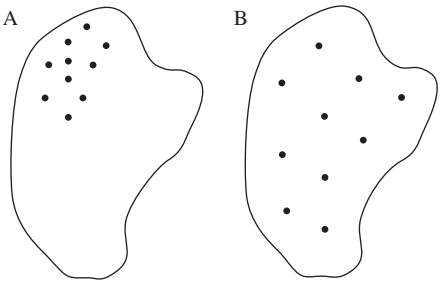


Figure 1 Alternative spatial sampling plans.

Alternative Spatial Sampling Designs

The most common spatial sampling designs are random, stratified, and systematic.

Random Sampling

Perhaps the most intuitive strategy for sampling locations in a defined study area is to choose random x and y coordinate pairs. Suppose a (possibly irregular) study area has minimum and maximum x coordinates of $x_{\min}$ and $x_{\max}$, respectively (see Fig. 2). Similarly, $y_{\min}$ and $y_{\max}$ are the minimum and maximum y-coordinates in the study area. To choose a random location, choose an x coordinate at random from the interval $(x_{\min}, x_{\max})$ and a y coordinate from the interval $(y_{\min}, y_{\max})$. If the point happens to fall outside of the study area (as with point A in Fig. 2), simply discard the point and try again. This is repeated to generate n locations. Note: If a random number generator returns a value (say, u) from a uniform distribution on the interval (a, b), this may be transformed into an x coordinate by taking

$$x = x_{\min} + (x_{\max} - x_{\min})(u - a)/(b - a).$$

The transformation is of course also used to find the y coordinate. This strategy can be used either to choose

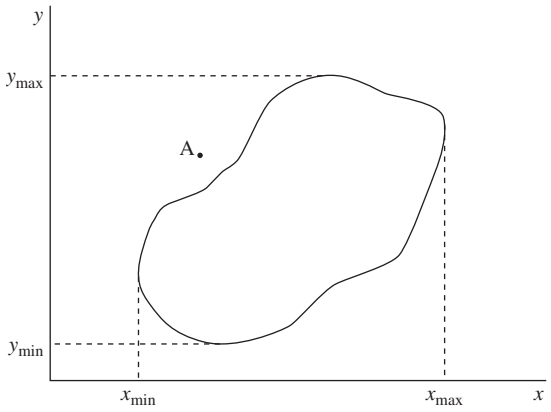


Figure 2 Hypothetical study area.

a set of point locations for variables defined over continuous space, or to randomly select spatial quadrats that in turn contain discrete objects (where in the second phase, either all objects or a sample of objects is taken from within the quadrat).

Stratified Spatial Sampling

The adequacy of estimates from spatial samples is based upon the amount of bias and the amount of precision, where the latter refers to the variance of the estimator. To reduce the variance of estimated parameters, it is often advantageous to stratify the population of sampling units according to some criteria known to affect the estimated quantity. For example, to estimate the proportion of residents in a community who have moved into their homes within the last year, it may be desirable to stratify on age, since age is a known covariate of mobility. Sampling from each stratum is then carried out, and the size of subsamples can either be proportional or disproportional to strata size. Stratified sampling can improve precision of population estimates by ensuring representation across a range of the stratifying variable known (or suspected) to be related to the estimate. Another use of stratification is to ensure precise estimates in each stratum, and in this case adequate sample sizes are necessary in each stratum.

With spatial sampling, it can be similarly advantageous to divide or stratify a geographic study area into subareas that are defined on the basis of some covariate. For example, a study of the recreational use of a county park might be carried out by first recognizing that frequency of use is likely to vary significantly with distance from the park. Concentric rings around the park would constitute strata, and the sample would consist of subsamples drawn from each stratum (see Fig. 3; note that the outermost stratum is noncircular in this example due to the irregular shape of the study region). One way to do this would be to select a subsample by randomly selecting individuals off of a list of all individuals residing in the stratum. If the goal was a precise estimate of average park usage, the number of individuals chosen in each stratum could be proportional to the number of individuals in the stratum. This would lead to an estimate of average park use that would be more precise than one obtained by random sampling. Suppose, however, that the goal was instead to obtain estimates of park use, disaggregated by distance from the park. If some strata had small populations, one could oversample in these strata to ensure more precise information about them (this is similar to the common practice of oversampling minority populations in samples stratified by race and/or ethnicity).

Systematic Spatial Sampling

With a list of N sample elements, a systematic sample of n may be chosen by dividing the list into n equal parts (where for simplicity it is assumed that N/n is an integer). The first observation is taken randomly from among the first N/n on the list; suppose we label this observation x, where $1 \leq x \leq N/n$. Then the remainder of the sample is chosen by taking as the next observations $x + N/n$, $x + 2\ N/n$, $x + 3\ N/n$, and so on.

Systematic sampling in such aspatial situations is often done as a matter of convenience. Generalizations of systematic sampling to the spatial case are desirable because they ensure comprehensive coverage of the study area. In addition, the likelihood of collecting redundant information from spatially dependent nearby locations is reduced to a minimum.

One approach to systematic spatial sampling is shown in Fig. 4. A study area is first divided into square cells, and then a point (e.g., point A) is chosen randomly within the first cell. Points are next chosen at the same relative locations in the remaining cells. There are numerous variations of this procedure. For example, when the sampling points are taken to be the center of each cell, the design is known as a centric systematic sample.

One potential though uncommon difficulty with systematic spatial sampling is that spatial periodicities may affect the estimate. For example, suppose that the housing

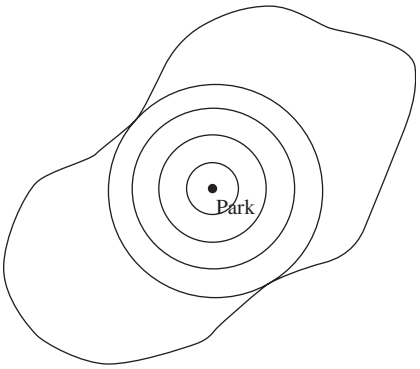


Figure 3 Sampling strata based upon distance from park.

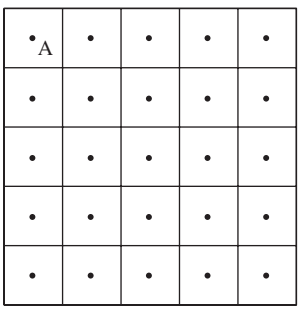


Figure 4 Systematic spatial sampling.

prices in an area are a function of location, and in particular are a function of elevation. There are some areas where housing prices are higher at higher elevations (because of the scenic amenities), and others where housing prices are higher at lower elevations (due to accessibility considerations). In either case, if hills and valleys are systematically spaced at roughly equal distances apart, it could be a mistake to take a systematic sample of housing prices because of the potential that the sampled locations would correspond entirely to high (or low) elevation locations. Judicious choice of the sampling interval is therefore called for. The problem may be avoided entirely where this possibility exists by geographic stratification.

Comparisons of the Three Spatial Sampling Methods

Many studies have compared the sampling methods described previously. When estimating the mean, all three methods yield unbiased estimates. Systematic and stratified spatial sampling are generally preferable to random spatial sampling, because the former two methods generally lead to estimates with smaller sampling variances. In addition, systematic spatial sampling is often found to have a slight advantage over stratified spatial sampling (in terms of a lower sampling variance).

Illustration

Figure 5 shows a hypothetical map of air pollution values. Suppose that it is decided to sample from the "true" map by taking observations at 16 point locations. The three types of sampling described previously were carried out as follows:

1. Random spatial sampling: coordinate pairs were chosen by randomly choosing x and y coordinates.
2. Stratified spatial sampling: the study area was partitioned into fourths by creating four square cells of equal area, and four observations were then chosen randomly within each cell.
3. Systematic spatial sampling: the study area was divided into 16 square cells of identical size. A point was chosen at random within the first cell, and the same relative positions within each of the other 15 cells were taken as sample points.

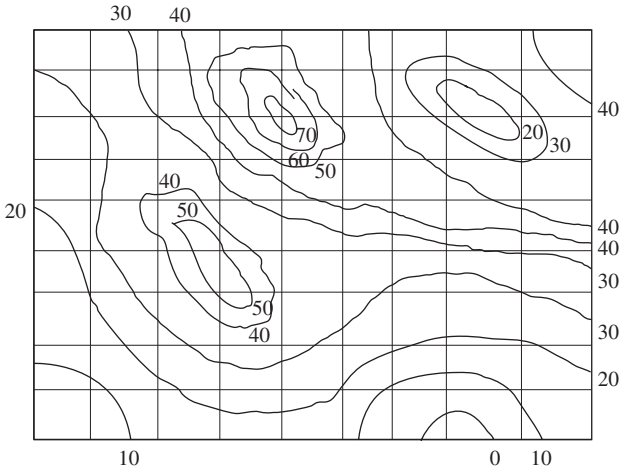


Figure 5 Hypothetical map of air pollution values.

One thousand samples of each type were taken. The results are shown in Table I. The sample means all compare favorably with the "true" mean of 30.36, found by overlaying a fine grid on the study area and sampling at the intersection of all grid lines. As expected, stratified and systematic methods led to lower variances than random sampling, and the variance of the mean under systematic sampling was less than that found under stratified sampling.

The stratified sampling used in the example may be further described as proportionate, since the size of the sample in each of the four strata was proportionate to the area of the strata. Disproportionate stratified sampling uses sampling fractions that are not proportional to, e.g., area. The 16 observations were next allocated disproportionately, under the alternative plans shown in Fig. 6; the figure also shows the mean and standard deviation associated with 1000 repetitions of each sampling plan. The sampling variance of the mean is lowest under the plan shown in Fig. 6A: this plan collects extra observations in the northwest corner of the study area, where air pollution values are most variable. Similarly, the sampling variance is highest under the plan shown in Fig. 6B; extra observations are "wasted" in the southeastern portion of the study area, where air pollution varies little from one location to the next. In general, it is useful to stratify into subareas when the subareas are very different from one another regarding the characteristic being measured. It is also useful to stratify into subareas when the variable of interest is more spatially variable within one or more subareas, as is the case here in the northwest quadrant.

Table I Comparison of Spatial Sampling Methods

Type of spatial sampling	*Mean*	*Standard deviation*
Random	30.38	3.08
Stratified	30.39	2.40
Systematic	30.24	1.95

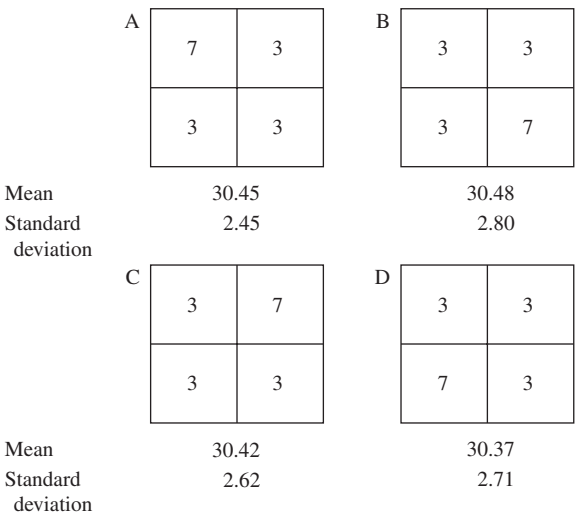


Figure 6 Alternative disproportionate stratified sampling plans. Numbers indicate the number of observations from each subarea.

This example used sampled point locations for a continuous variable; a comparison of corresponding areal sampling strategies (and then taking subsamples or censuses within those locations) would also typically result in the more precise estimates emerging from the stratified and systematic approaches.

Other Forms of Spatial Sampling

Cluster Sampling

Another type of spatial sampling is carried out via the hierarchical multistage sampling of spatial locations. For example, a sample of the census tracts in an urban area may be chosen in the initial phase (for example, via random selection of tracts, or via spatial or nonspatial stratification of tracts of different types). Then blocks may be sampled from the sampled tracts, and finally some or all of the households within the chosen tracts may be surveyed. The major benefit of this form of sampling is added convenience, since the cost of data collection can be substantially lower than alternative designs. For a given sample size, the resulting estimators will lack precision when, as is often the case, the variable of interest exhibits spatial dependence. Whether cluster sampling is effective for a fixed budget depends upon whether the additional sampling that is possible (due the lower cost of data collection) can increase the precision of the estimates sufficiently.

Adaptive Sampling

The sampling methods discussed to this point are conventional, in the sense that the method of sample selection does not depend upon the observations. One alternative to conventional designs is the set of adaptive sampling methods; these methods are particularly appropriate when the sampled characteristic is rare and spatially clustered. With adaptive sampling, the method of selecting observations may depend upon the observations. For instance, suppose that it is of interest to estimate the proportion of all individuals within a study region that walk to work. Because this is an uncommon mode of transportation, and because the location of such individuals is likely to be clustered spatially, adaptive sampling may be appropriate. One specific approach is to initially choose a census block randomly, and then canvas individuals within the block. If the sampled block reveals a high proportion who walk to work, it may then be advantageous to sample surrounding block locations, since positive spatial dependence will make it more likely that these adjacent areas will also contain individuals who walk to work.

Adaptive sampling methods can be operationally complex. Recall that in frequentist model-based approaches to inference, the observations are taken to be random variables, and for conventional designs, inference is independent of the particular (conventional) sampling strategy. With adaptive sampling, inference based on frequentist model-based approaches will depend upon both the model and the adaptive design.

Recent Directions in Spatial Sampling

The relationship between spatial dependence and spatial sampling has been the focus of recent research. For example, how spatial dependence may affect the optimal selection of a sampling method has been examined, when the objective was to estimate a map. This issue has also been addressed from the opposite perspective, to study the optimal sampling design for estimation of the nature of spatial dependence.

There is also increasing recognition of the need for space–time sampling methods, particularly within the context of environmental monitoring networks. For example, the topic of how monitoring stations may be added and/or deleted to achieve good estimations of environmental parameters that are changing over time has been discussed. Such methods are also of clear use in the context of social measurement (for example, in monitoring changes in neighborhood quality).

See Also the Following Articles

Clustering • Spatial Autocorrelation

Further Reading

Berry, B. J. L., and Baker, A. M. (1968). Geographic sampling. In *Spatial Analysis* (B. J. L. Berry and D. F. Marble, eds.), pp. 91–100. Prentice-Hall, Englewood Cliffs, NJ.

Bogaert, P., and Russo, D. (1999). Optimal spatial sampling design for the estimation of the variogram based on a least squares approach. *Water Res. Res.* **35,** 1275–1289.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Wiley, New York.

Das, A. C. (1950). Two dimensional systematic sampling and the associated stratified and random sampling. *Sankyha* **10,** 95–108.

Greig-Smith, P. (1964). *Quantitative Plant Ecology.* Butterworth and Co., London.

Griffith, D. A., and Amrhein, C. G. (1991). *Statistical Analysis for Geographers.* Prentice-Hall, Englewood Cliffs, NJ.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge University Press, Cambridge, UK.

Matern, B. (1960). *Spatial Variation.* Lecture Notes in Statistics, No. 36. Springer-Verlag, New York.

Quenouille, M. H. (1949). Problems in plane sampling. *Ann. Math. Stat.* **20,** 355–375.

Ripley, B. D. (1981). *Spatial Statistics.* Wiley, New York.

Stehman, S. V., and Overton, W. S. (1996). Spatial sampling. In *Practical Handbook of Spatial Statistics* (S. Arlinghaus, ed.) CRC Press, New York.

Thompson, S. K. (1992). *Sampling.* Wiley, New York.

Thompson, S. K., and Seber, G. A. F. (1996). *Adaptive Sampling.* Wiley, New York.

Van Groenigen, J. W. (2000). The influence of variogram parameters on optimal sampling schemes for mapping by kriging. *Geoderma* **97,** 223–236.

Wikle, C. K., and Royle, J. A. (1999). Space-time design of environmental monitoring networks. *J. Agric. Environ. Stat.* **4,** 489–507.

Spatial Scale, Problems of

Peter M. Atkinson
University of Southampton, Southampton, United Kingdom

Glossary

downscaling "Zooming in" or changing to a finer spatial resolution so as to reveal more detail in the data.
modifiable areal unit problem (MAUP) The problem associated with analyzing data defined on cells that vary in size, geometry, and orientation. The problem has two main components: aggregation and zonation.
regularization The operation of increasing the support (cell) on which a model (or data) is defined.
support The size, geometry, and orientation of the space on which a measurement, observation, or datum is defined.
upscaling "Zooming out" or coarsening the spatial resolution so as to reveal less detail in the data.
variogram A function that relates semivariance to lag (distance and direction of separation).

Spatial scale is an important concept in relation to social measurement, being inextricably tied with measurement and sampling. Importantly, the spatial variation evident in spatial data is a function of reality and the sampling framework. Therefore, it is important to distinguish between scales of measurement (i.e., those relating to the sample) and scales of spatial variation (i.e., those in data). The two most important scales of measurement are spatial resolution and spatial extent. These provide lower and upper limits on the scales of spatial variation that are detectable in data. Given a specific sample that can be modeled as a random field, the scale(s) of spatial variation in the data can be characterized using either functions of spatial resolution or functions of lag (distance and direction of separation), such as the variogram. It is possible to change the scale of measurement: Upscaling involves coarsening the spatial resolution through averaging, whereas downscaling involves increasing the spatial resolution, for example, through optimization or simulation. The variogram can be altered (regularized) as a function of support (the space on which each datum is defined). This amounts to scaling the model rather than the data. The major scale issue for social measurement is the modifiable areal unit problem. The problem is essentially that census and related data are defined on a variable support such that classical statistical techniques should not be applied directly without modification.

Defining Scale

Spatial scale has traditionally been defined by cartographers as the ratio between a distance on a map to the same distance in reality. This cartographic definition of scale is strictly correct. However, this definition may be confusing. For example, 1 : 10,000 is a larger cartographic scale (fraction) than 1 : 50,000, even though 1 : 10,000 is the smaller ratio and the 1 : 10,000 scale map covers a smaller ground area.

We often use scale in everyday language with a very different meaning. For example, a large-scale investigation, phenomenon, or process is simply a large investigation, phenomenon, or process. This use of scale, widely accepted and currently practiced in disciplines such as physics and ecology, simply equates scale to size (i.e., it renders the word "scale" redundant). This definition of scale is used in this article.

When describing spatial scale, it is important to distinguish between two types of scale: scales of measurement and scales of variation. Scales of measurement relate to sampling processes, whereas scales of variation relate to data. It is important to realize that the scales of variation observable in data are a function of (i) the sampling framework (and, therefore, the scales of measurement encompassed therein) and (ii) the phenomenon of interest in reality. Reality can never be observed independently

from a sampling framework (Fig. 1). All observation of the real world, even that obtained directly by our own senses, is essentially a filtered version of what exists in total. The importance of scale has to do with the way in which the sampling framework interacts with the phenomenon of interest in reality to produce (spatial) data.

This article draws heavily on geostatistics, a set of tools for the analysis of spatial data, meaning that the focus is on spatial variables that can be modeled as random fields. Point pattern data and object-based representations (e.g., points or cells grouped together to form higher order objects) are not covered in the same detail. The reason for this choice is that geostatistics provides a useful framework with which to explain the basic concepts relating to spatial scale, most of which can be applied to other types of data.

In addition, this article focuses on scales of variation in spatial data. Thus, scales of both temporal variation and process, although important, are not given the same attention. However, the basic concepts explained in relation to spatial data may be applied readily to scales of both temporal variation and process.

Scales of Measurement

There are two important scales of measurement; the spatial resolution and the spatial extent. To understand these properties, it is necessary to first discuss the sampling framework.

Sampling Framework

The sampling framework can be divided into properties that relate to a single observation (i.e., properties of the support) and other properties that relate to an ensemble of observations (i.e., properties of the sample).

Support

A measurement of a property defined spatially is always made on a positive space known as the support. This support has a size, geometry, and orientation. The size of support can be defined as zero (i.e., a point), but measurement is never actually possible on a point. An example is given by disease risk defined spatially. The risk is most often reported for a fixed areal unit, the support. If the risk were reported per 1×1 km cell, then the support would have a size of 1 km^2, it would be square, and it would have an orientation determined by the coordinate system used for sampling (i.e., the sampling grid) (Fig. 2A).

Sample Size

The sample size is simply the number of observations in the sample. For a sample of 100 respondents to a questionnaire, from a population of 10,000, the sample size is 100.

Sampling Scheme

The sampling scheme is the geometry of the sample. Examples include the random, stratified random, and systematic sampling schemes (Fig. 3). Systematic schemes include the square grid and equilateral triangular grid. Stratified random schemes can be stratified either by prior knowledge (e.g., disease risk is known to be less in nonvegetated areas) or by some prior sampling scheme (e.g., a square grid).

Sampling Density

The sampling density relates the actual distance units of the sample to distance units on the ground (in reality). It is an important property because it encapsulates some information related to cartographic scale.

Second-Order Properties

The properties of the sampling framework described previously are generally first-order properties because they relate to either a single observation or the sum of several observations. Two second-order properties may be defined as a function of these first-order properties; spatial resolution and spatial extent. To understand how these properties are constructed, consider the following

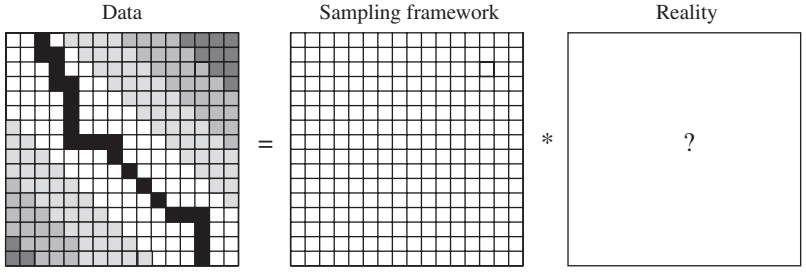


Figure 1 Data illustrated as a function of reality plus the sampling framework. Reality can never be observed independently of a sampling framework.

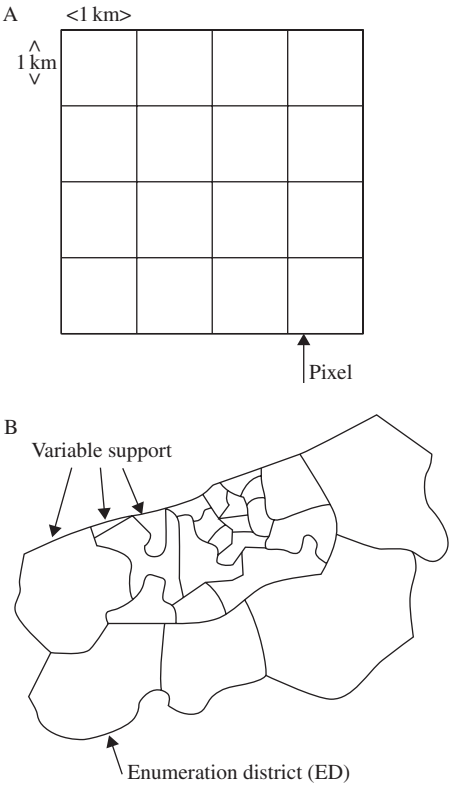


Figure 2 Data obtained on two different types of support: (A) a raster grid of square 1×1 km pixels and (B) census data for which the support is variable in size, geometry, and orientation.

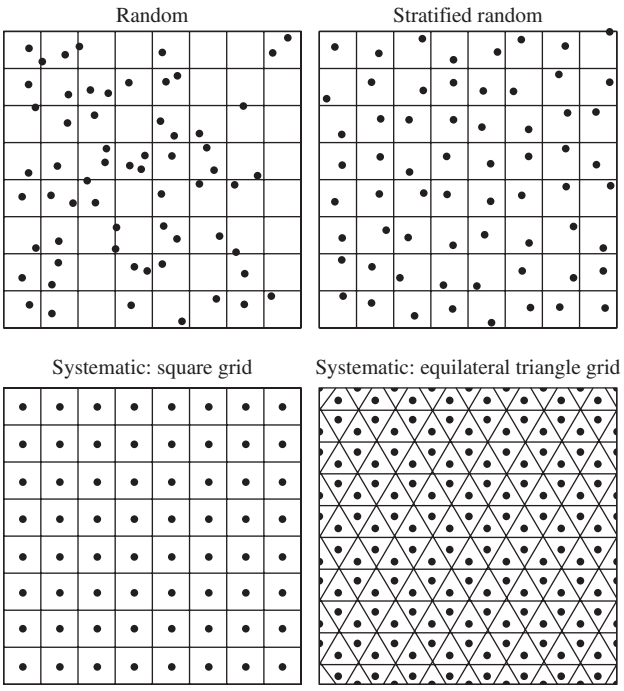


Figure 3 Four different sampling schemes.

example: Suppose that a map of disease risk has been produced with a support of 1×1 km, with 100×100 abutting observations on a square grid. The second-order properties (spatial resolution and spatial extent) are determined by the set of spatial distance and direction vectors (or lags) between each pair of constituent observations in the sample.

Spatial Resolution

The spatial resolution is defined as a function of the smallest lags between pairs of observations in the sample and the support. Since observations are often arranged to be nonoverlapping and abutting (as in the previous disease example), the spatial resolution is often equal to the support. However, it is quite possible for the spatial resolution to vary locally throughout a sample. Consider census data in which the size, geometry, and orientation of support may vary markedly from place to place (Fig. 2B). For such data, the spatial resolution depends more on the local lags between pairs of observations than on the support, and it varies locally. For point data (e.g., questionnaire respondents indexed to a point location), the support has no bearing: The spatial resolution is determined entirely by local lag. The important point is that spatial resolution is a second-order property because it depends on the relation or interaction between observations.

The spatial resolution defines a limit to the spatial detail that can be observed. It can thus be thought of as a scale of measurement. It is essentially a filter obscuring from the analyst variation that exists at a scale smaller than the smallest lag (sampling interval).

Spatial Extent

Like spatial resolution, the spatial extent is determined as a function of the set of lags between pairs of observations in a sample. Whereas spatial resolution depends on the shortest lags (sampling intervals), spatial extent depends on the largest lags. Spatial extent defines a second limit to the spatial variation that can be observed, filtering out variation that exists at a scale larger than the spatial extent. Thus, the spatial extent is a second fundamental scale of measurement.

The Set of Spatial Lags

Although it is useful to conceptualize the spatial resolution and spatial extent as providing lower and upper limits to the spatial variation that can be observed, it should be remembered that it is the set of spatial lags between all pairs of observations in the sample that provides the actual filter on reality. Why is this so? The answer lies in the nature of spatial variation. It is important to realize that spatial variation (and also spatial information) exists in the relations between data and not in them. Thus, if a value of 40 were realized by measurement, that value would

convey strictly zero information without comparison to either another value (e.g., it is less than 41) or to a priori knowledge (e.g., the measure is of human age, so the value conveys the meaning of "middle-aged adult"). In a spatial sample, the spatial variation and information exist as a function of the set of spatial lags between data pairs.

Scales of Spatial Variation

As explained previously, spatial variation exists in data as a function of the sampling framework and the phenomenon of interest in reality. It is not possible to measure reality independently of a sampling framework. Thus, when describing spatial variation (e.g., with summary statistics such as the variance) it is important to remember that the variation relates to data and not reality. It is for this reason that scales of measurement were discussed first.

The dispersion (or sample) variance s^2 will be familiar to readers. It is the square of the sample standard deviation s. The variance measures the spread of values around some unknown mean μ: this measures dispersion. Unfortunately, since the variance is aspatial (no locational information is provided) it describes only the amount or magnitude of variation and says nothing about the spatial scale of variation. To characterize the scale of variation, locational information must be used. There are several ways of doing this. Here, two are considered; the first is based on varying spatial resolution and the second on varying lag.

Characterizing Scales of Variation as a Function of Spatial Resolution

It is possible to obtain information on the scale(s) of spatial variation by making some statistic (such as the variance) a function of spatial resolution (or support). Essentially, the statistic is estimated for several spatial resolutions and the values are then plotted against spatial resolution. The plots convey some information on the scale(s) of spatial variation present in data. The techniques described in the following sections are applicable primarily to image data (e.g., a population surface model derived from census data).

Dispersion Variance

It is possible to vary the spatial resolution of data (by averaging cells successively) and calculate the familiar dispersion (or sample) variance at each step. The dispersion variance s^2 can be written more fully as $D^2(v, V)$ to indicate a variable defined on a support v within a region V. This may be obtained as

$$\hat{D}^2(v,V) = \frac{1}{N-1}\sum_{i=1}^{N}[\bar{z}_v(\mathbf{x}_i) - z_v(\mathbf{x}_i)]^2, \qquad (1)$$

where $\bar{z}_v(\mathbf{x}_i)$ is the mean of all values $\bar{z}_v(\mathbf{x}_i)$ at $\{i = 1, 2, \ldots, N\}$ locations $\mathbf{x}_i$. The plot of dispersion variance against spatial resolution reveals information about the scale(s) of variation. In particular, the rate at which the dispersion variance decreases with coarsening spatial resolution is determined by the scales of spatial variation present in the data. This relation (decreasing variance with coarsening spatial resolution) is important and is discussed later in relation to census data.

Scale Variance

An alternative to the dispersion variance is the scale variance. This is obtained by subtracting the variance obtained at a given spatial resolution from the variance obtained at the next finer spatial resolution. The scale variance $\hat{S}^2(v,V)$ may be obtained from $D^2(v, V)$ as follows:

$$\hat{s}^2\left(v \cdot 2^k, V\right) = \hat{D}^2\left(v \cdot 2^k, V\right) - \hat{D}^2\left(v \cdot 2^{k+1}, V\right), \qquad (2)$$

where 2^k varies between 0 and $\sqrt{N}$. This implies that the image is composed of $\sqrt{N}$ rows by $\sqrt{N}$ columns, and that $\sqrt{N} = 2^k$ for some value of k. One advantage of the scale variance is that peaks in the plot of scale variance against spatial resolution indicate where the scale of spatial variation is similar to the chosen spatial resolution (the plot is not monotonically decreasing as for the dispersion variance). A further advantage of the scale variance is that it is able to detect multiple scales of variation readily via multiple peaks.

Local Variance

The local variance σ^2_{vw} can be predicted for a moving $(2w + 1$ by $2w + 1)$ window applied to an image of L rows by M columns using

$$\hat{\sigma}^2_{vw} = \frac{1}{(2w+1)^2}\sum_{j=l-w}^{l+w}\sum_{k=m-w}^{m+w}\left[\bar{z}_v(j,k) - z_v(j,k)\right]^2, \qquad (3)$$

where w is usually set to 1. The local variance is different, in principle, from the previously described variances because it is local. This local variance, however, may be averaged over the whole image:

$$\hat{\bar{\sigma}}^2_{vw} = \frac{1}{L \cdot M}\sum_{l=1}^{L}\sum_{m=1}^{M}\hat{\sigma}^2_{vw} \qquad (4)$$

The average local variance $\hat{\bar{\sigma}}^2_{vw}$ can be predicted for different window sizes. The plot of $\hat{\bar{\sigma}}^2_{vw}$ against window size provides information that is similar to that provided by the dispersion and scale variances. Figure 4 shows

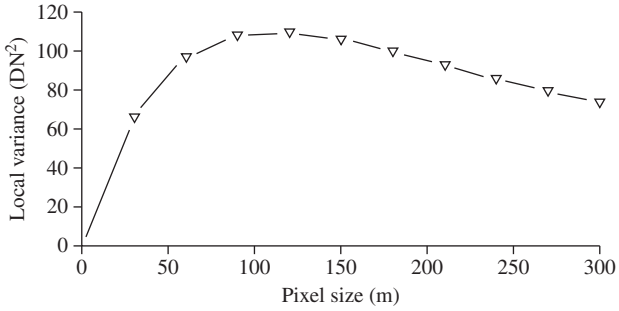


Figure 4 Average local variance plotted against pixel size. DN, digital number.

a hypothetical plot of average local variance against cell or pixel size.

Characterizing Scales of Spatial Variation as a Function of Lag

It can be difficult to interpret plots of variance against spatial resolution. Functions such as the autocovariance, autocorrelation, and variogram describe spatial correlation as a function of lag (distance and direction of separation). Of these three functions, the variogram is examined here because it is most often used in geostatistical analysis.

For continuous variables, the experimental or sample semivariance is defined as half the average squared difference between values separated by a given lag $\mathbf{h}$. Thus, the experimental or sample variogram $\hat{\gamma}_v(\mathbf{h})$ may be obtained from $\alpha = 1, 2, \ldots, P(\mathbf{h})$ pairs of pixels $\{z_v(\mathbf{x}_\alpha), z_v(\mathbf{x}_\alpha + \mathbf{h})\}$ defined on a support (or pixel) of size v at locations $\{\mathbf{x}_\alpha, \mathbf{x}_\alpha + \mathbf{h}\}$ separated by a fixed lag $\mathbf{h}$:

$$\hat{\gamma}_v(\mathbf{h}) = \frac{1}{2P(\mathbf{h})}\sum_{\alpha=1}^{P(\mathbf{h})}[z_v(\mathbf{x}_\alpha) - z_v(\mathbf{x}_\alpha + \mathbf{h})]^2. \qquad (5)$$

To provide a quantitative description of the character of spatial variation, a continuous mathematical model is fitted to the experimental variogram (most often using weighted least squares approximation). Variogram models can be divided into two general categories: unbounded and bounded. Unbounded models increase in semivariance monotonically with lag, without reaching a defined maximum. Bounded models reach a maximum value of semivariance (known as the sill variance, c) at a finite lag (known as the range, a). An example of a bounded model is given by the spherical model:

$$g(h) = c \cdot \left\{ \left(\frac{3h}{2a}\right) - \left(\frac{h}{2a}\right)^3 \right\} \text{ if } h \le a \qquad (6)$$

$$g(h) = c \qquad \text{otherwise,} \qquad (7)$$

where a is the nonlinear parameter, referred to as the range (Fig. 5A). The exponential model is bounded, but

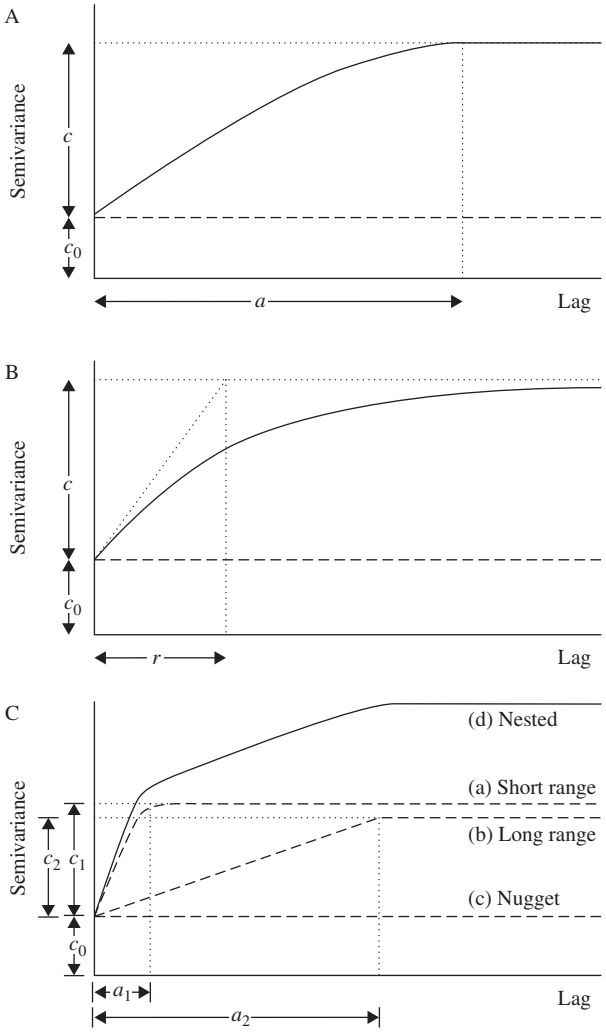


Figure 5 Examples of variogram models: (A) The spherical (plus nugget) model showing the structured (or sill) variance, c, the nugget variance, c_0, and the range, a; (B) the exponential (plus nugget) model showing the sill, nugget, and nonlinear parameter, r (which is approximately one-third of the effective range, a'); and (C) a hypothetical nested model in which short-range and long-range components are combined with a nugget component.

it never actually reaches the maximum because it is asymptotic. The exponential model is given by

$$g(h) = c \cdot \left\{ 1 - \exp\left(\frac{-h}{r}\right) \right\}, \qquad (8)$$

where c is the sill, and r is the nonlinear parameter, equal to approximately one-third of the conventional variogram model range a (Fig. 5B).

The sill variance c is equal to the *a priori* variance $D^2(v, \infty)$ (i.e., the variance of the property of interest defined on a given support in an infinite region). Furthermore, the dispersion variance $D^2(v, V)$ (i.e., the variance on a given support defined for a finite region) can

be obtained from the variogram model as the integral semivariance over that finite region. Thus, the sill variance conveys information on the amount of variation.

The range a provides a spatial limit to correlation. At lags larger than the range, pairs of data are expected to be unrelated, whereas at lags smaller than the range they are expected to be correlated. The range thus provides information on the scale(s) of spatial variation.

Often, more than one model is fitted to the experimental variogram in positive linear combination. If each model has a range, then the nested model describes the (multiple) scales of spatial variation (Fig. 5C). Nested variation is common in spatial data. In a few cases, the spatial variation is said to be self-similar or fractal, meaning that the same character of spatial variation exists at all scales of measurement. This fractal model is thus of key importance because it is the only example in which (at least the character of) spatial variation can be said to be independent of the sampling framework.

Commonly, a structured component model is fitted in combination with a nugget effect model. The nugget effect model is simply a constant semivariance with lag:

$$g(h) = c_0. \tag{9}$$

The nugget model has a sill (nugget) variance c_0 but no range (since it is flat) (Fig. 5). This so-called nugget variance describes unresolved spatial variation that exists at microscales (i.e., lags shorter than the shortest sampling interval) and measurement error.

In summary, the parameters (and type) of the fitted variogram model provide information on the amount and scale of spatial variation in sample data. However, the variogram is much more useful than simply as a description of spatial variation.

Changing the Scale of Measurement

When handling spatial data (e.g., within a geographical information system) it is often necessary to change the scale of measurement of the data (particularly the spatial resolution), most notably when one variable needs to be compared to another. Decreasing or coarsening the spatial resolution of the data (e.g., through averaging) is referred to as upscaling, whereas changing to a finer spatial resolution is referred to as downscaling (Fig. 6).

Upscaling

Upscaling is most often achieved through some form of averaging of original values to provide data at a coarser spatial resolution. For example, consider the situation in which the support is constant across space. Examples include the disease risk mapping described previously and a population surface model. Both data sets can be upscaled by calculating values for new larger cells as the averages of the two-by-two original cells that fit inside them. This kind of spatial averaging is the same as that used previously to calculate the image, scale, and local variances.

For census data (Fig. 2B), upscaling may also be achieved by averaging. However, the supports of the averages are constrained by the size, geometry, and orientation of the original supports. Essentially, the larger supports must be constructed to encompass perfectly the smaller original supports. This hierarchical system is commonly used for reporting census data. For example, in the UK census, enumeration districts (EDs) are encompassed within wards, which are themselves encompassed within metropolitan districts. Other forms of weighted averaging (e.g., the kernel density estimation used to create surface population models) involve interpolation and, thereby, some smoothing of the original data that occurs not only as a result of natural averaging but also as a result of interpolation. That is, the resulting support is larger than intended by the investigator and this causes problems in subsequent analysis.

Downscaling

Whereas upscaling is fairly straightforward given appropriate data, downscaling is relatively complex. It involves

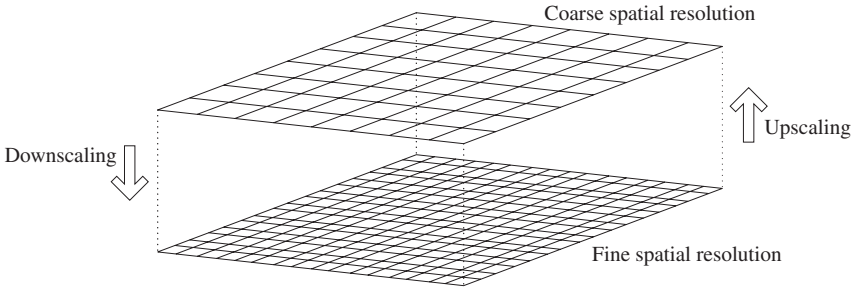


Figure 6 Two raster grids with different spatial resolutions illustrating upscaling (reducing detail) and downscaling (increasing detail).

increasing the spatial resolution of data. From theory, it is not possible to create new information (i.e., more than that provided in the original sample) by manipulation. However, in practice, downscaling may be achieved by transforming information from a nonspatial dimension into the spatial dimension or by simulating (thereby creating data but not new information about the property of interest).

Downscaling by Optimization

An example of downscaling was provided by the so-called superresolution mapping of land cover conducted by Tatem *et al*. in 2001. Multiple-waveband remotely sensed imagery was used to create a superresolution classification of land cover (e.g., woodland, grassland, and water)—that is, at a spatial resolution finer than that of the original imagery. This was achieved in two stages. First, multiwaveband values were used to predict land cover proportions per pixel using regression-type algorithms. Thus, a single pixel might be predicted to contain 70% woodland, 20% grassland, and 10% water. The prediction of land cover proportions (i.e., multivariate data) is made possible by the multi-waveband data provided in remotely sensed imagery. Second, the proportions in each pixel were allocated to subpixels within each pixel on the basis of some model of the character of spatial variation (e.g., variogram) and neighboring pixel values. This was achieved using an optimization algorithm.

Downscaling by Simulation

It is possible to create data at a smaller sampling interval than that of the original data by simulation. Simulation is most sensibly achieved if the model used takes into account both the actual values and the character of spatial variation in the original data. Such a model is provided by the geostatistical technique known as conditional simulation. Although no new information is provided, such simulated data can be useful for a variety of purposes (e.g., as input boundary conditions for spatially distributed dynamic models). For example, research has recently focused on dynamic modeling of the evolution of cities using cellular automata. Empirical data to drive such models are often provided by archive maps at coarse spatial resolution. Downscaling has the potential to provide realistic data with which to (i) run the model and (ii) evaluate uncertainty in the input data (since downscaling by simulation provides alternative realizations).

Scaling the Model

The geostatistical operation of regularization allows the variogram (or, alternatively, covariance or autocorrelation function) model to be scaled instead of the data. This is important because at the new spatial resolution, (i) the variogram describes the spatial variation evident; (ii) it is possible to obtain summary statistics, such as $D^2(v, V)$ and $D^2(v, \infty)$, from the variogram; and (iii) it is possible to (conditionally) simulate using the variogram. Thus, if it is possible to change the scale of measurement (i.e., the support) of the variogram, then it is possible to apply all of the previous functions at the new support without actually measuring on that support. A further example is provided by the image and scale variance statistics given previously, which can be predicted for different supports from the regularized variogram, without the need to coarsen the spatial resolution of the data.

The relation between the punctual or point semivariance and the regularized (defined on a support of positive size) semivariance at a lag $\mathbf{h}$ is given by

$$\gamma_v(\mathbf{h}) = \bar{\gamma}(v, v_{\mathbf{h}}) - \bar{\gamma}(v, v), \tag{10}$$

where $\bar{\gamma}(v, v_{\mathbf{h}})$ is the integral punctual semivariance between two supports of size $|v|$ whose centroids are separated by $\mathbf{h}$, and $\bar{\gamma}(v, v)$ is the average punctual semivariance within an observation of size $|v|$ (known as the within-block variance).

In words, the variation in the sample is equal to the variation in the property of interest minus the variation averaged out (i.e., lost) within the support. Thus, the variation discernible in spatial data is always less than that in reality because some variation is lost as within-block variance. This equation underpins the notion that data are always a function of the sampling framework and reality. It also explains why variance decreases as support size increases.

Figure 7 shows a punctual (point) variogram regularized to three different positive supports. Notice that the regularized variograms have a decreased sill variance (i.e., the within-block variance has been removed).

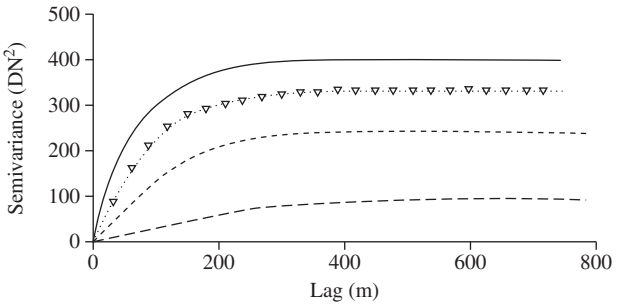


Figure 7 A punctual (point) exponential variogram model regularized to three different supports: 30 m (dotted line), 90 m (short dashed line), and 260 m (long dashed line). The symbols represent the experimental variogram observed on a support of 30 m.

Problems in Social Measurement

Here, two hypothetical examples from the social sciences in which individuals are averaged are considered; the first involves a constant support and the second a variable support.

Disease Risk

Disease risk is usually predicted as the number of infected individuals divided by the at-risk population:

$$\text{Risk} = \frac{\text{Infected}}{\text{At-risk population}}. \quad (11)$$

The at-risk population is often unknown, and in some cases it may be replaced by the total population. This value is reported for a given area (e.g., 1×1 km). There are several points to make about such a value.

First, the numbers involved affect the precision with which the value is known. The size of the at-risk population is inversely related to uncertainty: Small populations have greater uncertainty. Unfortunately, the at-risk population is likely to vary locally and, thus, uncertainty does also. For this reason, several authors have used a procedure known as empirical Bayes estimation in epidemiology to moderate extreme values that are less certain. Essentially, the information on risk (left-hand side of Eq. 11) and uncertainty (denominator in Eq. 11) is conflated into a single value (moderated risk). The smaller the denominator, the larger the moderation of the risk toward the mean. Uncertainty makes size of support an important consideration because generally, larger supports are likely to include a larger underlying at-risk population. For disease risk data, there is thus a trade-off between spatial resolution and precision of the variable.

Second, the number at risk (which inevitably will vary spatially) may affect the actual risk. For example, for contagious diseases the risk is often found to increase with population density (e.g., in urban, relative to rural, areas contagious diseases may spread more easily as a function of increased probability of contact). Thus, the scale of the process of infection may vary locally. Where the support is constant across space, this variation (in risk) may be difficult to separate from variation in uncertainty. Where the support varies locally to maintain a relatively constant underlying population, the risk and support effects may be confounded.

Despite the previously discussed issues, an advantage of disease risk data defined on a support that is constant spatially is that one cell may reasonably be compared with another, and classical statistical and geostatistical methods may be applied directly. Any scales of spatial variation existing between the 1×1 km supports may be detected and analyzed by the investigator. Spatial variation between point locations within each support will be averaged out and obscured from the investigator (Eq. 10).

The previous discussion implies that the support should be chosen carefully so as to resolve the spatial variation of interest. The tools discussed previously (e.g., the local variance plot and variogram) that describe the spatial variation in data can be helpful in making such a choice. Consider the plot of local variance against pixel size shown in Fig. 4. Most variation exists at scales of approximately 100–130 m. Thus, a new cell of up to 10 m on a side should be sufficient to resolve the spatial variation of interest. This new pixel size would be efficient because most of the information of interest would be conveyed at a much reduced data cost. The ability to make such choices about spatial resolution (i.e., scale of measurement) based on an understanding of spatial scale can be helpful to investigators.

If the support of the disease data were decreased to 100×100 m (i.e., an increase in the spatial resolution), then more variation would be revealed and, importantly, variation with a smaller range would be detectable. Problems inevitably occur as the support is decreased further: The risk per unit area is only defined for a limited range of support sizes. Although it may be possible to define risk on a 1×1 m support, it would be impossible to obtain suitable data with which to predict it. At the individual level, the actual property must be defined differently (i.e., risk per individual). In such cases, individual object-based models such as agent-based models must replace geostatistics.

Census Data

Census data are usually provided as values for census units that vary spatially in size, geometry, and orientation (Fig. 2B). This variation leads to the modifiable areal unit problem (MAUP). The main problem with such data is that since the support is not constant across space, it is not reasonable to compare values directly, and it is therefore not possible to apply classical statistical and geostatistical techniques directly to such data without modification. Why is this so? From Eq. (10), it can be seen that variation is a function of the support: larger supports lead to less variance (smoother variation) and *vice versa*. In census data, in which the supports vary hugely from place to place, comparison between data is problematic. The MAUP is often said to comprise two main components: the aggregation and zonation problems.

The Aggregation Problem

The aggregation component of the MAUP is similar to the regularization described previously. In the UK census, for example, data are presented for EDs. This is the smallest UK census unit. EDs are then aggregated into wards, metropolitan districts, and so on. The statistics associated with a given property (e.g., the number of cars per household per census unit) are affected by the level of

aggregation (i.e., EDs, wards, etc.). In particular, the variance is found to be less for larger units.

Much research in the social sciences has used correlation and regression analyses to describe the relations between variables. Researchers have found that the correlation coefficient r and regression parameters (α, β_i, $i = 1, 2, \ldots, n$) obtained were often a function of the level of aggregation. The specific case of the well-documented ecological fallacy arises when the results of an analysis conducted on aggregate data are used to describe individuals that form those aggregates. The ecological fallacy is thus a source of bias. Equation (10) provides an explanation, at least, for the ecological fallacy.

The aggregation problem for census data is much more problematic than previously implied. The ecological fallacy, for example, would hold true for image data. For census data, the problem is compounded by variation in the size, geometry, and orientation of the support across space. Essentially, a single data set of n units may comprise n different levels of aggregation and, therefore, n different variances. This problem is not resolved so readily, and it is the subject of current research.

The Zonation Problem

The zonation component of the MAUP is essentially a problem of small sample size (for aggregate statistics such as the variance). The problem is essentially that the actual realization of the sampling configuration (zonation) may have a major effect on the resulting statistics. For example, consider that a hot spot (in number of cars owned per household) exists in a given locality. If a census unit overlaps this area exactly, then the hot spot will show up clearly. If two units cross the hot spot and average its values with smaller values in neighboring areas, the hot spot will be greatly diminished in magnitude. Such effects are difficult to predict. In consequence, the single zonation provided by census bureaus such as the UK Office for National Statistics may be considered insufficient for mapping purposes. If many alternative realizations (zonations) were provided, the sampling may be adequate, and statistics such as the variance would converge to stable estimates. The problem then is that the spatial resolution is effectively increased and confidentiality may be compromised.

Conclusion

Problems of spatial scale are inevitably associated with problems of measurement and sampling. The scales of spatial variation detectable in spatial data are a function of (i) the intrinsic scales of variation that exist in the phenomena of interest in reality and also (ii) the scales of measurement encapsulated in the sampling framework. This statement is true whatever the nature of measurement: Even our own senses provide us with averages. After data are obtained, they represent the property of interest (and the statistics that describe it) defined on a given support and sampled with a given framework. Often, it is desirable to change that framework, but to do so requires some fairly sophisticated spatial statistical techniques. Of great theoretical interest is the geostatistical operation of regularization since this amounts to rescaling the model rather than the data. In the future, it would be useful to build the regularization operation into a spatial models to allow rescaling.

The problems associated with spatial scale for the social sciences are many and varied, not least because the underlying object of interest is usually the individual, and yet data are commonly reported as densities per unit area. Currently, the MAUP is the major scale issue for the social sciences because it is not possible to step beyond the initial impasse that cells should not be compared (therefore, variance, variograms, etc. cannot be estimated). It is possible to redistribute the varying areal units onto a raster (square) grid of equal cells (pixels). However, Atkinson and Martin have shown that this is valid only for very large new cells (i.e., there is much averaging involved) such that most information is thrown away. What is needed is a means of fitting a punctual (point support) model to variable areal unit data. This is the subject of current research.

See Also the Following Articles

Census Undercount and Adjustment • Census, Varieties and Uses of Data • Cognitive Maps • Computer-Based Mapping • Geolibraries • Multidimensional Scaling (MDS) • Spatial Externalities

Further Reading

Amrhein, C., and Wong, D. (1996). Research on the MAUP: Old wine in a new bottle or real breakthrough? *Geogr. Systems* **3,** 73–76.

Atkinson, P. M., and Martin, D. (1999). Investigating the effect of support size on population surface models. *Geogr. Environ. Modelling* **3,** 101–119.

Atkinson, P. M., and Tate, N. J. (2000). Spatial scale problems and geostatistical solutions: A review. *Professional Geographer* **52,** 607–623.

Journel, A. G., and Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press, London.

Martin, D. (1996). An assessment of surface and zonal models of population. *Int. J. Geogr. Information Systems* **10,** 973–989.

Moellering, H., and Tobler, W. R. (1972). Geographical variances. *Geogr. Anal.* **4,** 35–50.

Openshaw, S. (1984). The modifiable areal unit problem. In *Concepts and Techniques in Modern Geography CATMOG 38*. Geo-Abstracts, Norwich.

Quattrochi, D. A., and Goodchild, M. F. (eds.) (1997). *Scale in Remote Sensing and GIS*. CRC Press, New York.

Tatem, A. J., Lewis, H. G., Atkinson, P. M., and Nixon, M. S. (2001). Super-resolution target identification from remotely sensed images using a Hopfield neural network. *IEEE Trans. Geosci. Remote Sensing* **39,** 781–796.

Split-Half Reliability

Robert L. Johnson
University of South Carolina, Columbia, South Carolina, USA

James Penny†
CASTLE Worldwide, Inc., Morrisville, North Carolina, USA

Glossary

classical test theory A theory with the basis that an examinee's observed score on a test is the sum of a true score component and an independent random error component.
internal consistency Reliability of test scores based on the statistical association of responses to items within a test.
measurement error The difference between the observed score and the true score.
observed score A number that provides a description of an examinee's performance on a test.
parallel forms Tests that are considered interchangeable in that they have the same purpose, measure the same construct in the same manner, and use the same directions for administration. Statistically, parallelism requires the tests to have equal raw score means, standard deviations, error structures, and correlations with other tests.
parallel split method Estimation of reliability from the administration of a single test by the systematic division of test items to construct two interchangeable half-tests.
random error Nonsystematic measurement error that contributes to the variability of observed scores.
random split Estimation of reliability from the administration of a single test by the assignment of items by chance to construct two half-tests.
reliability Consistency of scores over repeated applications of a test across conditions that can include test forms, items, occasions, and raters.
reliability coefficient A statistical indicator that reflects the degree to which scores are free of random measurement error.
true/universe score The average of the scores that an examinee would obtain on an unlimited number of parallel tests.

Split-half reliability is a special case of the theoretical construct of reliability, which refers to the consistency of scores when a test is repeated on a population of examinees. The split-half procedure uses examinees' responses on a single test form to estimate score reliability. The test form is split by assignment of its items to two test halves. Scores from the two halves subsequently are used to estimate the degree to which scores would be similar if examinees were to complete a parallel test composed of similar items.

Introduction

Defining Reliability and Split-Half Reliability

The purpose of split-half reliability is to estimate the degree to which the results of a test would be similar if examinees were to complete a parallel test composed of a sample of similar items that measure the same construct. Split-half reliability is a special case of the theoretical construct of reliability, which refers to the consistency of observed scores when a test is repeated on a population of examinees or experimental subjects.

In classical test theory, the observed score (X) consists of two components, $T + e$, where T is a true score component of the observed score and e represents a random error component of the observed score. The true score is "conceptualized as the hypothetical average score resulting from many repetitions of the test or alternate forms of the instrument" (American Educational Research Association [AERA], American Psychological Association

[APA], & National Council on Measurement in Education [NCME]). The hypothetical difference between an examinee's observed score on a test and the examinee's true score for the instrument is referred to as measurement error.

Various sources of random measurement error contribute to the inconsistency of observed scores, and the testing community has developed procedures for estimation of reliability that take into account these error sources. In all these procedures, the purpose is to investigate the degree to which the observed scores for examinees reflect their true scores. For example, test scores for a population vary across the different occasions of a test administration. This source of score inconsistency addresses the temporal instability of examinees' test scores. When test developers want to estimate score consistency across different times, such a reliability investigation requires the use of test–retest procedures in which the same test form is administered to the same group of examinees on at least two occasions.

Additionally, observed scores of examinees vary across samples of items or tasks associated with a measure. Thus, error is attributed to the sampling of items in a content area, and reliability is assessed through the administration of parallel forms of a test. A third instance of measurement error reflects the variation of examinees' scores across items within one form of a test. It is the estimation of reliability in these instances that requires application of the split-half method or some other form of estimation of internal consistency.

Use of Split-Half Reliability

The estimation of score reliability across samples of items, as mentioned above, can be achieved through the use of the parallel form method. However, in some instances, the administration of parallel forms of a test is not possible. In such cases, test developers estimate reliability based on examinee responses to one test form. Feldt and Brennan indicated that reasons for the need to estimate reliability with only one form include (a) only one form of a test is produced because of the rare need for a second form; (b) the trait being measured is subject to rapid change, and (c) practical considerations might not permit the administration of more than one form of a test. It is these instances where the split-half method proves essential for the estimation of reliability. As Feldt and Brennan wrote,

> *For more than three quarters of a century, measurement theoreticians have been concerned with reliability estimation in the absence of parallel forms. Spearman (1910) and Brown (1910) posed the problem; their solution is incorporated in the well-known formula bearing their names.*

History

In 1910, Spearman and Brown simultaneously published articles in the *British Journal of Psychology* that outlined a method for splitting a test into two halves in order to investigate reliability. Spearman outlined the following method:

> *Let each individual be measured several times with regard to any characteristic to be compared with another. And let his measurements be divided into several—usually two—groups. Then take the average of each group; this we will term the "group average."* The division into groups is to be made in such a way, that any differences between the different group averages (for the same individual) may be regarded as quite "accidental." *It is further desirable that the sum total of the accidental variations of all the individuals should be not very unequal in the different groups; ordinarily, this will occur without further trouble, but in any case it can be arranged.*

Spearman subsequently described the method of splitting the test into two halves composed of odd items and even items that has become associated with the estimation of split-half reliability

> *A test of verbal memory, for instance, might well consist of memorizing twenty series of words. . . . Then series 1, 3, 5, . . . 19 would suitably furnish one group, while the even numbers gave the other. Any discrepancy between the averages of the two groups might, as a rule, be regarded as practically all due to the "accidents."*

Spearman offered the equation shown below for estimation of the increase of the reliability coefficient that was based on the division of measurements into two groups.

$$r_w = \frac{pr_{ab}}{1 + (p-1)r_{ab}} \tag{1}$$

where r_{ab} is the reliability coefficient based on the halves of the test, p is the number of times the test is lengthened or shortened, and r_w is the estimated reliability coefficient for the whole test. (When appropriate, subscripts across equations have been standardized to facilitate interpretation and comparison of the formulas.)

The focus of Brown's article was a study of the extent that a correlation exists between simple mental abilities and a general intellectual ability. The study was to examine the hypothesis of one single "central factor." In the study, many of the tests had two measures (parallel forms) that were amalgamated to form scores for correlation with other measurements. Brown labeled the reliability estimates for the measures as the *reliability coefficient* (r_2) *for amalgamated pair of tests*. The accompanying formula is the well-known

$$r_2 = \frac{2r_1}{1 + r_1} \tag{2}$$

where r_1 is the reliability for each test and "r_2 measures the extent to which the amalgamated results of the two tests would correlate with a similar amalgamated series of two other applications of the same test" (from Brown).

Methods for Calculation

Methods of Forming Split-Half Tests

Estimation of reliability through the use of the split-half procedure first requires the test developer to form the two halves of the test. The methods used to create the two halves include the assignment of odd items to one half and even items to another half, the random distribution of items to the two test halves, the allocation of items to form parallel halves, and the assignment of items to the two test halves to maximize item covariances.

Odd/Even Splits

Spearman described the development of the two test halves by the assignment of odd items to one half-test form and the even items to the other half-test form. More recently, Thorndike, Cunningham, Thorndike, and Hagen indicated that an even/odd split might be reasonable because items of similar form, content, or difficulty are likely to be grouped together. In essence, the even/odd split would produce two halves likely to be equivalent.

Random Splits

Early texts described the split-half method as forming chance halves of scores based on a single test administration (e.g., Brownell). In terms of random splitting, Cronbach used a test of vocabulary items to calculate the reliability estimates associated with all possible splits. He reported that the reliability coefficients ranged from 0.766 to 0.872 with a median of 0.823. He also noted that the random splits yielded tests that were not comparable in difficulty and variance. Cronbach recommended that the splits that are not comparable should be discarded. He also recommended that test developers should provide the means and standard deviations of the half tests.

Parallel Splits

The current *Standards for Educational and Psychological Testing* has taken the stance that the use of the split-half procedure to estimate reliability requires that the halves are parallel in content and statistical characteristics. Cronbach offered a method for the construction of the split-halves consistent with this guideline. Analogous to the parallel form method, the parallel split method requires the test developer to split the test items to construct two forms that are similar in content, difficulty, and range of difficulty. In order to do so, the test developer initially completes an item analysis of a subset of papers that will not be used in the subsequent reliability estimation. The information from the item analysis is used to select pairs of items with the same difficulty levels. Crocker and Algina indicated that test developers may rank order the items by difficulty and assign items with odd ranks to one form and even ranks to another form. Other considerations in the formation of parallel splits include (a) the assignment of items with similar content to the two test halves and (b) the pairing of items with the same format, such as multiple-choice with multiple-choice, true–false with another true–false. In the case of a group of items that deal with a single problem, such as a reading passage, the items are assigned intact to one half of the split.

Thus, the parallel split method illustrates that the split-half method of reliability estimation is theoretically similar to the parallel form method. This similarity is further reinforced by Cronbach in his consideration of the possibility that pairing of items in the parallel split method might introduce spurious factors in the reliability estimation. Cronbach posed the rhetorical question, "Would such a pairing be used in creating a parallel form of the test?" He then presented the parallel split method as analogous to the assignment of items to parallel forms.

Cronbach also applied the parallel splits method to the vocabulary data used for the random splits estimations of reliability. For the parallel split method, he reported that reliability coefficients ranged from 0.774 to 0.858 with a median of 0.825. In contrast to the random split method, the parallel split generally produced test halves of comparable means and variance. Given the similarity of the range of reliability estimates, Cronbach concluded the parallel split method did not offer much of an improvement over the random split for this one-factor test of vocabulary items. In 1951, Cronbach indicated that "marked variation in the coefficients obtained when a test is split in several ways can result only when (a) a few group factors have substantial loadings in a large fraction of items or (b) when first-factor loadings in the items tend to be very small or where they vary considerably."

Maximization of Item Covariances

Callender and Osburn described a method that optimizes the split-half coefficient by assigning items to the two test halves so that the sum of the item covariances is maximized. The procedure, referred to as MSPLIT, results in reliability estimates higher than KR20 and odd–even splits. The authors concluded that the procedure results in "somewhat inflated estimates of the corresponding population coefficients." If the method is applied to a subset of papers that will not be used in the subsequent reliability analysis, then the use of MSPLIT is analogous to

the use of item statistics in the formation of parallel splits. Thus, information about the test items is used to assign items in a manner that creates forms of similar content, difficulty, and variation.

Calculation of the Reliability Coefficient

Split-half reliability is typically estimated with the use of a Pearson correlation. Subsequently, the Spearman–Brown prophecy formula is applied to estimate the reliability of the full-length test. The Spearman–Brown method assumes that the two halves of the test are parallel. Parallelism requires that an examinee has the same true score across forms and the mean, variance, and error are the same across forms. If not, the estimated full length reliability for Spearman–Brown will be greater than obtained by other measures of internal consistency.

Not all calculations of split-half estimations use the Pearson correlation. Rulon provided two split-half formulas that he attributed to John Flanagan. One formula is based on the standard deviation of difference scores between the half-tests. The formula is

$$r_w = 1 - \frac{\sigma_d^2}{\sigma_w^2} \tag{3}$$

where $d = X_a - X_b$ and σ_w^2 is the variance for the whole test. Assumptions for this formula include (a) the difference between the two true scores for the two half-tests is constant for all examinees, and (b) the errors in the two half scores are random and uncorrelated.

The other formula is

$$r_w = \frac{4\sigma_a\sigma_b r_{ab}}{\sigma_w^2} \tag{4}$$

where σ_a is the standard deviation of scores for one test half and σ_b is the standard deviation associated with the other test half. Unlike the Spearman–Brown formula, these formulas do not require equivalent halves with equal variances. Both assume experimentally independent halves. Neither reliability estimate requires the application of the Spearman–Brown prophecy formula. Guttman offered the following contribution to the estimation of split-half reliability

$$r_w = 2\left(1 - \frac{\sigma_a^2 + \sigma_b^2}{\sigma_w^2}\right) \tag{5}$$

The terms σ_a^2 and σ_b^2 represent the variance associated with each test half.

If variances are equal on the two halves, the reliability estimate based on Spearman–Brown will be the same as achieved with the split-half procedures described by Rulon and Guttman. Moreover, the strict equality of variances is not required for convergence of reliability estimates across methods. According to Cronbach, if the ratio of the standard deviations for the two test halves is between 0.9 and 1.1, then Spearman–Brown gives nearly the same result as Eqs. (3) and (5).

Applications

To this point in the article, the discussion of the split-half method has focused on its application in the estimation of reliability based on a set of items administered as a single test form. Such a reliability estimate is completed to determine if results would be similar if students took a test composed of similar items. Cronbach indicated the purpose was to predict the correlation between two equivalent whole tests with two halves of a test. Thus, the method focuses on the consistency of performance across parallel sets of items.

An extension of the split-half method is seen in the estimation of interrater reliability. The scoring of constructed-response items, such as essays or portfolios, generally is completed by two raters. The correlation of one rater's scores with another rater's scores estimates the reliability scores based on a single rater. However, the reported score is often the average of the two scores assigned by the raters. If raters and observers are conceptualized as forms of a test (e.g., Feldt and Brennan), then pooling ratings is analogous to lengthening a test. Thus, the application of the Spearman–Brown prophesy formula to estimate the reliability for scores based on two raters is appropriate. As equivalence is assumed in the application of the split-half reliability estimation, the use of Spearman–Brown prophesy formula to estimate the reliability of scores based on two raters requires that the raters are equally qualified and producing ratings of similar means and variances. The addition of less qualified raters can weaken the reliability of the ratings.

Relationship to Other Reliability Coefficients

Cronbach's Alpha

In 1951, Cronbach introduced the alpha coefficient as an index of equivalence.

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_w^2}\right) \tag{6}$$

where k is the number of test items and σ_i^2 is the variance of item i. Using the split-half formula of Guttman, Cronbach demonstrated that alpha is the "mean of all split-half coefficients resulting from different splittings of a test." In addition, Cronbach

presented the split-half coefficient of Guttman as a special case of alpha. More specifically, in the case of $k = 2$, substitution in Cronbach's alpha results in

$$r_w = \frac{2}{2-1}\left(1 - \frac{\sigma_a^2 + \sigma_b^2}{\sigma_w^2}\right) = 2\left(1 - \frac{\sigma_a^2 + \sigma_b^2}{\sigma_w^2}\right) \quad (7)$$

Kuder–Richardson Formula 20 (KR-20)

In the instance where the items are dichotomous, item variance is written as $\sum p_i q_i$ and the above equation for coefficient alpha becomes Kuder's and Richardson's well-known KR-20,

$$r_w = \frac{k}{k-1}\left(1 - \frac{\sum p_i q_i}{\sigma_w^2}\right) \quad (8)$$

G-theory and Intraclass Correlations

The G- and D-coefficients of generalizability theory and Cronbach's alpha are special cases of intraclass correlations. Hoyt applied analysis of variance techniques to the estimation of reliability as the ratio of

$$r = \frac{MS_{persons} - MS_{pi}}{MS_{persons}} \quad (9)$$

where $MS_{persons}$ is the mean square for persons from an analysis of variance and MS_{pi} is the mean square for the person-by-item interaction. Hoyt's derivation, according to Cronbach, results in a formula equivalent to alpha. As such, intraclass correlations are linked to split-half reliability.

Limitations

Lack of a Unique Reliability Estimate

The lack of a unique reliability estimate received much attention in the early 1900s. For example, Brownell noted the wide variability of reliability estimates based on different splits. Cronbach wrote that "... chance element in splitting makes the reliability coefficient in error by an undetermined amount." Cronbach, however, noted this problem existed for all estimations of reliability for parallel forms. As an example, he noted that a test with four forms produces six reliability estimates, thus having no unique reliability estimate. This problem extends directly to split-half estimates of reliability owing to the different values that can arise from the different possible splits one might create for the computation.

Inflated Estimates of Reliability

Split-half reliability estimates based on an odd–even split will yield inflated reliability coefficients for speeded tests. Thus, tests designed to reflect the rate of work should use parallel form or test–retest approaches to assess reliability. Also, the use of split-half to estimate reliability when halves have unequal standard deviations can result in some inflation of estimates. When used in power equations to determine the sample sizes required to test the equality of two coefficients, Charter reported that the Spearman–Brown corrected split-half coefficients estimates may result in an underestimation of the requisite sample sizes.

Mixed Item Types

Although it is now common for an examination to be constructed using only one type of item (such as multiple-choice), more frequently national and state testing programs use tests composed of different item types, some of which do not produce dichotomous scores. If the split-halves technique is used to estimate the reliability of such tests, then the resulting reliability can be a function of the balance of item types in the two halves of the exam.

Several factors can depress the estimates of reliability for examinations containing mixed item types. These factors include (1) the relative restriction of range of dichotomous items as compared to scores produced for constructed-response items, (2) unanticipated changes in score distributions produced by mixed item types, (3) unexpected shifts in rater severity or leniency that can occur in the subjective evaluation of constructed responses but which are unlikely to influence the objective scoring of closed-form item types, and (4) unanticipated differences in difficulty that can arise with mixed item types. As mentioned earlier, the formation of parallel splits should take into account the item types in order to lessen any affect on the reliability estimates.

Item Difficulty

That test items can vary in difficulty is expected on most examinations, and this variability in difficulty can influence estimates of reliability produced using the split-half technique. If one split contains the easier items and the other split contains the more difficult items, the resulting reliability can be lower than it might otherwise be due to the potential disparate performance of students on the two sections of the exam. To lessen such influences, test developers can form parallel splits by using the statistical information from item analyses of a subset of tests that will not be used in the reliability estimation.

Item Placement

Even when a test is not explicitly designed as a speeded test, the possibility exists that elements of speededness will influence test results owing to the distribution of student ability. Students more deficient in the measured construct are often more likely to show fatigue earlier on the test than are the more able students, and for this reason, those less able students are more likely to perform more poorly on the items toward the end of the test. Hence, the latter test items have the potential to appear more difficult than the earlier items, producing the possibility of lower observed test reliability depending on which items are selected for the two halves.

Domain Sampling

Some examinations are constructed to assess a narrow domain. For example, one can easily find Italian vocabulary examinations targeted for particular levels of development. However, one can also find examinations designed to assess broader domains. To continue the example, one might consider an examination of students' grasp of the Italian language that addresses vocabulary understanding and their comprehension of reading passages.

The broader the domain sampled by the examination, the greater the possibility that the examination will exhibit multidimensionality, owing to the greater variety of cognitive skills required by students to successfully address the items and content. The degree of multidimensionality could influence the estimates of reliability produced by the split-half technique, depending on which items are assigned in which halves of the exam. Cronbach found that the parallel split is advantageous when group factors exist in a test and these group factors have loadings in the items that are larger than does the general factor.

Current Status and Directions for the Future

Nearly a century after the introduction of the split-half method, the role of split-half reliability is prominent enough to warrant explicit guidelines for its use in the current *Standards for Educational and Psychological Testing*. The literature in the social sciences continues to report reliability estimates based on the split-half method. In addition, test manuals often report split-half reliability estimates with other forms of reliability. A next step for the testing community is greater adherence to the recommendations advanced by AERA/APA/NCME for use of parallel splits in the formation of the test halves.

See Also the Following Articles

Alpha Reliability • Randomization • Reliability Assessment

Further Reading

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* American Educational Research Association, Washington, DC.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *Brit. J. Psychol.* **3,** 296–322.

Brownell, W. (1933). On the accuracy with which reliability may be measured by correlating test halves. *J. Exper. Education* **1,** 204–215.

Callender, J., and Osburn, H. (1977). A method for maximizing split-half reliability coefficients. *Education. Psychol. Meas.* **37**(4), 819–825.

Charter, R. (2001). Is it time to bury the Spearman–Brown "prophesy" formula for some common applications? *Education. Psychol. Meas.* **61**(4), 690–696.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Holt, Rinehart, and Winston, Fort Worth, TX.

Cronbach, L. (1943). On estimates of test reliability. *J. Education. Psychol.* **34,** 485–494.

Cronbach, L. (1946). A case study of the split-half reliability coefficient. *J. Education. Psychol.* **37,** 473–480.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16,** 297–334.

Cronbach, L. (1970). *Essentials of Psychological Testing,* 3rd Ed. Harper, New York.

Feldt, L., and Brennan, R. (1993). Reliability. In *Educational Measurement* (R. Linn, ed.), 3rd Ed., pp. 105–146. American Council on Education, Washington, DC.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* **10**(4), 255–282.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika* **6,** 153–160.

Kuder, G., and Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika* **2**(3), 151–160.

Rulon, P. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Education. Rev.* **9,** 99–103.

Spearman, C. (1910). Correlation calculated from faulty data. *Brit. J. Psychol.* **3,** 271–295.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., and Hagen, E. P. (1991). *Measurement and Evaluation in Psychology and Education,* 5th Ed. Macmillan, New York.

State Data for the United States

David M. Hedge
University of Florida, Gainesville, Florida, USA

Renée J. Johnson
University of Florida, Gainesville, Florida, USA

Glossary

analytic leverage The ability to explain social or political phenomenon with as few variables or little information as possible.
ecological fallacy Erroneously using group-level data to draw inferences about individuals.
measurement error Occurs when the procedure for assigning values fails to achieve either validity or reliability. The result is that differences in the values assigned to cases reflect flaws in the measurement process and not actual differences.
pooled cross-sectional, time series design Research design that examines variations across space and time.
significance tests A way of judging whether a particular value for a population parameter is plausible.

As the states assume greater authority for the governing of the United States, they truly can serve as "laboratories of democracy" by offering social scientists a unique context within which to examine basic questions concerning politics and policy. To achieve that potential, analysts must have access to valid and reliable state data and must address a number of issues concerning measurement and design. Fortunately, considerable progress has been made in recent years on both counts. There is an unprecedented amount of data available on the states, and social scientists are addressing several relevant methodological issues, including tests of statistical significance, ecological inferences, and the tradeoff between case studies and multistate analysis.

The States as Laboratories

The American states have frequently been hailed as "laboratories of democracy," a description often attributed to U.S. Supreme Court Justice Louis Brandeis in *New State Ice v. Liebmann* in 1932. The American states can also serve as laboratories about democracy. During the past several decades, there has been a marked resurgence of the American states. State governments have become more responsive to their citizens, the institutions of state government have been strengthened, and the states are assuming more responsibility for the day-to-day governing of the United States. However, what really commends the states as laboratories is their sheer diversity. Although a number of states have modernized their political institutions, many have not. Also, a review of the evidence suggests that some states are simply better at educating their children, reforming welfare, or regulating firms within their borders. It is those variations in politics and performance that offer scholars a unique opportunity to test fundamental propositions about politics and governance. Whether the states can play that role, however, depends in large part on the ability of scholars to obtain reliable and valid data on the states and successfully address key methodological issues surrounding the use of that information.

Sources of State Data

In 1982, Malcolm Jewell chided political scientists for doing too little systematic, comparative research on state politics. In seeking to reinvigorate the field, Jewell

argued that "the first need is for a much more comprehensive and systematic collection and analysis of comparative state political data" (p. 643). Although things were beginning to change when Jewell made his plea, few political scientists had the resources to mount a major data collection effort across multiple variables and states. Instead, scholars interested in doing comparative state politics research drew on just a handful of sources, most notably the "Statistical Abstract of the United States," the "U.S. Census of the Governments," and the Council of State Government's annual "Book of the States."

Much has changed since Jewell first issued his challenge. Today, political scientists have access to a stunningly large amount of data, much of which are available online or through various data archives scattered throughout universities and private organizations. The challenge today is to (i) identify available data and (ii) assess its validity and reliability. Neither task is easy. Although the amount of data on the states is seemingly boundless, there have been few efforts to catalog or archive the bulk of that information. The problems of reliability and validity are especially pronounced. Scholars who use what is essentially archival data have no control over what data are collected and, in many cases, too little information about how those data were collected.

Four kinds of data are available: data on (i) the states' economic and social characteristics (the "control variables" in much research), (ii) elections and parties, (iii) political institutions, and (iv) state policy outputs and outcomes. These data can be assessed through a variety of sources, including original data sources, data archives, and data links. Several online and published sources provide selected original data on some or all of the states in a particular area (e.g., health care or criminal justice statistics). In some cases [e.g., the Urban Institute's state database or data provided through the Inter-University Consortium for Political and Social Research (ICPSR)], data are available as downloadable data files. More typically, data are presented in tables and charts. In other instances, data and information have to be gleaned from narratives describing state programs. Data can also be accessed through a small number of data archives, including the ICPSR and Florida State University's State Politics and Policy Data Archive. Finally, a number of online sites provide links to a variety of data sources.

Original State Data Sources

U.S. Government

The U.S. government is the major source of original data on the American states. In addition to the Census Bureau, a variety of federal agencies provide considerable information on the condition of the American states and their citizens. The following is a partial listing of these agencies and the kinds of data they offer:

- Administration for Families and Children (Health and Human Resources): Data on Temporary Assistance for Needy Families (TANF) program spending, enrollment, recipient profiles, and workforce participation rates. Selected years available online at http://www.acf.dhhs.gov/acf_research_planning.html#stats.
- Bureau of the Census: "Statistical Abstract of the United States" (http://www.census.gov/statab/www). Online version of the "Statistical Abstract." Available for 1995–2002.
- Bureau of Economic Analysis, http://www.bea.doc.gov (Department of Commerce): State-level data on personal and gross state product, multiple years. Available online at http://www.bea.doc.gov/bea/regional/data.htm.
- Bureau of Justice Statistics, http://www.ojp.usdoj.gov/bjs (Department of Justice): Crime data since 1960 from the FBI's Uniform Crime Reports, homicide trends, and law enforcement activities. Available online at http://149.101.22.40/dataonline.
- Centers for Medicare and Medicaid Services, http://cms.hhs.gov/researchers/default.asp (formerly HCFA, Department of Health and Human Services): Links to extensive data sets on Medicare, Medicaid, and SCHIP enrollment and spending. Medicare enrollment data available online for the years 1985–2001.
- Energy Information Administration, http://www.eia.doe.gov (Department of Energy): "State Data Energy Report"—Data on state energy sources, consumption, and activities, 1960–1999. Available online at http://www.eia.doe.gov/emeu/sedr.
- National Center for Education Statistics, http://nces.ed.gov (Department of Education): "Digest of Education Statistics"—Comprehensive data set on enrollment, attendance, teacher/student ratios, teacher salaries, student achievement, expenditures, and reforms. Available online for the years 1996–2000.
- National Center for Health Statistics, http://www.cdc.gov/nchs/datawh.htm (Centers for Disease Control): Comprehensive source of state-level data on the health of the U.S. public, including mortality rates, women and children's health, the distribution of CDC expenditures, and environmental and occupational health concerns. Current data available online. Data for previous years available in hard copy.

Associations of State Officials

The National Conference of State Legislatures (NCSL), the National Governor's Association (NGA), the National

Center for State Courts (NCSC), and the Council of State Governments (CSG) do an excellent job of providing information on state politics, political institutions, and, increasingly, information and data on state policy issues and innovations. These associations also provide links to state agencies and other sources of information:

- CSG, www.statesnews.org: The CSG's annual "Book of the States" is a valuable source of information on the states' constitutions, political institutions, and finances.
- NCSC, http://www.ncsconline.org: Through its collaboration with the National Court Statistics Project, the NCSC has compiled an extensive caseload database for all states beginning with the year 1998.
- NCSL, www.ncsl.org: The NCSL offers a wealth of data on the states' legislatures, including information on campaign finance legislation, term limits, partisanship, and legislative compensation. Current data are available online. The NCSL also provides a limited amount of information concerning state policy initiatives across a number of areas.
- NGA, www.nga.org: The NGA offers online information summarizing state policy initiatives, including welfare, child health, the use of tobacco settlement funds, pharmaceutical regulation, and Medicaid programming.

Think Tanks/Research Organizations

Increasingly, some of the best data on state politics and policy are provided by think tanks located in Washington, DC, and elsewhere. Much of these data are online. The following is a sampling of these organizations and the kinds of data they offer:

- The Annie E. Casey Foundation, http://www.aecf.org/kidscount: KIDS COUNT is a project of the Annie E. Casey Foundation and represents a state-by-state effort to track the status of children in the United States. It provides various indicators of child well-being for each state.
- Environmental Defense Fund, www.edf.org: Environmental Defense Scorecard provides state rankings on health hazards from air pollution, air pollution levels, animal waste, and toxic chemicals (http://www.scorecard.org/ranking).
- The Henry K. Kaiser Family Foundation: State Health Facts Online (http://www.statehealthfacts.kff.org) Extensive online source of the latest state-level data on the health and health coverage of the states' citizens and state and federal health policy.
- Heritage Foundation, www.heritage.org: The foundation provides online information on charter schools in each of the states at www.heritage.org/schools. Includes Heritage's thumbnail sketch of each state's charter school legislation, an evaluation of state programs, and a count of charter schools/students in the state.
- Initiative and Referendum Institute, http://www.iandrinstitute.org: The institute offers, both online and in various publications, data on the incidence, nature, and outcomes of initiatives and referendum in each of the states for recent years.
- The National Institute on Money in State Politics: Follow the Money (http://www.followthemoney.org/database/enter.phtml) is an online source of data on campaign finance in the states. Covers the 1998, 2000, and 2002 elections.
- National Network of State Polls, http://www.irss.unc.edu/irss/nnsp/nnspindex.asp: The National Network of State Polls archive of state polls is the largest available collection of state-level data. It is part of the data archive of the Odum Institute for Research in Social Science. The archive contains approximately 60,000 items from more than 530 studies, contributed by 29 survey organizations in 22 states. More than 600,000 respondents contributed to the surveys.
- State Politics and Policy Quarterly Data Resource, http://www.unl.edu/SPPQ: Downloadable data set contains more than 40 state-level variables for multiple years. Set covers demographic, crime and policing, economic, state spending and taxing, and education data.
- State Policy Documentation Project (SPDP), http://www.spdp.org: A joint project of the Center for Law and Social Policy and the Center on Budget and Policy Priorities. SPDP tracked policy choices on TANF cash assistance programs and Medicaid in the 50 states and the District of Columbia from 1998 to 2000. The information presented on this Web site was collected through surveys completed by a key policy advocate in each state, confirmed by state agency staff, and verified against state statute and regulation by SPDP staff.
- The Urban Institute, http://www.urb.org: The Urban Institute's "Assessing the New Federalism" is a major, multiyear effort to assess the devolution of responsibility for social programs to the American states. Two major sources of data are available online—a large, downloadable state database and a case studies series that focuses on the efforts of 13 states. The state database (http://newfederalism.urban.org/nfdb/index.htm) includes more than 400 variables referencing state health and welfare programs, the states' poor population, and the states' economies. It is updated on a regular basis. The state case studies series (http://newfederalism.urban.org/html/state_ focus.html) reports additional data on

13 states—Alabama, California, Colorado, Florida, Massachusetts, Michigan, Minnesota, Mississippi, New Jersey, New York, Texas, Washington, and Wisconsin—in three case study reports. The first two involved site visits by institute researchers in 1996, and the third was compiled from the 1997 "National Survey of America's Families."

State Data Archives

- State Politics and Policy Data Archives—Florida State University, http://www.pubadm.fsu.edu/archives: During the past several years, scholars at the Askew School of Public Administration have maintained an archive on state data from published articles. A dozen downloadable data sets are available from articles published between 1992 and 1997.
- Inter-University Consortium for Political and Social Research (ICPSR), http://www.icpsr.umich.edu/org/index.html: The ICPSR is the largest archive of social science data in the world. Included in its vast holdings are the data files used in numerous published studies throughout the years, historical data on the states, as well as data from various federal agencies.

State Data Links

There are an increasing number of online sites that provide links to various data sources, many of which provide state-level data. The following are examples:

- *State Politics and Policy Quarterly*: In addition to its own data file, the journal also provides links to a dozen or so data sources at http://www.unl.edu/SPPQ/links.html.
- U.S. Bureau of the Census, Census State Data Centers, http://www.census.gov/sdc/www: The Bureau of the Census provides a link to each state's official data centers. These centers are official repositories of the census files for the state.
- University of California at Irvine, Social Science Data Archives, http://data.lib.uci.edu: This site is typical of sites offered at various universities in the United States that provide links to data sources, many of which report state-level data.

Methodological Issues

Although the states provide a unique and valuable setting for systematically studying politics and policy in the United States, there are several methodological challenges to researchers in this field, including the relative tradeoffs between case studies and multistate analysis, the use of significance tests, and issues of ecological inference.

Case Studies, Small-*N* Research, and Multistate Analysis

One of the principal reasons scholars choose to study state politics is its potential to serve as a laboratory of democracy. Given the variation in state politics, institutions, and policies, the states offer researchers a unique opportunity to examine the interactions between these variables in a rigorous fashion and, ultimately, test broader theories of politics and policy. The relative (to case studies and small N studies) advantages of 50 state analyses are obvious. Including all (or most) of the states allows the researcher to observe the often substantial differences in the phenomenon being studied (e.g., legislative professionalism, voter turnout, and educational achievement), analyze the influence of multiple variables, and draw upon statistical techniques to systematically sort their relative impacts. Also, of course, including each of the 50 states in the analysis allows the researcher to generalize study conclusions to the larger "universe" of American states.

In other words, the use of 50 states enhances the potential for analytical leverage. According to King *et al.* (1994), leverage entails "explaining as much as possible with as little as possible" (p. 29). Fifty-state studies attempt to utilize a few variables, usually institutional (e.g., party balance in legislatures or identity of the governor), economic (e.g., unemployment rate or fiscal health), and/or social (e.g., interest group organization and public opinion), to explain a state-level political phenomenon. In doing so, most 50-state studies are attempting to capitalize on the concept of leverage. There has been much debate about the success of 50-state studies in providing elegant explanatory models of state political processes. Some of this concern stems from the small R^2 that these models achieve compared with that obtained in other subdisciplines in political science. Nevertheless, many political methodologists argue that R^2 is not the most important way to measure the effectiveness of one's models.

The obvious advantages aside, 50-state studies often encounter methodological problems, some of which can neither be avoided nor corrected. First, many, perhaps most, multistate studies are cross-sectional, either by design or by necessity. In some instances, data on one or more key variables are simply unavailable. In other instances, the researcher's concern is only with a single point in time (e.g., the 2000 presidential election). In examining only one point in time, researchers can make mistakes in their inferences about the dynamics of the process. Where the necessary data are available, pooled, time series, cross-sectional analysis can and

should be used to more systematically model those processes. Pooled cross-sectional studies capture both the effects of space (variations across the 50 states at any point in time) and the effects of time to provide insights into the dynamic effects of political, social, and economic indicators. Nevertheless, pooled, cross-sectional, time series analysis has its own unique problems regarding modeling of both cross-sectional (state-by-state) variation and time-serial (over time) variation because each of these variations may have unique underlying explanations.

Because 50-state studies invariably rely on secondary data sources over which researchers have little control, these studies are especially susceptible to problems of incomplete and noncomparable data. Indeed, because of missing or invalid data many "50-state" analyses are not 50-state studies at all. Our 2002 analysis of state welfare policy provides an example of how easily this can happen. In examining the political and social correlates of state welfare policies, we, like researchers before us, included measures of the public's ideology and the level of interparty competition. As we proceeded with our analysis, our *N* quickly shrunk to just 44 states. Nebraska was excluded from the analysis because it is, after all, a nonpartisan legislature. Alaska and Hawaii were also excluded because of missing data on Erickson *et al.*'s public opinion variable. Also, for good measure, we excluded Nevada because of concerns about the validity of that state's public opinion score (Nevadans are not liberal despite what their score suggests). Two other states, Wisconsin and Vermont, were excluded from the analysis because they failed to report data on a child care variable that we were considering. Another concern is that states do not necessarily collect the same data in the same way. Criminologists, for instance, have long been skeptical of crime data reported by state and local officials because these data are often manipulated for political and fiscal reasons.

These and other data problems suggest the need to reconsider the role of the case study in state research. Historically, the bulk of research on the states concentrated on a single or small number of states. This approach is criticized on obvious grounds. Limiting the number of states included in the analysis limits the ability of the analyst to generalize his or her findings to other states and to sort out the influence of multiple variables. However, case studies can be particularly helpful in illuminating the richness and detail of state politics. Studies of one state can more accurately capture the nuances and complexities of political conditions. Often, a particular state may be unique in its institutional structure (e.g., Nebraska and its unicameral legislature) or innovative in its policy approach (e.g., Wisconsin and welfare reform) such that it warrants the type of in-depth examination a case study provides.

For example, Dan Smith analyzed roll call votes on three bills in Colorado to discover the determinants of legislative member voting on "counter-majoritarian" legislation. Colorado was an excellent domain for his analysis for several reasons. First, only 24 states allow direct legislation; Colorado has a long history of utilizing this process. Second, Colorado has had some very controversial issues on their ballot (school choice, abortion, term limits, etc.) that have received considerable national attention. Third, three bills introduced during the 1999 legislative session proposed legislation that was in direct opposition to recently passed ballot initiatives by the Colorado electorate. Thus, not only does Smith's analysis help state politics researchers to understand ballot initiatives in Colorado but also it provides a contextual understanding of how ballot initiatives might operate in other states. Studying one state also provides a good opportunity to compare local government performance while holding governmental structure constant. For example, Kevin Smith utilized school districts within the state of Florida to understand the dynamics of school choice.

Small and large *N* analyses are not mutually exclusive. State politics can benefit from combining 50-state analyses with more in-depth analysis of a small number of states. In his investigation of the relationship between state government and state economic performance, Brace utilized case studies of Arizona, Michigan, New York, and Texas to provide a rich description of different styles of state government interventions on the economy. In addition to case studies, Brace provided a comprehensive and concise model of the effect of state institutional, economic, and political characteristics on their economies across the 50 states.

The Question of Statistical Significance

An additional complication to the study of state politics is the question of statistical significance. There has been ongoing discussion regarding the appropriate use of null hypothesis significance testing in political science. Gill argues that the null hypothesis as traditionally taught and used in political science is deeply flawed and widely misunderstood. Although we do not intend to revisit this thoroughly discussed issue, we do think it is important to note how this debate uniquely affects the use of state data.

Despite the warnings of Gill and others, multistate analysis almost invariably proceeds by assessing claims against a null hypothesis (usually of no affect). However, because these studies usually encompass all 50 states, the data represent population, not sample, data. The test of statistical significance is typically taught and used to make inferences about a population using a subset of that population (preferably employing some form of random sampling technique). What is unusual about state data is that they often represent the population. As such, the normal process

of utilizing tests of statistical significance to make inferences to an unknown population parameter does not apply.

There is a counterargument to this logic. State politics and policy data, like much of political science and social science data more generally, are often plagued with measurement error. As such, the estimates that state research creates can still be thought to be estimates of unknown population parameters. With this conceptualization, hypothesis testing may still seem to be useful. Nevertheless, many argue that confidence intervals or, perhaps, more complex estimation techniques such as Bayesian methods are better suited to this type of data.

Ecological Inference

Problems associated with ecological inference frequently impede analysis of state data. According to King (1997), ecological inference "is the process of using aggregate (i.e., "ecological") data to infer discrete individual-level relationships of interest when individual-level data are not available" (p. xv). Often, 50-state studies collect aggregate state-level data on an outcome measure and its covariates and proceed to make inferences to individual-level behavior. In the study of public policy outcomes, ecological inferences are often utilized when recommending public policy action.

For example, in the field of state education policy, there is an extensive debate over the role of vouchers and school choice in improving education. One of the concerns in this literature is the effect that providing a school voucher to an individual to attend a private school has on that particular individual's performance in school. In order to accurately assess this situation, the researcher would need to compare individual student performance for those with vouchers and those without vouchers as well as those who would have attended private school without the voucher and those who only attended a private school because of the voucher. Because of confidentiality of individual student information, obtaining this information is either not legal or impractical. Instead, state politics researchers must rely on aggregate state data to make these inferences. Typically, researchers examine aggregate school-level variables, such as dropout rates, achievement scores, and percentage who attend college, and compare schools with a low proportion of voucher students to schools with a high proportion of voucher students.

There are several ways in which state-level problems with ecological inference can be solved. One way is to engage in the collection of individual-level data. State welfare policy researchers are often interested in making recommendations about which types of welfare programs are most successful at reducing the number of individuals on welfare and/or increasing the number of welfare recipients finding work. One of the ways in which this issue is studied is by using aggregate data. Typically, researchers collect data on welfare outcomes in the 50 states and analyze the effect that political, institutional, economic, and social factors have on welfare outcomes. From these findings, state policy researchers often make inferences to individual-level behaviors. However, using typical statistical methods, aggregate-level data do not provide the type of information necessary to make these individual-level inferences. Instead, researchers could survey former welfare recipients to determine whether they were subject to workforce participation requirements and whether they left welfare and found work. However, there are several limitations to this approach. First, generating a random sample of individuals that includes enough individuals who have been on welfare and have been subject to a program that involved workforce participation requirements is costly and impractical. Additionally, relying on individual survey responses may also introduce error into the model. Because being on welfare and out of work may be embarrassing to certain individuals, survey respondents may provide inaccurate information.

King argues that researchers can utilize information contained in aggregate data to solve some ecological inference problems. One problem with aggregate data is the assumption of constant parameters. In the school choice example, this means that aggregate models assume that voucher students across all types of schools attend college at the same rate. Obviously, characteristics of the individual schools might lead to some schools having high rates of college attendance and some having low rates of college attendance irrespective of the number of voucher students. To address this problem, King recommends specifying a model with random, rather than fixed, coefficients. A second problem is that standard regression models do not provide bounded parameter estimates even though results less than 0 or greater than 1 are not practically possible. A manifestation of this problem in the school choice example would involve predicting that more than 100% of voucher students attend college or fewer than 0% of nonvoucher students attend college. King's solution involves truncating the range of values to reasonable bounds for these coefficients. The reasonable bounds may include the entire range of possible values (0 and 100%) or, utilizing prior knowledge (often through Bayesian techniques) of school choice, the bounds may be limited to plausible values (20 and 80%).

See Also the Following Articles

Census, Varieties and Uses of Data • County and City Data • Ecological Fallacy • Federalism: Local, State, Federal and International Data Sources • Time-Series–Cross-Section Data

Further Reading

Achen, C. H. (1990). What does "explained variance" explain? *Polit. Anal.* **2,** 173–184.

Arceneaux, K. (2002). Direct democracy and the link between public opinion and state abortion policy. *State Politics Policy Q.* **4,** 372–387.

Brace, P. (1993). *State Government and Economic Performance.* Johns Hopkins University Press, Baltimore, MD.

Brace, P., and Hall, M. G. (1995). Studying courts comparatively: The view from the American states. *Polit. Res. Q.* **48,** 5–29.

Chubb, J. E., and Moe, T. M. (1990). *Politics, Markets, and America's Schools.* Brookings Institution, Washington, DC.

Erikson, R. S., Wright, G. C., and McIver, J. P. (1993). *Statehouse Democracy: Public Opinion and Policy in the American States.* Cambridge University Press, New York.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Polit. Res. Q.* **52,** 647–674.

Gill, J. (2001). Whose variance is it anyway? Interpreting empirical models with state-level data. *State Politics Policy Q.* **3,** 318–338.

Hedge, D. M. (1998). *Governance and the Changing American States.* Westview, Boulder, CO.

Hedge, D. M., and Johnson, R. (2002). It takes a village: Social capital and welfare reform. Paper presented at the annual meeting of the American Political Science Association, Boston, MA.

Hedge, D. M., and Scicchitano, M. J. (1994). Regulating in space and time: The case of regulatory federalism. *J. Politics* **56,** 134–153.

King, G. (1990). Stochastic variation: A comment on Lewis–Beck and Skalaban's "The R-Squared." *Polit. Anal.* **2,** 185–200.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press, Princeton, NJ.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton University Press, Princeton, NJ.

Nicholson-Crotty, S., and Meier, K. J. (2002). Size doesn't matter: In defense of single-state studies. *State Politics Policy Q.* **4,** 411–422.

Schneider, M., Teske, P., and Marschall, M. (2000). *Choosing Schools: Consumer Choice and the Quality of American Schools.* Princeton University Press, Princeton, NJ.

Smith, D. A. (2001). Homeward bound? Micro-level legislative responsiveness to ballot initiatives. *State Politics Policy Q.* **1,** 50–61.

Smith, K. (1994). Policy, markets, and bureaucracy: Reexamining school choice. *J. Politics* **56,** 475–491.

Smith, K. B., and Meier, K. J. (1995). *The Case against School Choice: Politics, Markets, and Fools.* Sharpe, Armonk, NY.

Teske, P. (1991). Interests and institutions in state regulation. *Am. J. Polit. Sci.* **35,** 139–154.

Statistical Disclosure Control

Mark Elliot
University of Manchester, Manchester, United Kingdom

Glossary

analytical validity The veridicality of a data set in terms of analytical results produced by its users.
attribution The certain association or certain disassociation of a variable level (attribute) with a particular population unit.
data intruder An individual, group, or organization seeking to identify population units within an anonymized data set.
identification The association of a record within an anonymized data set with a particular population unit.
key variable Information known to a data intruder about a population unit that is also present on an anonymized data set.
reidentification A method of assessing the disclosure risk through matching experiments.
table linkage The combining of multiple smaller tables into larger tables with a higher potential for disclosure.
unique(ness) A record that is distinct from all other records with respect to a set of variable levels within a sample (sample uniqueness) or population (population uniqueness).

Statistical disclosure control concerns preventing the identification of individual population units and/or the disclosure of information about individual population units through statistical processes such as matching identification information to records within anonymized data sets. Research into statistical disclosure control is concerned with the development of statistical and computational methods in three categories: disclosure risk assessment methods, disclosure control methods, and information loss assessment methods. All three present complex and difficult challenges to researchers whose work encompasses diverse academic disciplines, including statistics, mathematics, computer science, psychology, and social policy.

Introduction

The issues of statistical disclosure have become increasingly important with the exponential increase in computing power and the near-universal availability of Internet connectivity. The concept of statistical disclosure could be defined as follows.

The revealing of information about a population unit through the statistical matching of information already known to the revealing agent (or data intruder) with other anonymized information (or target data set) to which the intruder has access, either legitimately or otherwise.

Disclosure is viewed as a potential problem because information that is released in an anonymized form has invariably been collected with assurances of confidentiality by the data gatherer. This is perceived by national statistical agencies as being particularly important for national censuses, where there is a legally enforced obligation for all members of a given population to participate and where a breakdown of trust in the confidentiality of the process might lead to a reduction in the cooperation of the population and therefore undermine the purpose of the census.

Disclosure, Identification, and Attribution

Statistical disclosure has two components: identification and attribution. Identification is the association of a particular record within a set of data with a particular population unit. Attribution is the association or disassociation of a particular attribute with a particular population unit. The relationship between the two is complex and they can take place independently of one another.

Attribute disclosure can occur without identification if it can be inferred from the anonymized data that all

population units who possess a set of attributes $(X_1, \ldots, X_n)$ also possess attribute X_{n+1} (positive attribute disclosure) or if all population units who possess a set of attributes $(X_1, \ldots, X_n)$ do not possess attribute X_{n+1} (negative attribute disclosure). Therefore, if a data intruder has an identification record including attributes $(X_1, \ldots, X_n)$, then X_{n+1} can be attributed or disattributed to the identification record. Conversely, if an intruder already knows all attributes within a record (or cell within a table), then identification can take place without attribution. These two situations encapsulate the disclosure risk problem for aggregate tabular data, where one is more concerned about attribution. Such data are often released as full population data, making verification of attributions simple.

Identification disclosure, as represented in Fig. 1, is usually regarded as paradigmatic of the disclosure risk situation for microdata (that is, files of records of individual population units). The situation in Fig. 1 can best be described as follows: a data intruder has a set of information or identification variables, which identifies a population unit, and a further set, the key or the key variables, which is also present in a target data set. The association of the key for the population unit with that of a record in the target data set leads to the inference of identification and the attribution or disclosure of the target variables for the population unit.

Actual and Perceived Risk

Another key distinction is between actual and perceived risk. Perceived risk is a complex psychosocial process and impacts at three different loci: the data intruder, the data gatherers, and the population.

The perception of disclosure risk by the data intruder clearly affects whether an attempt will be made. However, disclosure may be a secondary goal for an intruder in service of a primary goal, such as embarrassing the data gatherer. In 1999, Elliot and Dale developed an 11-point taxonomic system for assessing the risk that an intruder will attempt an intrusion, based on the perception that the intruder's own goals will be achieved.

The relationship between actual and perceived disclosure risk is nonlinear and attempts to model these processes have been limited. Although some progress has been made, most practical disclosure control and risk assessment research assumes that the probability of an attempt being made is unity and that the consequences of a successful attempt will be catastrophic in terms of loss of public confidence. Even with this substantial simplification, the processes of assessing and controlling disclosure risk are highly complex. Although the research issues have become considerably clearer, no definitive conclusions have been reached. The situation has been further complicated by the demands for measuring the impact of disclosure control on data quality.

Key Selection

One difficult issue in the analysis of disclosure risk is that of key variable selection. This is particularly an issue for microdata files, which can contain hundreds of variables, and thus analyzing all the possible combinations of variables produces a combinatorial explosion and is fcomputationally intractable. Although advances using high-performance computing have enabled a more comprehensive approach, the issue of which combinations of variables to select is still complex. The usual resolution of this issue is the development of "scenarios of attack." These are scenarios that specify the information that an intruder is likely to have access to. Another approach is to set a base key of common variables and then add each other variable in turn to examine its additional impact. This is useful in assessing which variables contribute most to the overall risk of the file.

Whatever decision is made about key variable selection one needs, when engaged in practical disclosure control for microdata, to be aware that it is always possible that the intruder will have access to information that lies outside the specified key.

Disclosure Risk Assessment

The first step in any well-formed disclosure control regime is an appropriate type of disclosure risk assessment. In the early days of disclosure control, measurement of risk was not well understood and proxy measures

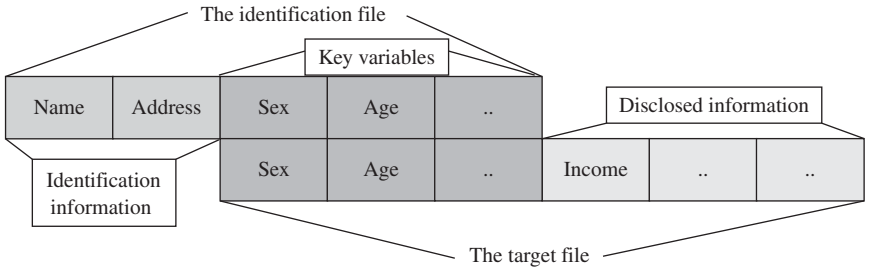


Figure 1 An illustration of the key-variable-matching process leading to disclosure.

such as thresholds on raw frequency counts, were often used. However, measures have become more sophisticated and, although not fully conceptualized, the problems are well defined. Furthermore, increases in computing power have allowed for more in-depth analyses.

Risk Assessment for Microdata

There are essentially two analytical frames for measuring risk for microdata, which both have their uses. The first (file-level risk) focuses broadly on the whole of the data set in question; the second (record-level risk) focuses on individual population units. A third theme in microdata risk assessment is the use of matching or reidentification experiments.

File-Level Metrics

Measuring risk at the file level provides a broad way of summarizing risk for a set of data. This allows the appropriate decisions to be made regarding the level of coding for variables and sampling fraction for a file. The archetypal file-level measure is the level of population uniqueness. This is the proportion of population units that are unique on a given key. The rationale for this metric is as follows: if the intruder knows that an individual is unique in the population on a given set of attributes, and she or he finds a record with such attributes in a target file, then identification disclosure has taken place.

A variation of the population uniqueness metric is the proportion of sample unique records that are also population unique. The argument for this metric is that an intruder with access to a target data set will focus on the records that are unique within the sample on a given key and then attempt to find population units that correspond to those records.

However, these measures are flawed since they presuppose that the intruder information is 100% compatible and there is no obvious systematic way of adjusting them to allow for more realistic assumptions. Furthermore, the measures require that the data holders themselves will have access to population data, which, although a reasonable assumption for full census outputs, will not be the case for outputs from microcensuses or surveys. A final issue with these measures is that they are insensitive to variations in risk across a data set; i.e. they do not identify which records within a data set are risky.

Risky Records

To understand the risky records problem, consider the following situation: a neighbor of yours is a 16-year-old widow; examining a data set you find a record for 16-year-old widow who is coded as residing in your area. You infer that the record is that of your neighbor, because your common-sense demographic knowledge tells you that 16-year-old widows are very rare and therefore there are unlikely to be other such individuals in your locality. Such a record is called a risky record or a special unique.

A crucial aspect of disclosure risk assessment is to identify risky records. Some approaches, such as that of Skinner and Holmes, use statistical models of two-way interactions. Elliot and Manning have used a computational modeling technique that combines the sample uniqueness of combinations of variables within a key. These techniques appear to be successful in identifying sample uniques that are more likely to be unique within the whole population and therefore more risky. The advantage of these approaches is that disclosure control techniques can be targeted on the identified risky records and therefore hopefully do less damage to the data.

Matching

An important method for assessing disclosure risk pioneered by Muller *et al.*, and refined by Winkler, is the use of matching or reidentification experiments. In this method, the risk assessor attempts to match records on two files for which she or he has identification information, in order to establish the empirical probability of matching correctly. These experiments are very useful because they are closer to what an actual data intruder would do if attempting to identify a person in an anonymized data set. However, they have the major disadvantage in being *ad hoc*, i.e., specific to the data sets concerned. It is therefore difficult to draw general conclusions from these experiments.

The *ad hoc* and time-consuming nature of matching experiments indicates that it would be useful if a means could be used that allowed such experiments to be simulated in a more general way. Skinner and Elliot have developed a method for doing just this. By using a bootstrapping-like method called data intrusion simulation, Skinner and Elliot have shown that it is possible to obtain an accurate estimate of the bottom-line matching probability—the probability of a correct match given a match against a sample unique.

Risk Assessment for Tabular Data

Where data are released as a set of tables (or aggregated data) rather than as a sample of microdata, different issues present themselves. In principle, it is possible to view microdata as being derived from a single large table. However, release practices regarding the two forms of output are different and consequently the disclosure risk issues are also different.

(1) Aggregated data are often released as a set of tables (often many in number), whereas microdata are usually released as a single file. Each table will usually contain only a small number of variables, typically one to four.

(2) Aggregated data are often released at much lower levels of geographical detail than microdata.

(3) Aggregated data are often released with counts of whole populations, whereas microdata are usually released as a sample (typically less than 5%).

Many of the metrics that have been discussed here thus far are meaningless for aggregate data. The specific issues that arise are as follows: (1) avoiding cells within tables with low counts, particularly 0's (which can lead to attribute disclosure) and 1's (which give rise to identification), and (2) taking care that tables cannot be easily linked.

Table linkage occurs when information within cells from two or more tables is linked to produce a larger table or fragments of microdata. This can arise directly from uniques or empty cells; however, it need not do so, as Tables I–IV demonstrate. In Table I, there are three individuals who have attribute A and who also have attribute D. In Table II, there are only two individuals who have attribute A and who also have attribute E (and therefore do not have attribute F). Therefore, one can infer that at least one individual has attribute triple (A,D,F). Similarly, one can also infer that at least two individuals have attribute triple (A,C,F) and that eight individuals have the triple (B,C,E) and one individual has the triple (B,D,E).

As the number of tables increases, so does the possibility of linkage. In Table III, there are nine individuals who have attributes E and C. Of these, between seven and nine individuals must have attribute B (from Table II). One can already see from Tables I and II that at least eight individuals have this combination, so one can infer that at least one individual has the attribute triple (B,C,F). In fact, the combination of these three tables allows one to infer the three-dimensional Table IV.

This type of linkage is a simple example of a more general form, which might allow potential data intruders to produce more complex data structures than were originally intended.

A key issue for understanding disclosure risk for tables of counts is the notion of bounds. For each cell within a linked table or one that has been perturbed by disclosure control, there is effectively a feasible maximum (upper bound) and a minimum (lower bound). If the bound's width (upper bound minus lower bound) is too narrow, then it is relatively easy for an intruder to recover the exact counts and perhaps disclosive tables; however, a wide bound width may be achievable only through strong perturbative disclosure control, which may render the tables unusable. The exact calculation of the bounds on a cell is important but difficult and much research has been devoted to developing an efficient algorithm for their calculation.

The foregoing assumes that the aggregate data are in the form of tables of counts. Where cell entries are in the form of magnitudes (such as is frequently the case with business data), a different set of issues arise. One problem is that it might be possible for a firm make quite detailed inferences about their competitors by subtracting themselves from the totals. Estimating whether a cell is sensitive is difficult. Although several different sensitivity rules have been proposed, there is no agreement on the best approach. In 2001, Cox provided a useful discussion of the issues.

Table I Artifical Aggregate Data

	Variable 1	
Variable 2	*A*	*B*
C	4	10
D	3	3

Table II Artifical Aggregate Data

	Variable 1	
Variable 3	*A*	*B*
E	2	11
F	5	2

Table III Artifical Aggregate Data

	Variable 2	
Variable 3	*C*	*D*
E	9	4
F	5	2

Risk Factors

For any given set of data, there are several central factors that affect the overall risk of the file. The major ones are the sampling fraction, the size of key (the number of potential key variables and the level of detail on the key variables), and data divergence.

Sampling Fraction

The impact of sampling fraction is fairly straight forward—universally, the larger the sampling fraction, the larger the risk. There are two elements to this. First, the larger the sample, the higher the probability that any given

Table IV Extended Table Derived from Tables I–III

	Variables 1 and 2			
Variable 3	*A and C*	*A and D*	*B and C*	*B and D*
Definitely E	0	1	8	2
Definitely F	3	1	1	0
E or F	1	1	1	1

population unit is in the sample. This has a linear impact on risk. Second, as records are added to a sample, uniques that are not population uniques may have their statistical twins added to the sample and thus the proportion of sample uniques that are population uniques increases. For census-type aggregate data, the sampling fraction is not usually relevant as such data are most frequently released as 100% population data.

Data Divergence

Data divergence is a term used for inconsistencies between data sets (data–data divergence) or between a data set and the world (data–world divergence). There are many reasons for such inconsistencies, such as errors in response, errors in coding, errors in data entry, and data aging. Data divergence generally reduces the probability of correct attribution or identification.

Size of Key

In general, it is understood that the size of the key, in terms of both the number of variables and the number of possible values that a variable might take, is directly related to the disclosure risk. However, because variables tend to be heavily intercorrelated, the impact of adding an additional variable to a key on any risk metric decreases as the size of the key increases, whereas the divergence effect of adding an additional key is closer to linear. Therefore, larger keys can become difficult for an intruder to handle effectively. For aggregate data, the size of the key is limited by the number of dimensions in the table (or the number of variables in a set of linked tables).

Disclosure Control Methods

Disclosure control methods do not neatly divide into aggregate and tabular types. Most techniques can be applied to both. With aggregate data, it is possible to apply pretabulation disclosure control. This essentially means that disclosure control is applied to the microdata before it is converted to tables. This has the advantage that all tables produced will be consistent in margins and totals, which is not always the case when posttabulation methods are applied.

Recoding

Recoding is a basic tool of disclosure control. The essence of recoding is to take raw data and collapse categories of a variable so that the variable's attributes with low frequencies are conjoined. A typical example is age, where single years of age might be recoded to age groups of 5 or 10 years, or occupation, where specific occupations might be grouped together (usually in accordance with a recognized occupational classification). A special case of recoding is topcoding, in which frequencies tend to become smaller toward the top of the variable's range. Again, age is an obvious example and it is common practice to group all ages above, say, 90 into a single attribute.

The usual practice is to apply all recodes universally across a file, so-called global recoding. However, it is possible to use localized recoding if a variable is highly correlated with location. One example is country of birth. In many populations, people born outside their country of residence are concentrated in urban areas, so providing the detail of country of birth for individuals outside these areas is potentially disclosive. One way to circumnavigate this is to provide two variables: a crude recoded variable covering the whole population and a more detailed variable for those areas in which the nonindigenous population is relatively prevalent.

The major advantage of recoding is that the impact on data quality is visible. Although some secondary analysis will be prevented by the loss of information, there is no possibility of the method introducing analytical invalidity, which some of the perturbative methods risk doing. Against this, global recoding, because it affects the entirety of a data set, can sometimes be of relatively small benefit in terms of risk reduction for the cost in terms of information loss. With localized recoding, this problem is ameliorated, but at the cost of complicating subsequent data analysis.

Cell Suppression

Cell suppression is a method of disclosure control that is specific to aggregate or tabular data (that is, where data are released in the form of tables of counts rather than full individual records or microdata). Suppressing a cell simply means leaving the cell blank. A decision is made that a cell is sensitive because it has a low or zero count and therefore could give rise to attribute disclosure.

A problem with this method is that it is never sufficient to suppress simply the sensitive cell. First, if only cells of a small range of values (e.g., 0 and 1) are suppressed, then the cells that are suppressed are identifiable as having a small range of values. Second, by simple subtraction of known individuals, it may be trivial to recover cell values. Therefore, it will often be necessary to suppress a larger range of cell counts to disguise which cells are the sensitive ones. To give a simple example, in Table V it is desired to suppress the cell (C,A) with a frequency of 1. However, it will be simple to infer its value from the marginal totals and the other cells in the table; this means that it will also be necessary to suppress at least one other cell in each row and column containing the sensitive cell (complementary suppressions). In order to ensure that these cannot be recovered, further cells

may also need to be suppressed (see Table VI). The result is that many cells in any given table may end up being suppressed and the table would have limited analytical value.

Rounding

Rounding methods of disclosure control are applicable to aggregate data. The idea is to disguise the exact frequency count for a cell by rounding every cell in a table to a given base (typically 3, 5, or 10). Thus, Table V could be rounded to base 5, as shown in Table VII.

There are several variants on rounding—the variant shown in Table VII is random rounding, where cell counts can be rounded (either up or down). One problem with this approach is that cell counts may not add up to the subtotals and totals, which is analytically awkward and can also be used to make more accurate inferences about the range of possible counts (or bounds) in the unrounded table, in some instances leading to the recovery of the exact counts. Controlled rounding circumnavigates this by enforcing additivity, ensuring that all counts within a table add up to their respective totals. However, even this may be problematic if multiple overlapping tables are rounded separately, with common marginal cells being rounded to different values.

Table V Table with Sensitive Cell (C,A)

	Variable 1			
Variable 2	*A*	*B*	*C*	*Total*
A	13	22	13	48
B	5	7	19	31
C	**1**	22	7	30
Total	19	51	39	109

Table VI Table V with Sensitive Cell Suppressed and Further Complementary Suppressions Made

	Variable 1			
Variable 2	*A*	*B*	*C*	*Total*
A	13	22	13	48
B	X	7	X	31
C	**X**	22	X	30
Total	19	51	39	109

Table VII Table V Randomly Rounded to Base 5

	Variable 1			
Variable 2	*A*	*B*	*C*	*Total*
A	15	20	15	45
B	5	10	20	30
C	**5**	25	10	30
Total	20	50	40	105

Masking and Blurring

Masking and blurring describe a class of methods for protecting confidentiality by adding noise to the data. With tabular data, this can be achieved through adding and/or subtracting numbers to cells (a process called barnardization). With microdata, values can be manipulated by changing them to adjacent (or simply different) values (postrandomization). Another technique used with microdata, called suppress and reimpute, uses various techniques to overwrite sensitive values with values that are likely in the context of the data set.

Data Swapping

Data swapping is the exchange between records in a data set of values for particular variables (either sensitive/target variables or key/matching variables). If values of potential key variables are swapped, then the intruder's matching task is much more difficult. If sensitive information is swapped, then the intruder cannot be certain if, for a given match, the information disclosed is accurate. Swapping can be performed randomly. However, it is more usual for data to be swapped with records that share the same values on other variables. This leads to less distortion of the data and reduces the risk of inconsistent value combinations.

A particular type of data swapping is termed record swapping. This is the simple exchange of a record with another in a different geographical area, effectively data swapping the records' geographical codes. This can be particularly good for maintaining the overall structure, while disguising a particularly important key variable. As with the more general data swapping, the swaps are often with records that share values on certain variables.

Measuring the Impact on the Data

A great deal of research has been conducted on disclosure risk assessment and disclosure control methodology. However, the impact on the data of disclosure control is much less defined. There have been attempts to obtain measures of information loss to indicate how much information the disclosure control processes have discarded. However, these can be misleading. The mathematically based models do not necessarily correspond to what users want to do with the data. Some user analyses may be unaffected by a given disclosure control regime, whereas others may be badly affected. It is generally assumed that the more complex an analysis, the more likely it is to be affected by the statistical disclosure control methods

applied to the data, although this has not been thoroughly tested.

Alternatives to Disclosure Control

As the demand for data and the technological possibilities for disclosure continue to increase, data holders seek alternative ways to provide information to bona fide data users. The two most explored of these are safe settings and simulated data.

Safe Settings

Another approach to disclosure risk is to look at the users in terms of who is allowed access and how and where the analysis is conducted. The so-called "safe settings" approach looks to set up "data laboratories," where users can run analyses in a controlled environment. This has a major advantage in that the data holders will know exactly who is using the data, what they are doing with it, and so on. It has a severe disadvantage for the users in terms of ease of access and does not encourage exploratory data analysis, which is a key virtue of much microdata release. A variant on this, which is being explored by several statistical agencies, is the virtual safe setting, where the data are stored behind a firewall, but can be indirectly interrogated by the user who can run analyses without accessing the actual data.

Simulated Data

There has been much interest in the possibilities presented by simulated data, which are microdata that have been produced artificially from one or more statistical or computational models. The simulated data are produced through a variety of techniques, such as multiple imputation and mutation algorithms. Sometimes, simulated data are combined with real data and sometimes the data used are fully simulated. Clearly, the disclosure risk with such data is considerably lower. The key question is whether such data can produce analytically equivalent results to real data and most work with simulated data is concerned with assessing this issue. Results are mixed; simulation seems to work well when analyses are simple, but more complex analyses' results can be problematic.

Concluding Remarks

Research into statistical disclosure control has made considerable advances. The parameters of the field are now well defined and there is a wide a range of methods for dealing with issues of confidential data release. As yet, the relationship between disclosure risk and data utility/analytical validity is not well understood. The goals of the field, to be able to specify precisely the disclosure risk for a given data release and to optimize the playoff between data quality and disclosure risk, are still some way from being realized. However, it is possible to look forward to a situation where the data release methodology becomes fully specified.

See Also the Following Articles

Confidentiality and Disclosure Limitation • Missing Data, Problems and Solutions • Risk and Needs Assessments • Statistical Disclosure Control • Statistical/Substantive, Interpretations and Data Limitations

Further Reading

Cox, L. H. (1982). Data swapping: A technique for disclosure control. *J. Statist. Planning Infer.* **6,** 73–85.

Cox, L. H. (2001). Disclosure risk for tabular economic data. In *Confidentiality, Disclosure and Data Access* (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), pp. 167–183. Elsevier, Amsterdam, The Netherlands.

Cox, L. H. (1995). Network models for complimentary cell suppression. *J. Am. Statist. Assoc.* **90,** 1453–1462.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., and Roehrig, S. F. (2001). Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure and Data Access* (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), pp. 135–166. Elsevier, Amsterdam, The Netherlands.

Domingo-Ferrer, J. (ed.) (2002). *Inference Control in Statistical Databases*. LNCS 2316. Springer-Verlag, Berlin, Germany.

Domingo-Ferrer, J., and Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In *Confidentiality, Disclosure and Data Access* (P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, eds.), pp. 111–133. Elsevier, Amsterdam, The Netherlands.

Duncan, G. T., and Lambert, D. (1989). The risk of disclosure from microdata. *J. Bus. Econ. Statist.* **7,** 207–217.

Elliot, M. J., and Dale, A. (1999). *Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk.* Netherlands Official Statistics.

Elliot, M. J., Manning, A. M., and Ford, R. W. (2002). A computational algorithm for handling the special unique problem. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.* **5,** 493–509.

Fienberg, S. E., and Makov, M. M. (1998). Confidentiality, uniqueness and disclosure limitation for categorical data. *J. Official Statist.* **14,** 395–398.

Fischetti, M., and Salazar, J. J. (1999). Models and algorithms for the 2-dimensional cell suppression problem in statistical disclosure control. *Math. Program.* **84,** 283–312.

Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *J. Bus. Econ. Statist.* **6,** 487–500.

Skinner, C. J., and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *J. R. Statist. Soc. Ser. B* **64,** 855–867.

Skinner, C. J., and Holmes, D. J. (1998). Estimating the re-identification risk per record in microdata. *J. Official Statist.* **14,** 361–372.

Willenborg, L., and de Waal, T. (2001). *Elements of Statistical Disclosure Control.* Springer-Verlag, New York.

Winkler, W. E. (1998). Reidentification methods for evaluating the confidentiality of analytically valid microdata. *Res. Official Statist.* **1,** 87–104.

Statistical/Substantive, Interpretations and Data Limitations

Ronda Priest
University of Southern Indiana, Evansville, Indiana, USA

Glossary

alpha (α) A predetermined probability level such that one rejects the null hypothesis.
alternative hypothesis (H_A) The hypothesis that contradicts the null hypothesis, stating that the parameter is a value opposite that of the null.
null hypothesis (H_0) The hypothesis that is tested, stating that the parameter has no effect.
power of the test The probability of rejecting the null hypothesis when it is false.
***P*-value** The calculated probability of the observed data or data more extreme, given that the null hypothesis is true.
type II error The null hypothesis is not rejected even though it is false.

Statistical/Substantive: Data Limitations and Interpretations is the general discussion of statistical significance tests, their overuse and sometimes misuse in the social sciences, and the need for researchers to focus more on practical and theoretical interpretations of their findings.

Introduction

This article discusses the difference between statistical significance and substantive significance. A measure of statistical significance reveals the probability that the value of a statistic could have resulted from chance factors associated with the random manner that observations are gathered for analysis. Substantive significance refers to the extent that a statistic suggests the presence of an important theoretical or practical finding. In evaluating research findings, it is important to clearly distinguish between these two types of significance. Statistics with high statistical significance may have varying degrees of substantive significance, and statistics with low (or even no) statistical significance may or may not have low substantive significance.

Substantive significance has not received as much attention and is study/variable/theory specific. Mentioned only briefly in standard statistics texts, substantive significance (a.k.a. theoretical, practical, meaningful) refers to the theoretical importance or practicality of the result. The social sciences need to pay more attention to reporting the substantive importance of their statistical findings. The goals of social research are reached when the scientific community's main focus is on the size of effects and their theoretical and practical significance utilizing techniques such as descriptive statistics and confidence intervals.

Statistical Significance Testing

Overview of the Process

Tests of statistical significance provide measures of the likelihood that differences among outcomes are actual, and not just due to chance. All significance tests have these basic elements: assumption, null hypothesis (H_0), theoretical or alternative hypothesis (H_A), test statistic (*e.g.*, t), *P*-value, and conclusion. First, regardless of the type of statistical significance test, certain assumptions must be met. These assumptions vary somewhat, but include factors such as: (1) type of data, (2) theoretical form of population distribution, (3) method of sampling (usually random sampling), and (4) sample size.

Second, one develops a null hypothesis about some phenomenon or parameter. For example in its simplest form

$$H_0: \mu = \#$$

A null hypothesis is the opposite of the research hypothesis

$$H_A: \mu \neq \#$$

Next, data are collected that bear on the issue and an appropriate statistic (e.g., proportions, mean, regression coefficient) is calculated that measures the association between an independent and dependent variable. A statistical test of the null hypothesis that determines there is no relationship between the predictor (independent) variable and the dependent variable is then conducted via a test statistic. The test statistic typically involves a point estimate of the parameter to which the hypothesis refers, for example the Student's t test:

$$t = \frac{\bar{x} - \mu_0}{\{s/\sqrt{(n-1)}\}},$$

where $\bar{x}$ is the sample mean, μ_0 is the value of the parameter assumed under the null hypothesis, S is the standard deviation, and n is the sample size.

The test statistic is used to generate a P-value. The P-value is the probability, when H_0 is true, that the test statistic value is at least as contradictory to $H\mathrm{d}_0$ as the value actually observed.

Finally, the P-value is compared a predetermined cutoff value (α) that is usually set at 0.05 or 0.01. When a P-value below α is attained, the results are reported as statistically significant. From this point several interpretations of P often are made.

Interpretation of the *P*-value

Sometimes the P-value is interpreted as the probability that the results obtained were due to chance. For example, small P-values are taken to indicate that the results were not just due to random variation. A large value of P, say for a test that $\mu = \#$, would suggest that the sample mean $\bar{x}$ actually recorded was due to chance, and μ could be assumed to be value assumed under the null hypothesis.

The P-value may also be said to measure the reliability of the result, that is, the probability of getting the same result if the study were repeated. Significant differences are often termed "reliable" under this interpretation. Ironically, while tests of statistical significance measure the reliability of a test statistic, measures of statistical significance sometimes prove to be unreliable themselves.

Finally, P can be treated as the probability that the null hypothesis is true. This interpretation is the most direct, as it addresses the question of interest. These three common interpretations are all incorrect. Small values of P are taken to represent evidence that the null hypothesis is false. However, several studies have demonstrated this is not necessarily so. In reality, a P-value is the probability of the observed data or data more extreme, given that: (1) the null hypothesis is true, (2) the sample size was adequate according to the type of test, and (3) the sampling was done randomly.

Problems with Null Hypothesis Testing

Is the Null Hypothesis Really True?

Most null hypotheses tests, are designed to determine if some parameter equals zero (e.g., $\mu = 0$), a specific value (e.g., $\mu = 100$), or that some sets of parameters are all equal (e.g., $\mu_1 = \mu_2$). Yet rarely is the true value of the parameter exactly equal to the value listed in null hypothesis.

Second, truly random samples are seldom collected in the social sciences. Generally, sampling is done without replacement, resulting in differences in probabilities for each element. Nonresponse and missing data are also key problems in securing true randomness. Suffice it to say, the sampling procedures must be critically examined before conclusions of significance can be made. Replication of findings across studies can help reduce the problems of strict nonrandomness in social sciences (discussed below).

Effects of Sample Size

Finally, the probability of rejecting the null hypothesis when it is false (as it typically is) is greater as the sample size increases. The power of the test is the ability to detect a false null hypothesis. Specifically, for a particular value of the parameter from the alternative hypothesis:

$$\text{Power} = 1 - P(\text{Type II error}),$$

where a Type II error is the failure to reject H_0 when it is, in reality, false.

Holding α constant, power increases as sample size increases. It is generally recommended for a test to have high power. However, in an attempt to raise the power of a test, the P-value can be made as small as one wishes. In other words, to guarantee a rejection of the null hypothesis, one only needs a large enough sample. For example, in testing to see if the mean of a population (μ) is 100. The null hypothesis then is $H_0: \mu = 100$, versus the alternative hypothesis of $H_A: \mu \neq 100$. One might use a Student's t test

$$t = \frac{\bar{x} - 100}{\{s/\sqrt{(n-1)}\}}.$$

Clearly, t can be made arbitrarily large (and the P-value associated with it arbitrarily small) by making either $(\bar{x} - 100)$ or n large enough. As the sample size increases, $\bar{x}$ and S will approximately stabilize at the true parameter values. Hence, a large value of n translates into a large value of t, which generates a small P-value. While minimum sample sizes are strictly adhered to in choosing an appropriate test statistic, maximum sample sizes are not set. Alternatively, α is not usually adjusted according to the sample size. For example, for larger sample sizes (e.g., $n > 1000$), the set significance should be adjusted to a smaller value (e.g., $\alpha = 0.01$ or 0.001).

Even more arbitrary is the tendency for researchers to adhere to a standard set significance of 0.05 or 0.01. P-values less than or equal to α are deemed significant; those greater than α are nonsignificant. This rule promotes the nonsensical distinction between a statistical significant finding if $P = 0.04$, and a nonsignificant finding of $P = 0.06$. Such minor differences are delusive anyway, as they derive from tests whose assumptions often are only approximately met (e.g., a random sample).

Statistical significance is often an imperfect method for determining the reliability of a statistic. However, if the assumption of the tests are met, attention is paid to sample size, and if care is taken to interpret P-values in relation to confidence intervals, statistical significance tests can be the starting point for further analysis, but never the end point.

Moreover, under the best of circumstances, a statistical significance test never indicates the effect size or whether the results are of any theoretical or practical use.

Substantive Significance

Researchers want to answer three basic questions: (1) Is an observed effect real or should it be attributed to chance? (2) If the effect is real, how large is it? and (3) Is the effect large enough to be theoretically interesting or practically useful?

The first question concerning whether chance is a possible explanation for an observed effect is usually addressed with a null hypothesis significance test. As stated earlier, a null hypothesis significance test tells us the probability of obtaining the effect or a more extreme effect if the null hypothesis is true. A significance test does not tell us how large the effect is or whether the effect is important or useful. To do so, one needs to examine substantive (practical/meaningful) significance. Suggestions for analyses regarding the last questions include (1) effect size, (2) confidence intervals, and (3) replication.

Effect Size Measures

The general approach of obtaining some kind of scale-free effect size measure as the indicator of practical meaningfulness or importance has become popular and its use has been widely advocated in recent years. Effect size is generally reported as some proportion of the total variance accounted for by a given effect. Also, termed measures of association strength, R^2 or η^2 are the most common examples. The general formula can be expressed as

$$R^2 = SS_{\text{a source}}/SS_{\text{total}}.$$

The numerator represents the sum of squares from a source of interest. In a model that contains one explanatory factor, the source of interest is obviously the only explanatory variable in the model. For a model containing multiple factors, the source of interest may include either a subset of factors or all the explanatory factors in the model. In the former case, η^2 is used, in the latter, R^2. R^2 is interpreted as the proportion of variance explained by all the factors in the model. Due to an upward bias in R^2 (the maximization property of the least-square principle), numerous "bias correcting" counterparts have been proposed such as ω^2 and ε^2. However, in interpretation, they are conceptually the same.

In interpreting the result, Cohen recommends that an effect size between 0.10 and 0.25 be considered small, one between 0.25 and 0.50 be considered medium, and one above 0.50 be considered large. However, Cohen also stated that these are arbitrary standards and should be interpreted in terms of the theoretical relevance of the effect size for the problem being investigated. Therefore, R^2 and other measures of association suffer some of the same limitations of statistical significance testing. In other words, there is still a question about what constitutes a substantive finding. Too strong of an adherence to effect size conventions tend to cloud the distinction between the size of an effect and substantive significance. In fact, Cohen also stated that effects as small as 1% of explained variance can be theoretically important and large effect sizes may be rather uninteresting, depending on the theoretical hypotheses.

Confidence Intervals

While a significance test provides a way of judging whether a particular value for the parameter is creditable the estimation method of confidence intervals provides a range of the most plausible values for a parameter. A confidence interval is defined as a range of numbers within which the true population parameter is believed to fall. They have the form

$$\text{Point Estimate} \pm (z - \text{score})(\text{standard error}),$$

where the point estimate is the value calculated from the sample (e.g., a mean or proportion).

Generally, testing a hypothesis about a parameter value is not as informative as estimating that parameter

using a confidence interval. Yet researchers have been reluctant to use confidence intervals as widely as tests of significance. A confidence interval contains all the information provided by a significance test, and in addition provides a range of values within which the effect parameter is likely to fall. A confidence interval is just as useful as a null hypothesis significance test for deciding whether chance or sampling variability is an explanation for an observed effect.

Unlike a test statistic, for example a difference among means, point estimates and confidence intervals use the same unit of measurement as the data. This facilitates the interpretation of results and makes inconsequential effects harder to ignore. It should be noted that confidence intervals require the same assumptions as statistical tests, they can suffer from the same violations (e.g., population distribution and nonrandom sampling). However, if we accept significance tests as proof of a finding, why not use the more informative interval estimate?

Replication

All social scientists agree that replication and meta-analysis are better vehicles for accumulating knowledge than statistical significance testing. Multiple studies not only provide replication of specific empirical findings, they also provide a means for eliminating rival hypotheses and establishing the boundary conditions of observed outcomes. The best evidence of replicability involves conducting a true replication of results. Next best alternatives include thoughtfully comparing the effect sizes in a given study with the effect sizes reported in previous research. Finally, internal replicability analysis of sample results can be conducted.

Conclusion

Unfortunately, all too often the primary focus of research is on rejecting a null hypothesis and obtaining a small *P*-value. The focus should be on what the data tell us about the phenomenon under investigation. This is not a new idea. Critics of significance testing have been saying it for years. For example, Yates, a contemporary of Ronald Fisher, observed that the use of the null hypothesis significance test has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, and too little to the estimates of the magnitude of the effects they are investigating. The emphasis on statistical significance testing, and the consideration of the results of each study in isolation, have had the unfortunate result that researchers have often regarded the execution of a test of significance as the principal goal. Too strong an emphasis on null hypothesis significance tests detracts researchers from interpreting the theoretical relevant outcomes of research.

Science is only thoroughly accomplished when researchers focus on theoretical or practical significance. These questions are best addressed with a combination of effect size measures, descriptive statistics, confidence intervals, and replication.

One of the appeals of null hypothesis significance testing is that it appears to be an objective, scientific procedure. On the other hand, deciding whether effects are useful or theoretically significant obviously involves an element of subjectivity. However, researchers have an obligation to make this kind of subjective judgment. No one is in a better position than the researcher who collected and analyzed the data to decide whether the effects are trivial or not. Ironically, researchers make a variety of complex decisions in the design and execution of a research study, but in the name of objectivity they are not expected nor even encouraged to decide whether the effects are practically or theoretically significant. Unfortunately, there are no statistics that directly measure the practical significance of effects. Yet, reporting the results of a significance test alone surely do not.

See Also the Following Articles

Confidence Intervals • Hypothesis Tests and Proofs • Type I and Type II Error

Further Reading

American Psychological Association (1994). *Publication Manual of the American Psychological Association*, 4th Ed. American Psychological Association, Washington, DC.

Cohen, J. (1994). The Earth is round ($p < 0.05$). *Am. Psychologist* **49,** 997–1003.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. Erlbaum, Hillsdale, NJ.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver & Boyd, London.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Education. Psychol. Meas.* **56,** 746–759.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training researchers. *Psychol. Methods* **1,** 115–129.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Education. Res.* **25,** 26–30.

Yates, F. (1951). The influence of "Statistical Methods for Research Workers" on the development of the science of statistics. *J. Am. Statistical Assoc.* **46,** 19–34.

Stone, Richard

Flavio Comim
St. Edmund's College, Cambridge, United Kingdom; and
Universidade Federal do Rio Grande do Sul, Brazil

Glossary

balance sheets A method of clearing up some problems of definition that arise from many different ways of defining national income.
classes of consistency Different requirements used to assess the coherence of national accounts with basic characteristics of economic variables, such as arithmetic identities, accounting identities, knowledge of past behavior and technology, expectations about future behavior and technology, transitional possibilities, all remaining aspects of the problem, and all long-term aims.
equivalent subsets of transactions A process of selection and aggregation of variables based on an operational definition of economic variables that makes measurement possible.
national accounts A system of accounts that provides a display of the basic structure of an economic system reduced to its simplest terms. The main message conveyed by the measurement framework of an accounting structure is of mutual interdependence among its parts. An important feature of this system is consistency in treatment of concepts.
system of multiple classifications A system designed to transform theoretical distinctions into different ways of organizing the balance between subdivision and consolidation.

This article explores John Richard Nicholas Stone's contribution to the creation of national accounts. Its objective is to examine Stone's views on measurement by focusing on his work on national accounts. It is organized into three parts. The first part presents a very brief account of the earlier history of national accounting, centered on Stone's views of it, with subsequent reference to the role of war in the development of national accounts. The second part examines Stone's contribution to the conceptualization and elaboration of a general structure of national accounts and his early views on measurement. Finally, the third part investigates his work on related areas, such as regional accounts and demography, and the significance of these topics to the development of his views on measurement.

Introduction

Throughout his work, John Richard Nicholas Stone expressed a sustained interest in measurement issues. His contribution to economics covered a wide range of areas, such as: empirical analysis of consumer behavior, with emphasis on the measurement of market demand and linear expenditure systems; economic growth, with the construction and analysis of large disaggregated macroeconomic models; economic demography, where he developed new approaches to the issues of population and migration forecasting; and education, where he developed mathematical models of educational planning. His work also covers topics such as index-numbers, time series and cross-section surveys, environmental and sociostatistics, models of financial markets, and optimization problems related to economic growth. Stone's contribution to these topics was shaped by his concern with integrating theory and facts, abstract and practical reasoning, and expressing the results in a measurable form so that they could be used to solve concrete problems.

In no other area did Stone's work achieve more recognition than in the field of national accounts. His Nobel Prize contribution to the development of systems of national accounts has been widely acknowledged by economists and Stone has been called "the father-figure

of national accounting," "the leading international authority in the field of national accounts," "a genius at the National Income and Expenditure," "the leading English political arithmetician, arguably the most distinguished of those who have followed the tradition started by Sir William Petty," and praised for being "one of the discipline's most creative and productive applied economists." In addition, Stone's work on national and social accounting has, according to Deaton, "had a profound influence on the way that measurement is carried out in economics, and his econometric model building has changed the way that economists analyze those measurements." The relevance of Stone's work to measurement in economics should not be underestimated. Even John Maynard Keynes, who was skeptical about measurements in economics, is reported to have exclaimed—about Stone's work on national accounts—that "We are in a new era of joy through statistics."

But Stone did not simply "measure" economics, if by that we understand straightforward "quantification" of phenomena, that is, the direct expression of economic variables in terms of number, degree, or amount. Measurement for Stone meant something more complex, involving not a mere listing of economic quantities or a conceptual mould to which all statistics must conform, but rather a coherent framework for observing behavior. Together with Colin Clark and James Meade, Stone was responsible for changing attitudes toward measurement in economics.

British Empiricists and the Development of National Accounts

The origins of national accounting can be traced back to William Petty's early estimates of capital, income, and expenditure for 17th century England. With Petty, the general boundaries of *Political Arithmetick* were established and an emphasis on measurement, as a way of objectively settling arguments, was advocated. Subsequently, Petty's views on economic measurement were consolidated by Charles Davenant's and Gregory King's work on national income. Stone delves into the history of early British empiricists and argues that there is a strong element of continuity between his work and the work of previous generations. It is interesting to examine his interpretation of the contribution of these early political arithmeticians. Issues of precision and rigor seem to have a minor role in Stone's assessment of their measurements. For instance, Stone does not seem to be bothered by the fact that Petty assumed that all income is spent on consumption or by the highly speculative nature of his estimates. He notes that, "In any case the fact that Petty had poor data and that many of his estimates were guesses and often biased guesses does not invalidate the usefulness of his method." Stone refers here to Petty's innovative use of quantitative data to formulate an argument. Similarly, Stone praises Davenant's contribution to political arithmetic, arguing that "though not original, [it] was invaluable." The lack of precision of Davenant's estimates was less important for Stone's assessment of his contribution than Davenant's "list of principles to be kept in mind by the political arithmetician."

Stone's assessment of Petty and Davenant can be partially explained by his evolutionary attitude toward measurement, according to which, the actual data were relevant not merely for their precision but for the knowledge they transmitted. His appraisal of Gregory King's contribution to measurement of economic aggregates follows the same lines. Stone notes that King "must have relied to a large extent on his power of estimation based on a wide acquaintance with all kinds and conditions of man and a good understanding of how national accounting magnitudes should fit together" and that "Without aiming at strict accuracy one could hope to talk sense." Though Stone conflated the meaning of the words accuracy and precision in his discussion of the British empiricists, we could, perhaps, better understand the message he conveys if we distinguish between them. If we use the word accuracy to express the correctness or truthfulness of something and the word precision to refer to exactness or rigor of estimates, then it could be said that good measurement for Stone seems to be mainly about achieving accurate estimates rather than merely achieving precise ones. This notion is reinforced by his criticism of Lindert and Williamson's attempts at replacing King's *guesses* by more precise statements. Stone's argument is that King's estimates, though imprecise, were accurate and coherent among themselves. It is interesting to note that Stone criticized Petty's and King's estimates of the world population for not being accurate.

From this very brief outline of Stone's assessment of the work of early British empiricists, we can form a first picture of his views on measurement. The first, though obvious, point is that measurement provides a good perspective with which "to talk sense" about the world. The second point is that the validity of measurements should not depend exclusively on their precision but rather on their accuracy, that is, on their ability to express the truthfulness of something. Finally, the third point is that coherence is an important element behind the accuracy of estimates. The above references to the British empiricists seem to suggest that guesses and speculations in measurement are not bad things per se, but that their validity depend on the coherence of the story they are telling. These appear to be the lessons that Stone assimilated from the early British empiricists.

The history of modern national accounting starts with A. L. Bowley's, Colin Clark's and Simon Kuznets's estimations of the principal macroeconomic accounting values for the United Kingdom and United States, respectively, for some years in the 1920s and 1930s. Both have been described as coauthors of "the statistical revolution" that followed the revolution in macroeconomic theory of the 1930s. Production of unofficial statistics boomed during the 1930s, in countries like Hungary, Germany, Sweden, Canada, Australia, Netherlands, United States, and United Kingdom, with important contributions being provided by Ragnar Frisch, J. B. D. Derksen, and Erik Lindahl. During the 1940s, when new institutions were created with the aim of calculating national statistics, many contributors, such as Milton Gilbert, Morris Copeland, George Luxton, E. F. Denison, O. Aukrust, and J. Tinbergen, to quote just a few, helped the development of national accounting. It was not uncommon to find a variety of diverse isolated statistics—such as production levels of food or minerals—being used to characterize the evolution of national income. Indeed, the beginnings of this "statistical revolution" seem to be characterized by statistics being produced without much coherence.

The development of the national accounts was very much influenced by the occurrence of the Second World War (see, for instance, the work of Ruggles). Being a "centrally planed operation," the war involved the coordination of a vast amount of activities within unified systems of control. The assembly of aggregate statistics in a clear, easy-to-read, easy-to-use form was important during the war, when centrally planned operations had to be decided and production and finance had to be restructured in a coordinated manner. In the United Kingdom, the war produced a change in attitude toward the compilation and use of statistics. But what possible uses could national aggregates estimates have during the war? Stone discusses this issue in detail. The budgetary aspect influenced decisions concerning taxation and the evolution of the national debt; it also indirectly concerned demand pressures due to the government's increasing expenditures of war. The allocation of resources in a war economy involved a whole range of physical controls, and financial disequilibria could through exchange rates and price increases invalidate these attempts to control the allocation of resources.

For Stone, in war or peacetimes, any important measure of social control and government intervention depended on the information provided by the description of the economic system as whole offered by national income statistics. Measurement of these aggregates could not necessarily indicate the right sort of policies to pursue, but would, according to him, bring out clearly the nature of the problem and for this reason would be essential for administrative policy purposes under any circumstances. It was also the war that brought James Meade and Richard Stone together and provided the environment for the development of the first stages of national accounting. Stone's involvement with national accounts is abundantly commented and illustrated by himself and by others, such as Meade, Deaton, Pesaran, Harcourt, and Pesaran and Harcourt. The brief outline that follows draws on these references.

When Colin Clark—Stone's greatest influence—left for Australia in 1937, he bequeathed to Stone and his first wife, the statistical supplement called *Trends*, which appeared in the monthly *Industry Illustrated*. In *Trends*, Stone produced statistics and graphs of British economic time series with occasional articles on a topical subject. Partly for this work, when the war became imminent, earlier in 1939, Stone was invited to join the Ministry of Economic Warfare. Later, by suggestion of Austin Robinson, he left the Ministry in August 1940 to join James Meade for Central Economic Information Service of the Offices of the War Cabinet to work on national income accounts. Meade's system, as Deaton puts it, "was a system of empty boxes." Though Meade had developed the original conceptual framework of national accounts, it was through Stone that it was operationalized. They worked together until April 1941, when the new budget split the Economic Information Service into an Economic Section, to which Meade was attached, and a Central Statistical Office (CSO), where Stone became responsible, thanks to Keynes's intervention, for the national accounts. It was during this period that a new conception of measurement emerged from the collaboration between Stone and Meade: a conception that saw the measurement of national income not merely as a quantification of isolated single magnitudes but as a quantification of an *integrated accounting system* in which magnitudes from different sources had to agree. Although Stone and Meade were not the first to create the concepts of national accounts or to estimate national aggregates, they were the first to put forward a notion of measurement based on criteria of systematization and consistency among aggregates. When Meade learned that Stone was leaving the CSO to assume the post of the first director of the Department of Applied Economics at Cambridge, he considered it "a very serious blow" and "a real disaster for Whitehall."

At the end of the war, before assuming his new post, Stone accepted an invitation by Winfield Riefler to visit the Institute for Advanced Study in Princeton. At the Institute he met Alexander Loveday, the head of the Intelligence Department of the League of Nations, who asked him to write a report on national income statistics for the League. The report, published by the United Nations in 1947, was a landmark in the literature of national accounts; it established the measurement standards for the field for many years ahead and contributed to a reconceptualization of measurement in economics.

The Measurement of National Accounts

Richard Stone started his work on national accounts just like everybody else did in the 1930s, by measuring different indices of industrial product and comparing them to other aggregate measures. In 1939, Stone and his first wife Winifred calculated a new measure of mining and manufacturing output using the census of production and employment statistics. They assembled information on movements in production from three different sources and discussed the measurement problems involved in putting together statistics with different degrees of reliability. Their solution to these problems consisted in defining as a standard the material considered most accurate (as for instance, the census and import duty inquiry data) and dismissing those that were extremely difficult to measure (such as the engineering group indices). Assessment of reliability of individual items was an intrinsic part of measurement. It was pursued directly, by confronting different indices that purported to represent the same things, and indirectly, by analyzing their logical significance. In the latter case, the accuracy of adjustments was of crucial importance. Their emphasis on the importance of the plausibility and reliability of data was compatible with what could be called the *Principle of Broad Limits*. Although adjustments were an intrinsic part of measurement, they should be pursued within broad limits, and rough estimates should be used even when they were not very (relatively) precise—as long as they were accurate. In this work, Stone's concern was with the quality of individual series. Expressions such as confidence and creditability were used as a criterion of assessment of these individual and isolated series.

A distinct approach emerged from Stone's first work in collaboration with Meade. In 1941, they put forward the notion of balance sheets as a method of clearing up some problems of definition that arose from many different ways of defining national income. Agreement on definitions, such as those of net investment, direct and indirect taxes, became the first stage in the measurement of national accounts. These definitions should be settled in ways that could be of interest and use to economists, allowing statistical crosschecking. Major problems of definition should be solved in order to assure proper measurement. As a logical consequence, tables would have to balance. Minor problems of definition could be solved through adjustment, since it would be impossible to treat all of them at length. It is not an exaggeration to say that Meade's and Stone's 1941 paper represented a watershed in the literature of national accounts. While the previous focus on measurement had been on the assessment of the reliability of individual series, they put forward the idea of considering all series together in a consistent way. Their tables of national income were not national accounts in the strict sense of the word, but contained the basic measurement principle of coherence which would come to characterize the accounts later.

Interest in studying the variations in economic activity in a peacetime economy led Stone, in 1943, to investigate U.S. figures. He joined the discussions between Simon Kuznets and the U.S. Department of Commerce on the definition and measurement of national accounts, adding his own disagreements to theirs. Stone's discordances with methods used in American estimates revealed that in his view: (i) measurement criteria should not be decided *a priori*. If one were dealing with a situation facing consumers, he would agree with Gilbert's proposition (that market prices were more important than costs to the determination of equilibrium prices). However, if the situation to be analyzed was about productivity, then he would consider factor cost as a more relevant criterion than market prices. Concepts should be rearranged according to the purpose in hand; (ii) measurement of government activity indicated people's attitudes to government. Kuznet's interpretation of government, as if it were a commercial activity, was rejected by Stone as being "a thoroughly inconvenient way of looking at the matter." It might be speculated that wartime experience provided a different perspective on government activities that influenced Stone's disagreement with American (peacetime) position. Finally, Stone criticizes Kuznets's use of maximum errors on the basis that (1) Kuznets does not define the range of his maximum errors and that (2) the reliability of estimates could be better assessed by using concepts analogous to probable or standard errors. Incomparability problems between the American and British figures led Meade and Stone to emphasize what could be called the *Principle of Flexibility* in national accounts, according to which there are many admissible ways of defining the national income, and that there is nothing absolutely right or wrong about any of these definitions. The national income must be measured according to the definition which is most suitable for the particular purpose in view.

A major breakthrough in the measurement of the national accounts came with the memorandum on *Definition and Measurement of the National income and Related Totals* that Stone wrote for the United Nations in 1947. The boundaries of measurement of national aggregates were extended from a simple concern with measuring a collection of single magnitudes to a more complex concern with measuring a coherent system of magnitudes. In the model of an advanced industrial economy that Stone adopted as his working system, he defined three basic forms of economic activity: production, consumption, and accumulation, which became four when transactions with the rest of the world were added. By recording the incomings and outgoings of the basic forms of economic

activity in four accounts, Stone suggested a systematic method of collecting information. This conceptual basis provided a greater uniformity of content in the estimates for different nations. It also permitted Stone to advocate a change of emphasis from the measurement of individual aggregates to the measurement of structures of transactions. The basic forms of economic activity when put together would provide a display of the basic structure of an economic system reduced to its simplest terms. The main message conveyed by the measurement framework of an accounting structure is of mutual interdependence among its parts. An important feature of this system is consistency in treatment of concepts—even for those that are recognized as arbitrary. To a certain extent, all concepts involve a degree of arbitrariness and conventions that are intrinsic to measurement.

Stone suggests that measurement and economic theory should be tailored to each other's needs. On the one hand, the social accounting system should preserve conceptual distinctions that are needed for economic analysis. On the other hand, economic analysis should restate its needs in a terminology that could be measured. These elements could be regulated through a process of selection and aggregation based on the notion of equivalent subsets of transactions. By using this notion, the general meaning of economic variables could be defined operationally in a way that measurement becomes possible. Since there is no uniquely right way of combining accounts, the principle of flexibility could be used to establish the proper combination between subdivision and consolidation for each case. Yet, some types of accounts are never to lose their identity under consolidation. A degree of invariance is needed in order to cope with the complexity of irrelevant features of economic systems. What could be called the *Principle of Invariance* is of particular importance in the cases of nonmonetary transactions and systems of taxation, since it provides a very useful assumption of homogeneity that could be used to measure economies where a market basis exists.

Stone puts forward a more elaborate proposal for multiple classifications in social accounting. According to him, the problem of choosing suitable criteria of classification could *prima facie* have three solutions: (1) the limited solution, where the transactor classification is removed or reduced to a minimum; (2) the solution of Procrustes, where a single classification of transactors is applied; and (3) the proper solution, where we can choose many classifications according to their usefulness. He claimed that only the accounting system could endow the system of classifications with some flexibility. No doubt, flexibility was, for Stone, an essential element of proper measurement. How could he otherwise explain and measure different systems of classification? How could he harmonize a consumers' classification of products with a producers' classification? Or to connect, for instance, government expenditures on health, education, etc., with the different industries that are producing these goods? Thus, a system of multiple classifications could transform theoretical distinctions into different ways of organizing the balance between subdivision and consolidation. Stone summarized these distinctions by claiming that he was "not impressed either by purely theoretical arguments which do not concern themselves with the problems of data collection and processing nor with purely practical arguments which do not concern themselves with theoretically desirable distinctions."

Now, it is important to note that the criteria Stone suggested for measurement were not enough for the solution of all conceptual and practical problems involved in the actual measurement of national accounts. Stone was then forced to recognize that when measurement barriers cannot be overcome, conventions must be introduced, not as a matter of principle but of convenience.

Stone's conceptualization of measurement seems to have evolved from a simple emphasis on the reliability of isolated series, passing through the use of balance sheets to cross-check statistics, to an emphasis on coherence or consistency as the landmark of his contribution to the national accounts. While commentators like Deaton appear to suggest that consistency is defined at a logical value, it seems closer to Stone's views to suggest that consistency was achieved at a conceptual level. When discussing the meaning of consistent projections in multisector models, Stone considers consistency (i) in a restricted sense, in which certain identities must be satisfied and (ii) in a wider sense, "to include consistency with everything we know, everything we expect, and everything we desire to achieve." In order to constrain this very broad definition, Stone formulates seven classes of consistency based on: arithmetic identities, accounting identities, knowledge of past behavior and technology, expectations about future behavior and technology, transitional possibilities, and all remaining aspects of the problem and all long-term aims. Models will meet the different requirements in different degrees according to their uses. Stone does not discuss how a balance could be achieved among these different classes but observes that following theory might be a way of keeping consistency. He comments that,

> *Measurement is important in economics, which is largely a quantitative subject. But left to themselves facts are not very coherent, they need interpretation by the investigator. Theory helps him to do this in a way which makes them consistent with what he knows.*

It is interesting to note how Stone uses *prima facie* the notions of coherence and consistency conveying the same meaning. However, a closer reading reveals that he uses coherence to express a feature of the systems and consistency to express a property of the investigator's assessment

of the systems. As discussed below, this distinction is important to analyze Stone's discussion of adjustment techniques.

The Development of National Accounts

It is also interesting to note that Stone, though recognizing the importance of stepping out of the economic ivory tower to contextualize the accounting framework, pursued a strategy of measurement based on a progressive evolution from the simplest terms of economic structures. In concrete terms, a tension between standardization and extension was manifested in the way Stone handled the inclusion of social and demographic variables into the general framework of national accounts.

Stone did not introduce these characteristics directly as classification criteria in the economic accounts, but rather treated them separately in independent socio-demographic accounts. This was not merely a question of convenience or flexibility of the systems, as it could *prima facie* seem to be the case. This was mainly a consequence of his emphasis on starting measurement from simplest structures and then making the picture more complex by progressively adding new information. Yet, consistency of national accounts, if pursued in a wider sense, would have to refer to a more complex picture of reality. Two of the main areas explored by Stone were regional accounts and demographic accounts.

In broad terms, the regional accounting system developed by Stone was an extension of the system usually applied to individual countries. While a series of new problems arose, such as those concerning the definition of regions or lack of information about interregional flows, old problems related to international accounts became irrelevant, such as the different accounting systems used by different countries and their different units of account. Stone proposed a model where accounts for regions were ordered by type of account and region. When simple national schemes were applied to regions they generated independent relationships that could be used either to make indirect or residual estimates of some of the flows or to adjust an inconsistent set of direct estimates. Yet, the concept of region and of regional structure was not without its difficulties. As Stone put it, "we can apprehend, though we may not be able to formulate precisely, the concept of regional structure." The solution he proposed consisted in finding a geometric analogy between the concept of region and the concept of distance. But because regions should not be distinguished merely by their size, he suggested that this measure should be divided by the region's population. Moreover, to reduce the impact of large transactions he suggested that every transaction should be normalized, and to take into account the correlation between transactions, he transformed them into a set of hypothetical orthogonal transactions. Thus, by using a geometric analogy he was able to give a definition of cluster for measurement purposes. In many cases, regional differences in price levels could not be ignored. Conventional methods, based on pairwise comparisons, could not produce coherent results if more than two countries were involved. The solution would lie in a version of the principle of invariance, consisting in imposing the concept of an average quantity structure. The application of national accounting techniques to regional matrices implied a rejection of the assumption of proportionality between inputs and outputs in favor of linearity, which demanded much more information. As Stone observed, these models suffered from problems of standardization and communication (among their diverse components) that required new mathematical tools, such as matrix algebra. The extension of SNAs would put pressure on further standardization of national accounts.

Stone's work on demographic accounting started with his interest in including education and demographic trends in the Cambridge Growth Model. He repeatedly mentioned the importance of combining demographic and environmental with economic statistics for a proper study of society. However, the statistical approach he used to measure demographic variables followed very closely the approach used to measure economic variables. Stone acknowledged that "The statistical problems encountered in constructing socio-demographic matrices are, *mutatis mutandis*, similar to those encountered in constructing economic matrices." The unit of measurement of demographic accounts, instead of being the pound, was the human individual. The categories used to group units, instead of being industries and products, were based on age groups and within-age-groups based on activities and occupations. Instead of total output, total population; of intermediary product, surviving part of population; of final output, deaths and emigrations; of primary inputs, births and immigrations, and so on. Apart from a straightforward inversion of the role of inputs and outputs in demographic models, the main difference between them and the narrow national accounts consisted in the notion of *life sequences* (such as medical sequence, changes of marital status, and regional migration) used by Stone. They were used to produce a framework of dynamic accounting structure. His work on demographic accounts marked a tendency, pointed out by Johansen, according to which "Stone gradually turned from accounting formulations towards the representation of national accounts in the form of a large 'social accounting matrix' or 'transaction matrix'." It seems that extensions of the SNA, through the inclusion of regional and demographic accounts, would lead to an abandonment of the (formal)

original principle of coherence behind the concept of national accounts. The formal basis of coherence would also be extended to accompany the new bases of measurement. The use of matrix algebra and transaction matrices replaced the former notion of accounts, providing the flexibility needed to assemble complex sets of information. Nevertheless, by using the same standards of measurement, extension of SNAs appeared to involve more, rather than less, standardization of entries.

Final Remarks

Basic data, untailored by human hands, were for Stone, incomplete, inaccurate, inconsistent, and subject to many types of errors. He recognized what he called the "problem of measurement" as pervasive in all sciences. Quite often, according to Stone, the problem of data collection is to organize the large quantities of economic statistics available. In concrete terms, as he argues, "we never start from a *tabula rasa* and the practical problem is not to devise an ideal data collection scheme *ab initio* but to introduce more design and coherence into the one that already exists." Measurement starts then with attention to the existing methods of collection and tabulation and the use of common definitions and classifications and standard dates and intervals. In addition, new types of data can be collected. Boundary regions should be delimitated before what Stone, Champernowne, and Meade called "the practical work of measurement" begins.

The practical work of measurement was a reference to the transformation of direct into consistent measurement. Because initial estimates could not alone produce a consistent and complete set of measures, the boundaries of measurement would have to include the practical work of transforming quantities into empirical facts. Stone lamented that "yet, even nowadays, it is not generally accepted that the task of measurement is unfinished until estimates have been obtained that satisfy the constraints that hold between their true values."

Ultimately, measurement for Stone was about defining systems or stories that could "talk sense about the world." As he remarked: "Just because a theory is coherent there is no particular reason why it should also give a good account of reality." Although it is difficult to distinguish the precise sense in which Stone used the word "coherent" here, there is evidence to suggest that he held a stochastic view of measurement. He did not elaborate on that. Yet, it could be inferred from his writings that he saw measures as estimators of true values of variables. And, it seems, according to him, that this is the best we could aspire to. Accuracy and truth were beyond the potentiality of measurement, because at the end, the best measures would still be conditional to our best subjective impressions. It could be suggested that in the face of the increasing complexity of the measurement of national accounts, the notion of practice (or practical imperative) developed into an important element in Stone's concept of measurement. Thus, it can be said that measurement for Stone was an intrinsic element in this search for integration between theory and practice.

See Also the Following Article

Accounting Measures

Further Reading

Deaton, A. (1987). Stone, John Richard Nicolas. In *The New Palgrave: A Dictionary of Economics* (J. Eatwell, M. Milgate, and P. Newman, eds.). Macmillan, London.

Johansen, L. (1985). Richard Stone's contributions to economics. *Scan. J. Econ.* **87**(1), 4–32.

Pesaran, M. H., and Harcourt, G. C. (2000). Life and work of John Richard Nicolas Stone 1913–1991. *Econ. J.* **110**(461), F146–F165.

Stone, J. R. N. (1986a). Nobel Memorial Lecture 1984, The accounts of society. *J. Appl. Econ.* **1,** 5–28.

Stone, J. R. N. (1971). *Demographic Accounting and Model-Building.* O.E.C.D., Paris.

Stone, J. R. N. (1970). *Mathematical Models of the Economy and Other Essays.* Chapman and Hall, London.

Stone, J. R. N., and Stone, G. ([1961] 1977). *National Income and Expenditure*, 10th Ed. Bowes & Bowes, Lowes.

Stone, J. R. N., Champernowne, D. G., and Meade, J. E. (1941–42). The precision of national income estimates. *Rev. Econ. Studies* **9,** 111–125.

Suzuki, T. (2003). The epistemology of macroeconomic reality: The Keynesian Revolution from an accounting point of view. *Account. Organ. Soc.* **28,** 471–517.

Stratified Sampling Types

Garrett Glasgow
University of California, Santa Barbara, California, USA

Glossary

disproportional allocation The allocation of a sample to strata in a way that does not reflect the actual proportion of the strata in the population; also known as oversampling.
optimal allocation The allocation of a sample to strata in a way that minimizes the variance of estimated population parameters; in some cases known as Neyman allocation.
poststratification The process of allocating the sample to strata after the sample has been drawn.
proportional allocation The allocation of a sample to strata in proportion to the actual proportion of the strata in the population.
quota sampling A nonrandom sampling technique for allocating the sample to strata.
strata A set of mutually exclusive and collectively exhaustive subpopulations of some population; one such unit is a stratum.
stratum weight The proportion of the population a stratum represents.

Stratified random sampling is the process of dividing the sampling units within a population into a set of mutually exclusive and collectively exhaustive groups, known as strata. Simple random samples are then drawn from each strata and combined to form a stratified random sample. Because strata are usually selected to be more homogeneous than the population as a whole, stratification can lead to large improvements in the precision of estimated parameters. Stratified sampling is also used when one or more strata in the population are relatively rare, and an oversample of this subpopulation is desired.

What is Stratified Random Sampling?

Introduction to Stratified Random Sampling

As with any other type of sampling, stratified random sampling is a method by which some observations are drawn from a population in order to make inferences about the population as a whole. Simple random sampling accomplishes this by giving each sampling unit in the population an equal probability of being drawn, and then drawing a sample of the appropriate size. Stratified random sampling begins by dividing the sampling units into a set of mutually exclusive and collectively exhaustive groups. These subpopulations are known as strata. Once the strata have been determined, a simple random sample is independently drawn from each, and the resulting subsamples are weighted and combined to form a stratified random sample.

For example, it may make sense to use stratified random sampling in a survey of individual opinions on tax expenditures on law enforcement, and stratify the sampling units into urban and rural households. It might be expected that individuals in relatively high-crime urban areas will have different opinions on the proper level of tax expenditures on policing than residents of rural areas. In this case, stratification would lead to more precise estimates of public opinion on this issue. Populations can of course be stratified by variables other than geographic location (such as gender or age), and populations can be stratified by multiple variables.

Why Stratified Random Sampling?

There are a number of reasons why a researcher may opt for a stratified random sample. First, stratification may

produce a gain in precision for estimates of characteristics of the whole population. Imagine a population that is heterogeneous, and thus requires a large simple random sample in order to get a reasonably precise estimate of the population characteristics. It may be possible to divide this heterogeneous population into a number of homogeneous strata. If this is the case, precise estimates of the characteristics of each of these homogeneous strata can be obtained with relatively small simple random samples, because the characteristics being measured vary little from one unit to another within each strata. These estimates can then be combined to form a more precise estimate of the population characteristics as a whole. Thus, stratification will be beneficial whenever a heterogeneous population can be divided into relatively homogeneous strata.

Second, a stratified random sample may be superior in terms of cost or administrative convenience. For example, a survey research center may have offices in major cities, and thus create strata based on metropolitan region, with each office responsible for sampling only in its area. This may also reduce the cost per observation in the survey, enabling a larger sample size than a simple random sample.

Third, stratification may be used to address differences in sampling problems across different parts of populations. For example, different sampling methods might be required or desired for different strata. Telephone interviews might be most convenient for rural strata, and face-to-face interviews most convenient for urban strata.

Fourth, estimates of population parameters may be desired for certain subpopulations within the general population. Stratification allows the researcher to place each subpopulation in a stratum and draw an independent sample for this group. This approach is especially useful if one or more strata are relatively rare in the population. In this case, an oversample of this strata can be used to estimate population parameters for these strata.

Properties of Stratified Random Sampling

Notation

In a stratified sample, the population of N sampling units is divided into H exhaustive and mutually exclusive subpopulations, such that $N_1 + N_2 + \cdots + N_H = N$. Once the strata are determined, independent simple random samples are drawn from each strata, denoted by $n_1, n_2, \ldots, n_H$, respectively. The total sample size is denoted n. It is standard to let the subscript h denote the stratum and i denote the individual unit within the stratum. Then y_{ih} denote the value of variable y obtained for unit i in stratum h.

Let μ represent the true population mean of some parameter of interest y in the stratified population, and let μ_h represent the true mean of y in stratum h in this population. Further, let σ_h^2 indicate the true variance of y in stratum h. Note that the formulas for variance presented here ignore the finite population correction (fpc). In finite populations, estimates of the variance are multiplied by the term $1 - (n/N)$, which adjusts the variance downward as the sample becomes a larger fraction of the population. In practice, the fpc is usually ignored if N is large, which is what will be assumed here.

Finally, let W_h refer to the stratum weight of stratum h. This is simply the proportion of the total population contained in the subpopulation defined by stratum h.

$$W_h = \frac{N_h}{N}. \tag{1}$$

Estimation of Population Parameters in Stratified Samples

The purpose of sampling is to obtain information about a population. This is most often done by estimating population parameters using the information in the sample. This section discusses the properties of parameter estimates obtained from stratified random samples. The estimation of a population mean is presented here. Estimates of population totals (population means multiplied by N) and population proportions (population means with ones indicating observations of interest and zeros indicating observations not of interest) follow directly from this discussion.

The true population mean μ for some parameter of interest y in a stratified random sample is simply the sum of the individual stratum means, weighted by their stratum weights:

$$\mu = \sum_{h=1}^{H} W_h \mu_h. \tag{2}$$

An unbiased estimate of this population mean is denoted $\bar{y}_{st}$ (st stands for stratified). Let $\bar{y}_h$ be the sample mean of y in stratum h. Then the population mean is estimated as

$$\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h. \tag{3}$$

This is simply the sum of the individual stratum means, weighted by their stratum weights. This estimate is not necessarily the same as the sample mean $\bar{y}$, which is calculated as

$$\bar{y} = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h. \tag{4}$$

The difference between these estimators is in the weights placed on the strata. In $\bar{y}_{st}$, the sample means from each strata receive their population weights, W_h.

However, the weights placed on the sample means from each strata in $\bar{y}$ will depend on the fraction of the sample allocated to the different strata in the population. Obviously, $\bar{y}_{st}$ and $\bar{y}$ will be equivalent when

$$\frac{n_h}{n} = \frac{N_h}{N} = W_h \tag{5}$$

for all strata.

Setting the sample fraction of each stratum to match the stratum weight is known as stratification with proportional allocation, since strata sample sizes $n_1, n_2, \ldots, n_H$ are proportional to population stratum weights $W_1, W_2, \ldots, W_H$. Proportional allocation is only one way in which to allocate the sample to the strata; other allocation strategies are discussed below.

The variance of the estimate $\bar{y}_{st}$ is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h}. \tag{6}$$

Of course, the true variance of y in stratum h is not generally observed. An unbiased estimate of σ_h^2 is given by

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{ih} - \bar{y}_h)^2, \tag{7}$$

which makes the estimate of the variance of $\bar{y}_{st}$

$$\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{s_h^2}{n_h}. \tag{8}$$

This formula reveals that the variance of $\bar{y}_{st}$ depends only on the variances of the estimates of the individual stratum means $\bar{y}_h$. This means that were it possible to divide a heterogeneous population into perfectly homogeneous strata, the population mean μ could be estimated without error. From this, it is easier to see how stratification into groups that are more homogeneous than the population as a whole can improve the precision of parameter estimates.

Further, this formula reveals that altering the relative sample sizes for the strata (n_h) will alter the variance of $\bar{y}_{st}$. In fact, for any fixed sample size, there exists a stratified random sampling strategy that yields an estimate of μ of minimum variance. This is known as stratification with optimal allocation. This and other sample allocation strategies are discussed in the next section.

Sample Allocation among Strata and the Construction of Strata

An important consideration in stratified random sampling is the allocation of sampling units among the strata. This allocation strategy will depend on the goals of the survey. A common goal in stratified random sampling is to improve the precision of estimates of population parameters. If properly used, stratification nearly always results in more precise estimates of population parameters than those obtained from a comparable simple random sample. However, if strata sample sizes are poorly chosen, stratified random sampling may produce samples that yield less precise estimates of population parameters than simple random sampling. Another common goal in stratified random sampling is to obtain precise parameter estimates for a subpopulation of interest. Three sample allocation strategies are discussed in the following sections: proportional allocation, optimal allocation, and disproportional allocation; the construction of strata is also considered.

Proportional Allocation

Proportional allocation sets the sample size in each stratum equal to be proportional to the number of sampling units in that stratum. That is, $n_h/n = W_h$. Proportional allocation yields a self weighted sample (no additional weighting is required to estimate unbiased population parameters). For example, $\bar{y}_{st} = \bar{y}$, as previously discussed. This was regarded as an important advantage in the past, but modern computational power makes this less of a concern.

Proportional allocation will yield population parameter estimates at least as precise as those obtained from simple random sampling. Depending on the differences between the strata means, the gain in precision from stratified random sampling can be vary large, with gains increasing as the differences between the strata means increase. Proportional allocation is useful if precise estimates are desired for the larger strata in the population, as large sample sizes are allocated to the large strata.

However, proportional allocation often will not produce the most precise parameter estimates possible. The precision of parameter estimates within each stratum is determined by the sample size, not the ratio of the sample size to the population size. Thus, the precision of the estimates can often be improved by allocating more of the sample to the smaller strata. This can greatly improve the precision of the estimates from the smaller strata without greatly reducing the precision of the estimates from the larger strata, improving the overall precision of the population parameter estimates.

Optimal Allocation

As previously discussed, for a fixed sample size there exists an allocation of sample sizes across strata that minimizes the variance of estimated population parameters. This allocation is known as optimal allocation.

Optimal allocation will yield population parameters estimates at least as precise as those obtained from a simple random sample of the same size, and usually

these estimates are much more precise. Further, optimal allocation often yields parameter estimates that are more precise than a stratified random sample of the same size that relies on proportional allocation, although the gain in precision here is usually less than that realized by switching from a simple random sampling approach to a stratified random sampling approach. The gain in precision for an optimal allocation over a proportional allocation will depend on the strata standard deviations, with gains increasing as the differences between the strata standard deviations increase.

For a fixed sample size n, the sample size in stratum h under optimal allocation is given by

$$n_h = n\frac{W_h\sigma_h}{\sum_{k=1}^{H} W_k\sigma_k} = n\frac{N_h\sigma_h/N}{\sum_{k=1}^{H} N_k\sigma_k/N} = n\frac{N_h\sigma_h}{\sum_{k=1}^{H} N_k\sigma_k}. \quad (9)$$

Each strata is sampled in proportion to the product of the standard deviation of the parameter of interest and the fraction of the population the stratum represents. Strata that are more heterogeneous are sampled more heavily in order to reduce the variance of parameter estimates from those strata, thus reducing the variance of estimates of population parameters. This allocation is also known as Neyman allocation.

This allocation strategy assumes that the costs of sampling are equal across strata. If this is not the case, a slightly more complicated optimization problem must be solved. Optimal allocations when sampling costs differ across strata can seek to minimize the variance of population parameter estimates for a fixed cost, or minimize costs for a fixed variance. The optimization problem is the same in either case.

Let c_h be the cost of obtaining one observation from stratum h. Thus, the entire cost of the survey will be $C = c_0 + \sum_{h=1}^{H} n_h c_h$, where c_0 represents any overhead cost for the survey. Under this cost constraint, the sample size in stratum h under optimal allocation is given by

$$n_h = n\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{k=1}^{H} N_k\sigma_k/\sqrt{c_k}}. \quad (10)$$

Each stratum is sampled in proportion to the product of the standard deviation of the parameter of interest and the fraction of the population the stratum represents, and in inverse proportion to the square root of the sampling cost in that stratum. In general, larger fractions of the sample will be allocated to strata that are relatively large, relatively heterogeneous, and less costly to sample from. Note that when sampling costs are identical across strata, the cost terms cancel out, and optimal allocation for a fixed cost becomes equivalent to optimal allocation for a fixed sample size.

Disproportional Allocation

In some cases, the sample may be deliberately allocated across strata in a disproportionate way (meaning the sample size in each stratum is not proportional to the number of sampling units in that stratum). Disproportionate allocation is often used if a particular stratum is of special interest but is too small a proportion of the population for reliable statistical inference without increasing the number of observations from this stratum. This is also known as oversampling, because observations from this stratum are overrepresented in the sample relative to its stratum weight. For example, a study might seek to compare the opinions of World War II veterans to the opinions of the U.S. population as a whole. As most sampling strategies would likely end up with very few World War II veterans, a stratified random sample with disproportionate allocation might be used, placing World War II veterans (or a subpopulation likely to contain a high proportion of World War II veterans) in one stratum, and oversampling this stratum to ensure enough World War II veterans in the sample for a meaningful comparison.

Oversampling will not bias estimates of population parameters as long as the appropriate strata weights are used. However, poorly selected strata sample sizes can result in less precise estimates of population parameters than even simple random sampling. Thus, disproportionate sampling is generally used only to gather a large sample on a particular stratum of interest, and not as a strategy to improve the precision of estimates on overall population parameters.

The Construction of Strata

How many and what type of strata to define in a stratified random sample will be determined by the goals of the study. In some cases, the strata will be determined by a desire to examine a particular subpopulation of interest, such as in the study described in the preceding section. In other cases, the strata will be defined in such a way as to minimize the variance of estimates of population parameters. It is this case that is considered here.

In order to gain the most from stratification, strata should be selected so that the differences between strata means are as large as possible and so that each stratum is as homogeneous as possible. Theoretically, increasing the number of strata will improve the precision of estimates of population parameters. However, in most practical applications, little improvement is seen beyond $H = 6$ or so. Further, additional strata may add to the cost of the study. Thus, most stratification strategies rely on a relatively small number of strata.

Once the number of strata is determined, boundaries between the strata must be selected. One method that has been found to be practical and efficient is known as the

cumulative square root of the frequency method. This method works when the stratifying variable y can be organized into ordered categories. The square root of the frequency of each category of y is calculated, and the cumulative distribution of these terms is examined. This cumulative distribution is denoted cum $\sqrt{f(y)}$, and the dividing points between strata are then selected to create equal intervals on the cum $\sqrt{f(y)}$ scale. For instance, if three strata were desired, and the square root of the cumulative frequency of the last category of y was 60, the dividing points between the three strata would be the values of cum $\sqrt{f(y)}$ closest to 20 and 40. In cases where the frequency distribution of y is not available, the frequency distribution of a variable highly correlated with y can be substituted. This method has been found to perform well for a variety of distributions of y.

Obviously, this method will not work for many variables that could be used for stratification (such as geographic variables). Like the determination of the number of strata, the construction of strata boundaries will often depend on judgement, trial and error, and knowledge of the characteristics of the sampling problem at hand in order to define relatively homogeneous strata with means that differ from each other.

Other Stratification Issues

Poststratification

In some instances, it is inconvenient or impossible to divide the population into strata before sampling. This is most often the case when the variable used to stratify the population can only be observed after sampling. For example, it will be impossible to stratify a public opinion poll by gender if the sample is drawn using randomly dialed telephone numbers, as it will be impossible to determine the gender of respondents until they are in the sample. Poststratification is often used when a simple random sample does not reflect the distribution of some known variable in the population.

In this case, a simple random sample is conducted, and then observations are placed in strata. Estimates of population parameters are carried out as with a stratified random sample. For example, a population mean would be estimated as the sum of the individual stratum means, weighted by their stratum weights. This estimate will be similar to that obtained from a stratified random sample with proportional allocation as long as the sample size is large in each stratum (generally larger than 20). Note that poststratification should not be used if the stratum weights are not known or cannot be closely approximated, because inaccurate stratum weights can lead to very poor estimates of population parameters.

The poststratification estimator will not have the same variance as an estimate obtained from a stratified sample, as the sample size in each stratum is no longer fixed but is a random variable. The approximate variance of $\bar{y}_{\text{pst}}$ (pst stands for poststratified) is given by

$$\hat{V}(\bar{y}_{\text{pst}}) = \frac{1}{n}\sum_{h=1}^{H} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{H}(1 - W_h)s_h^2. \qquad (11)$$

The first term in this equation is equivalent to the variance of the estimate of $\bar{y}_{\text{st}}$ that would be obtained from a stratified random sample under proportional allocation. The second term gives the increase in the variance of $\bar{y}_{\text{pst}}$ due to poststratification. This second term is always nonnegative, although it will be small as long as n is large.

Double Sampling for Stratification

The preceding discussion assumes that the strata weights W_h are known constants before sampling begins. However, in many instances the strata weights will be unknown. In some cases, the relative size of the strata may be determined through non-sample information (voter registration rolls, census information, etc.). For cases in which this is not true, preliminary information will have to be gathered from the population to construct strata before stratified random sampling can begin.

Double sampling (also known as two-phase sampling) is the process of first gathering preliminary information on which to base stratification, and then drawing a stratified random sample from this first sample, using information obtained in the first sample to determine the appropriate strata weights. This strategy will be viable in cases in which observations on the variables on which to base stratification are easy to obtain. The phase 1 sample is generally a large simple random sample used to estimate the strata weights. The phase 2 sample gathers the information of central interest to the study by drawing a smaller stratified random sample from the elements first drawn in the phase 1 sample.

Stratum weights for each stratum h are estimated by

$$\hat{W}_h = \frac{n'_h}{n'}, \qquad (12)$$

where n' is the sample size of the phase 1 sample, and n'_h is the number of observations falling into stratum h in the phase 1 sample. $\hat{W}_h$ is an unbiased estimator of W_h, assuming the phase 1 sample is random.

The phase 2 sample then randomly draws the appropriate number of elements (n_h) from the n'_h elements identified as belonging to stratum h. Estimates of population parameters are then obtained from the phase 2 sample in a straightforward way, replacing W_h with $\hat{W}_h$.

The variance of population parameter estimates using the phase 2 sample will not be the same as that of parameter estimates from a stratified random sample for which stratum weights are known. This is because stratum weights are no longer fixed, but are random variables. The approximate variance of $\bar{y}'_{st}$ (an estimate of μ using the phase 2 sample) is given by

$$\hat{V}(\bar{y}'_{st}) = \sum_{h=1}^{H} \hat{W}_h^2 \frac{s_h^2}{n_h} + \hat{W}_h \frac{(\bar{y}_h - \bar{y}_{st})^2}{n'}. \qquad (13)$$

The first term in this equation is identical to the variance of the estimate of μ with a stratified random sample that did not employ double sampling, except $\hat{W}_h$ replaces W_h. The second term gives the increase in variance due to the estimation of the stratum weights. Although this second term grows larger as the differences in the stratum means increase, the advantages of stratification over simple random sampling also increase as the differences between the stratum means increase. Thus, even though double sampling for stratification increases the variance of estimated population parameters, it may still result in large increases in precision over simple random sampling.

Quota Sampling

Quota sampling is the nonprobability equivalent of stratified random sampling. Like stratified random sampling, the population is first divided into strata. For a fixed sample size n, the n_h required in each stratum for proportional stratification is determined. A quota is set for each stratum of n_h observations, and the researcher continues sampling until the quota for each stratum is filled. For instance, if a population is known to be 70% men and 30% women, a survey of 100 people using quota sampling would ensure that 70 of the interviews were with men and 30 were with women. Subjects for the interviews are selected based on convenience and the judgement of the interviewer.

Quota sampling is generally less desirable than stratified random sampling for two reasons. First, because the selection of sampling units is non-random, the usual sampling error formulas (such as the estimation of variances on our estimated parameters) cannot be applied to the results of quota samples with any confidence. Second, because the observations included in a quota sample are selected nonrandomly, this may introduce bias into the sample that a random sample would not. That is, while a quota sample will be representative of the population on the variables used to define the strata, it may not be on other variables. Randomized samples will most likely be more representative on uncontrolled factors than an equivalent quota sample.

See Also the Following Articles

Observational Studies • Randomization • Sample Size • Surveys

Further Reading

Cochran, W. G. (1977). *Sampling Techniques,* 3rd Ed. John Wiley & Sons, New York.

Kish, L. (1995). *Survey Sampling.* Wiley-Interscience, New York.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.* **97,** 558–606.

Scheaffer, R. L., Mendenhall, W., and Ott, L. (1995). *Elementary Survey Sampling,* 5th Ed. Duxbury Press, Belmont, CA.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys.* John Wiley & Sons, New York.

Structural Equation Models

David Knoke
University of Minnesota, Minneapolis, Minnesota, USA

Glossary

confirmatory factor analysis A multivariate equation model with one or more unobserved common factors describing or explaining the relationships among empirical measures.
random error Unpredictable error resulting in normally distributed variation around a measure's true value.
reliability The extent to which different operationalizations of the same concept produce consistent results.
structural equation model A multivariate equation model combining relations among unobserved constructs with links to empirical indicators.
validity The degree to which the operationalizations of a variable accurately reflect the concept that they purport to measure.

Structural equation models (SEM) are a family of analysis methods that represent translations of a series of hypothesized cause–effect relationships among variables, for making quantitative estimates of model parameters and their standard errors, for assessing the overall fit of a model to data, and for determining the equivalences of model parameters across several samples. The techniques for analyzing multivariate relationships among systems of equations build directly on multiple regression, exploratory factor analysis, and path models. Although SEM methods can be applied to complex problems, such as nonrecursive models that estimate reciprocal causal effects, space constraints allow only a basic exposition.

Reliability and Validity Issues

An important advantage of structural equation models (SEM) is their capacity to combine empirical observations with relations among unobserved constructs into a single integrated system. Measurement theory seeks to represent a latent (unobserved) construct with one or more observable indicators (operational measures or variables) that accurately capture a theoretically intended concept. Two desirable properties of any empirical measure are high levels of reliability and validity. Reliability indicates the extent to which different operationalizations of the same concept produce consistent results. Reliability refers to the replication of measurement results under the same conditions; a perfectly reliable instrument must generate identical scores when the re-measurement conditions are unchanged. Alternative or multiple measures are reliable indicators of the same construct to the extent that they correlate highly. Validity is the degree to which the operationalizations of a variable accurately reflect the concept that they purport to measure. Many validity issues concern how well or poorly a particular instrument, whether consisting of a single or multiple empirical indicators, represents its intended latent concept. To be valid, a measure must demonstrate at least moderate reliability. In the extreme, if a measure has zero reliability, its validity would be attenuated relative to a more reliable measure. Multiple indictors may vary in their validity as measures of the unobserved concept they are intended to measure. Some measures may be very reliable but not valid; that is, an instrument might very precisely measure a particular phenomenon, yet be invalid for some purposes. For example, individual height can be very reliably measured, yet is worthless as an indicator of a person's physical health. A multiple-item health battery is less reliably measured, yet is far more valid for measuring physical health. Unfortunately, researchers never obtain perfect measurements in the real world; every empirical measure is subject to some degree of measurement error. Measurement theory is therefore also a theory about how to estimate magnitudes and sources of errors in empirical observations.

Measurement reliability assumes random errors. If random error occurs when a measure is repeated several times on the same cases under the same conditions, then the resulting variations in scores form a normal distribution about the measure's true value. The standard error of that distribution represents the magnitude of the measurement error: the larger the standard error, the lower the measure's reliability. In classical test theory, the observed score (X) of respondent i on a measuring instrument (such as an aptitude test score or a survey item) arises from two hypothetical unobservable sources: the respondent's "true score" and an error component:

$$\text{True}_i \longrightarrow X_i \longleftarrow \text{Error}_i$$
$$X_i = T_i + \varepsilon_i.$$

On the assumption of random error, the error term is assumed to be uncorrelated with the true score. Both sources make unique contributions to the observed score variance in a population: $\sigma_X^2 = \sigma_T^2 + \sigma_\varepsilon^2$. The ratio of true score to observed score variances is defined as the reliability of measure X:

$$\rho_X = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2}.$$

This formula demonstrates that reliability ranges between 0 and 1: if the entire observed variance is error, $\rho_X = 0$; but if no random error exists, then $\rho_X = 1$. Rearranging the reliability formula also reveals that the true score variance equals the observed score variance times the reliability: $\sigma_T^2 = \rho_X \sigma_X^2$. Similarly, for two parallel measures (i.e., items having equal variances), the true score variance can be estimated as the product of their correlation ($\rho_{X_1X_2}$) and the variance of either measure; that is, $\sigma_T^2 = \rho_{X_1X_2}\sigma_X^2$. Hence, reliability equals the correlation of two parallel measures, $\rho_X = \rho_{X_1X_2}$, while the correlation between a true score and its indicator equals the square root of the reliability: $\rho_{TX_1} = \sqrt{\rho_X}$. The measurement theory principles summarized in this section are encompassed within structural equation models and are used in the next section on the confirmatory factor analytic approach to modeling the relationships between observed indicators and latent constructs.

Confirmatory Factor Analysis

Factor analysis refers to a family of statistical methods that represents the relationships among a set of observed variables in terms of a hypothesized smaller number of latent constructs, or common factors. The common factors presumably generate the observed variables' covariations (or correlations, if all measures are standardized with zero means and unit variances). In confirmatory factor analysis (CFA), a researcher posits an *a priori* theoretical measurement model to describe or explain the relationship between the underlying common factors and the empirical measures. Then the analyst uses statistical fit criteria to assess the degree to which the sample data are consistent with the posited model, that is, to ask whether the results confirm the hypothesized model.

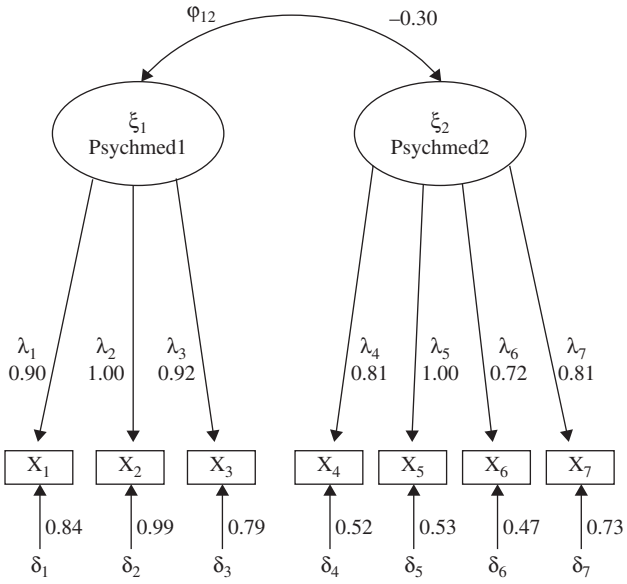


Figure 1 A two-factor confirmatory factor analysis model with seven psychiatric medicine indicators.

Figure 1 hypothesizes that observed measures of three harmful and four beneficial effects of psychiatric medicines (X_1 to X_7) load on separate but correlated latent factors (ξ_1 and ξ_2), labeled Psychmed1 and Psychmed2. Data for the estimates are from 1070 respondents in the 1998 General Social Survey. The seven λ_i are the factor loadings of each observed variable on the two common factors, and the seven δ_i are the observed variables' unique error terms. This diagram implies that the latent constructs are responsible for the covariation among the observed variables. Each observed score is a linear combination of its shared unobserved factor plus its unique error term. These relationships can also be seen by writing the implied measurement equation for the first and seventh indicators: $X_1 = \lambda_1\xi_1 + \delta_1$ and $X_7 = \lambda_7\xi_2 + \delta_7$. Note the similarity of each factor analytic equation to classical test theory's representation of an observed score as a sum of a true score plus an error term.

Figure 1 assumes that all seven error terms are uncorrelated with both factors and among themselves (although alternative models allow such specifications). Hence, the only sources of an indicator's variance are its common factor ξ and its unique error term,

$$\sigma_{X_i}^2 = \lambda_i^2\sigma_{\xi_k}^2 + \Theta_{\delta_i}^2,$$

where $\Theta^2_{\delta_i}$ signifies the variance of the error in X_i. Because ξ_k is unobserved, its variance is unknown and because it is unknown, it may be assumed to be a standardized variable with a variance equal to 1.0. Therefore,

$$\sigma^2_{X_i} = \lambda^2_i + \Theta^2_{\delta_i}.$$

Again, note that this formula closely resembles the classical test theory in which the variance of a measure equals the sum of two components—the true score variance plus the error variance. When both components are standardized, their sum must equal 1.0. A CFA model exhibits another similarity to the classical test theory. The reliability of indicator X_i is defined as the squared correlation between a factor and the indicator (if that indicator loads on only one factor). This value is the proportion of variation in X_i that is statistically "explained" by the common factor (the "true score" in classical test theory) that it purports to measure, $\rho_{X_i} = \rho^2_{\xi X_i} = \lambda^2_i$. Finally, the covariation between any two indicators in a multiple-factor model is the expected value of the product of their two factor loadings times the correlation between the factors. Because the error terms are uncorrelated with the factor and with each other, this simplifies to $\sigma_{X_i X_j} = \lambda_i \lambda_j \varphi^2_{\xi_k \xi_l}$.

As noted above, an unobserved common factor has no definite scale, meaning that both the origin and the unit of measurement are arbitrary. Researchers usually fix a factor's origin by assuming it has a mean of zero. The measurement unit must be scaled by one of two ways: (1) fixing the unobserved factor's variance to unity; or (2) forcing the factor loading of one indicator (λ_i), called the reference indicator, to take a specific value (typically by setting it equal to 1.00). This latter procedure forces the factor's true score variance to equal the reliable portion of the reference indicator's variance.

The CFA example for psychiatric medicine effects in Fig. 1 illustrates the second technique for setting the two factor scales by constraining the factor loadings of the X_2 and X_5 indicators equal to 1.00. Although the seven estimated loadings are all positive, the two factors have a negative covariation (−0.30). This inverse relationship is not surprising, given the substantive wordings of the seven GSS items, emphasizing harmful or beneficial effects, respectively. Because respondents generally do not regard psychiatric medicines as simultaneously harmful and beneficial, a negative covariation occurs between the latent constructs represented by these two sets of empirical indicators.

CFA solutions can represent relationships in both unstandardized and standardized forms. Because a structural equation model consists of both structural and measurement levels of analysis, standardization may be done separately at each level: (1) the standardized solution scales the factors to have standard deviations of one, but leaves the observed variables in their original metrics, and (2) the completely standardized solution transforms the standard deviations of both latent and observed variables to unity. Figure 2 displays the completely standardized solution for the two-factor psychiatric medicine model. The correlation between the two latent factors is −0.57, indicating that they share 32.5% of their variation [$r^2 = (-0.57)(-0.57) = 0.325$]. Unlike the factor model in Fig. 1, the completely standardized solution does not require constraining any indicators to have loadings equal to one. Hence, their magnitudes can easily be compared to assess the indicators' relative importance. Further, in both standardizations, the sum of a squared factor loading plus its square error term equals 1.00, showing that all the variation in an observed indicator is determined by these two sources. For example, the first indicator in Fig. 2 has a standardized factor loading of 0.63 and an error term of 0.78; the sum of their squared values is $(0.63)^2 + (0.78)^2 = 0.397 + 0.608 = 1.00$, within rounding.

Assessing Model Fit to Data

Parameter significance and overall correspondence between the data and the model's parameters are two important concerns of model testing. SEM computer programs estimate standard errors for all free parameters in confirmatory factor analysis or structural equation models. Thus, analysts can test the null hypothesis that a particular parameter is zero in the population, using

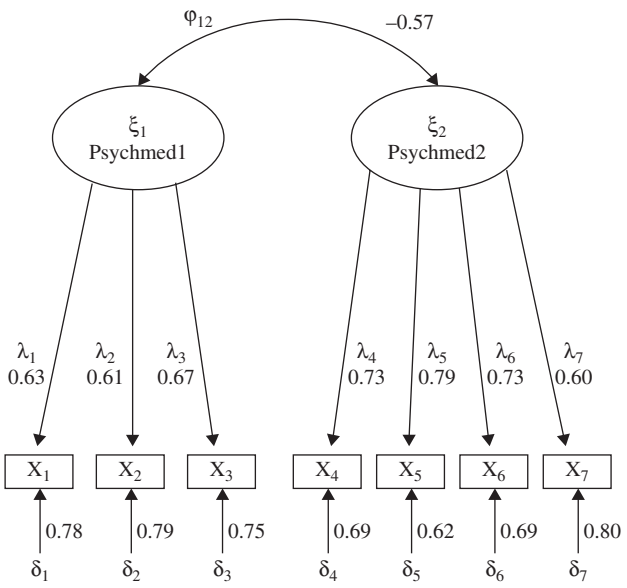


Figure 2 Completely standardized solution for a two-factor confirmatory factor analysis model with seven psychiatric medicine indicators.

appropriate one- or two-tailed t-tests or Z-tests, depending on sample size. All the factor loadings, residuals, variances, and covariances in the CFA model in Fig. 1 are significant at $p < 0.05$. However, testing the significance of individual parameters cannot reveal whether the model as a whole fits the sample data.

Statistical tests for overall model fit involve a comparison of two variance-covariance matrices: (1) the observed matrix ($\mathbf{S}$) of covariances among the K empirical indicators in the sample data; and (2) the expected matrix [$\mathbf{\Sigma}(\boldsymbol{\theta})$] of covariances among the same K indicators, computed from the model's estimated parameters ($\boldsymbol{\theta}$). An SEM program fits a model to the data by minimizing a fit function $F[\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta})]$. (Iterative maximum likelihood estimation is the default procedure of most programs, but alternative methods may be more appropriate for some model specifications, such as generalized least squares or weighted least squares.) The fit function involves discrepancies between the observed and predicted matrices: $F[\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta})] = \ln|\mathbf{\Sigma}| - \ln|\mathbf{S}| + \mathrm{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - p$; where $|\mathbf{\Sigma}|$ and $|\mathbf{S}|$ are determinants of each matrix, tr indicates "trace," the sum of the diagonal elements of the matrix, and p is the number of observed variables in the model. The fit function is always nonnegative and equals zero only if a perfect fit occurs; that is, if $\mathbf{S} - \mathbf{\Sigma} = 0$. For a large sample N, multiplying $F[\mathbf{S}, \mathbf{\Sigma}(\boldsymbol{\theta})]$ by $(N-1)$ yields a test statistic distributed approximately as a χ^2 with degrees of freedom equal to $d = [k(k+1)/2] - t$, where t is the number of estimated parameters. Because the minimum fit function χ^2 test statistic increases proportional to sample size, N, obtaining low χ^2 values with large samples often proves difficult. The CFA model in Fig. 1, based on $N = 1070$ cases, has $\chi^2 = 24.1$ for df (degrees of freedom) $= 13$ ($p = 0.03$), indicating that the model does not fit the data perfectly.

SEM computer programs print numerous goodness-of-fit indexes that can be used to assess overall model fit. Many indices are normed within a 0 to 1 range, with higher values reflecting better fits, but others have arbitrary metrics. Some fit indexes are functions of sample size, like χ^2, while others vary with degrees of freedom. For example, the widely used root mean square error of approximation (RMSEA) measures the mean of the squared discrepancies between observed and predicted matrices per degree of freedom. Small RMSEA values (<0.05) indicate a "close fit." The RMSEA for the CFA model in Fig. 1 is 0.029, indicating that the model fits the data quite well. One useful classification system distinguishes absolute, relative, and adjusted goodness-of-fit measures. Absolute indexes assess whether a specific model leaves appreciable unexplained variance. Relative fit indexes compare the specific model to possible baseline or null models estimated using the same data. Adjusted measures ask how well the model combines both fit and parsimony, taking into account the degrees of freedom used in the model specification. Analysts remain divided about criteria for selecting fit index and evaluating good fit. Several major points of consensus have emerged: (1) a strong substantive theory is the best guide to assessing model fit; (2) χ^2 should not be the sole basis for determining fit; (3) analysts should not rely on a single measure of overall fit; (4) other model components, such as equation R-squares and magnitudes of coefficient estimates, should be taken into account; and (5) rather than attempt to assess a single model's fit in some absolute sense, several models should be examined for plausible alternative structures.

Structural Equation Models

This section extends confirmatory factor analysis models to models with two or more latent variables having multiple indicators. Structural equation models combine factor analysis principles with path analysis and other path modeling methods in specifying a set of linear equations representing hypothesized relations among latent constructs and their multiple indicators. Structural equation models consist of two interrelated components, a measurement model and a structural model. The measurement model, which specifies how the latent constructs are indicated by their observed indicators, describes these indicators' measurement properties (reliabilities and validities) and is analogous to CFA. The structural equation model specifies causal relationships among the latent variables, describes their direct and indirect effects, and allocates explained and unexplained variance of the dependent constructs.

The observed indicators are partitioned into exogenous variables whose variation is predetermined outside the model, and endogenous variables whose variation is explained within the model. In the matrix algebra notation, a generic system of structural equations is denoted by $\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$, where $\boldsymbol{\eta}$ is a vector of unobserved endogenous variables, $\boldsymbol{\xi}$ is a vector of unobserved exogenous variables, $\boldsymbol{\varsigma}$ is a vector of unobserved errors, and $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ are the matrices of structural parameters to be estimated. The measurement model is specified by two equations: $\mathbf{Y} = \boldsymbol{\Lambda}_{\mathbf{Y}}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$ and $\mathbf{X} = \boldsymbol{\Lambda}_{\mathbf{X}}\boldsymbol{\xi} + \boldsymbol{\delta}$, where $\mathbf{Y}$ and $\mathbf{X}$ are vectors of the observed endogenous and exogenous indicators, the two $\boldsymbol{\Lambda}$ parameter matrices specify how the observed indicators are linked to the unobserved constructs (equivalent to factor loadings in a CFA model), and the $\boldsymbol{\varepsilon}$ and $\boldsymbol{\delta}$ vectors contain the error terms of the indicators.

An SEM implies a covariance structure for the observed variables. Estimating the model assumes empirical data from a random sample of N cases for which all the indicators have been measured. A computer program then uses an iterative algorithm to fit the specified SEM to the sample covariance matrix ($\mathbf{S}$) of the indicators. The program simultaneously estimates the free parameters of

both the structural and the measurement models, estimates standard errors for each parameter, and calculates various goodness-of-fit indexes for the whole model. Several SEM computer programs perform these computations, for example, LISREL, AMOS, EQS, MPLUS, and SAS CALIS. Most programs no longer require analysts to specify their models in formal matrix algebra language, but use simple programming instructions to denote the hypothesized relations among latent and observed variables. Some SEM programs allow a researcher to draw a diagram on a computer screen, then translate it into software commands that estimate the model parameters.

Diagrams are indispensable tools for conceptualizing and interpreting a SEM. In the simple example in Fig. 3, the structural model depicts an exogenous "political ideology" construct causing variation in an endogenous "federal help" construct. The measurement model consists of two observed indicators of politics (conservative political views and party identification) and four indicators of help (attitudes against the federal government's responsibility for solving social problems: not helping with poverty, not helping with any problems, not helping with medical bills, and not helping African Americans). Parameter estimates for the completely standardized solution were computed by LISREL from a covariance matrix computed for 1594 respondents in the 1998 GSS. The overall model fit is very good ($\chi^2 = 15.7$, df = 8, $p = 0.052$; other fit indexes have quite acceptable values), and the individual parameter estimates are all highly significant. The estimated structural parameter (0.63) means that a difference of one standard deviation in political ideology is associated with a three-fifths standard deviation difference in attitude toward the federal government's role in solving social problems. Given that all indicators were measured with conservative responses scored high and liberal responses scored low, the positive sign means that more ideologically conservative respondents favor more individualistic solutions to social problems (i.e., less federal government involvement).

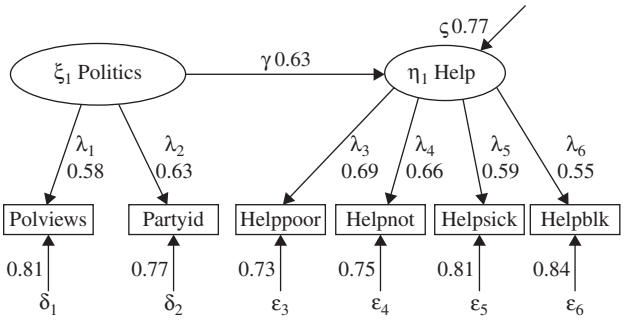


Figure 3 Completely standardized solution for a simple structural equation model.

SEM programs allow an explicit statistical test of the hypothesis that two or more parameters are equal in the population. Constraining a pair of parameters to be equal (rather than letting them freely take differing values) requires estimating only one parameter instead of two. As a result, one degree of freedom is then available to assess whether constrained and unconstrained models' χ^2 statistics differ at a chosen α-level. If no significant difference occurs between the two models, then the more parsimonious version with equal parameters (i.e., the model with fewer unconstrained parameters) would be preferred. In the model in Fig. 3, the standardized parameters of the four help indicators seem to have roughly similar magnitudes (ranging between 0.55 and 0.69). Some alternative models that specified equal loadings for several pairs of indicators did not produce significantly worse fits to the data. However, a model that hypothesized equal loadings for HELPOOR and HELPBLK was rejected (χ^2 difference of $24.9 - 15.7 = 9.2$ for $df = 1$, $p < 0.01$). Another model, hypothesizing equal factor loadings for the first three indicators (HELPPOOR, HELPNOT, HELPSICK), did not produce a significantly worse fit compared to the model with no equality constraints (χ^2 difference of $19.2 - 15.7 = 3.5$ for $df = 2$). The three help indicators each had estimated parameters equal to 0.76, while the HELPBLK indicator had a lower value (0.66).

Model Identification and Modification Strategies

For a model to be estimable, both its measurement and its structural equation portions must be identifiable. An SEM or CFA model is identified if every unknown parameter has a unique value that can be estimated by fitting the model to the data. A model is "underidentified," and not estimable, if the number of unknown parameters to be estimated exceeds the available degrees of freedom (the number of indicator variances and covariances). In such instances, a model can be respecified to ensure identification by constraining sufficient numbers of the unknown parameters to fixed values (typically set to zero). "Just identified" models, with precisely as many unknown parameters as available degrees of freedom, always produce trivially perfect fits but may provide useful baseline estimates against which to test other models with positive degrees. "Overidentified" models, with positive degrees of freedom, reveal whether the model specifications reasonably represent relationships in both the measurement and structural models. For a complicated SEM, a researcher's *a priori* identification of all parameters may become difficult because fulfilling all the formal requirements to assure identification can often be quite complicated. SEM computer programs usually ascertain whether a hypothesized

model is not identified, if they cannot calculate unique estimates with standard errors for the unknown parameters. Nevertheless, model and parameter identification remain relevant concerns, because SEM computer programs may occasionally produce solutions for unidentified models. Model modification strategies contain additional important concerns. Unless a model fits the data well, researchers seldom fit a single hypothesized model to their data, then stop after making the decision to accept or reject that specification without proposing any alternative. The more common practice involves model generation strategy, an exploratory approach that incrementally respecifies parameters and fits a series of alternative CFA or SEM to the same data. The analyst's ultimate objective is to find an overidentified model that fits the data at an acceptable level (using the various goodness-of-fit indexes), while also yielding plausible and meaningful interpretations of the estimated parameters. Exploratory modifications of a tentative initial model should not rely entirely on statistical criteria, but should also take into account existing theory and empirical knowledge about a substantive area. Important procedures for locating sources of model misspecification include examining parameter estimates for unrealistic values or anomalous signs inconsistent with theoretical expectations; assessing squared multiple correlations (R^2) for each equation for evidence of weak or nonlinear relations; and inspecting residuals, standardized residuals, and model modification indices to pinpoint expected parameter changes and fit improvement, for example, by correlating error terms. By repeating these steps for successively modified models, analysts may obtain a final version that fits the sample data reasonably well and provides a plausible interpretation.

Unfortunately, a final SEM or CFA model with an improved fit to the data is unlikely to be the "true" or "best-fitting" model, in the sense that its successive improvements involved capitalizing on chance covariation in the sample data. Instead, it is probably one of several alternative models of equivalent overall fit that approximate the unknown true population SEM. A more robust approach to SEM generation cross-validates the modified model results with an independent sample. Alternatively, researchers can randomly split a sufficiently large sample in half, using the first subsample to estimate the modified model and the second subsample to cross-validate that specification. Recent developments in automated algorithms, such as TETRAD and TABU, assist researchers in their search for true model specifications and parameter estimates.

Strengths and Limitations of SEM

Both CFA and SEM methods, implemented in a variety of computer packages, provide researchers with powerful data analysis tools. Applied judiciously, these methods have important advantages over traditional multivariate methods, such as linear regression, that assume no errors in observed measures. If an SEM model is true, then its structural parameters linking the latent constructs take into account the biases of less reliably and validly measured indicators. However, the price paid for these advantages is susceptibility to erroneous parameter estimates and model fits if analysts misspecify the true measurement and structural relationships. For example, covariation in cross-sectional data offers no clues to asymmetric or reciprocal causation; even the temporal sequences among repeated measures in longitudinal panel designs are not an infallible guide to causal order. Because SEM methods by themselves do not enable researchers to distinguish among many alternative models with statistically equivalent fits, analysts face heavy requirements to apply logic and theory jointly to distinguish incredible from plausible alternative model specifications. The protean qualities of SEM methods should spur researchers to work harder at improving their theoretical understanding of the social processes they seek to explain.

Acknowledgments

The author thanks George W. Bohrnstedt, Francisco J. Granados, and three anonymous reviewers for their comments on previous drafts.

See Also the Following Articles

Factor Analysis • Measurement Theory • Reliability • Validity Assessment

Further Reading

Bollen, K. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Bollen, K., and Long, J. (1993). *Testing Structural Equation Models.* Sage Publications, Thousand Oaks, CA.

Breckler, S. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychol. Bull.* **107,** 260–273.

Cliff, N. (1983). Some cautions concerning the application of causal modelling methods. *Multivar. Behav. Res.* **18,** 115–126.

Cudeck, R., and Browne, M. (1983). Cross-validation of covariance structures. *Multivar. Behav. Res.* **18,** 147–167.

Jöreskog, K., and Sörbom, D. (1996). *LISREL 8: User's Reference Guide.* Scientific Software International, Chicago.

Kenny, D., Kashy, D., and Bolger, N. (1998). Data analysis in social psychology. In *Handbook of Social Psychology* (D. Gilbert, S. Fiske, and G. Lindzey, eds.), 4th Ed., Vol. 1, pp. 233–265. McGraw-Hill, Boston, MA.

Knoke, D., Bohrnstedt, G., and Mee, A. (2002). *Statistics for Social Data Analysis*, 4th Ed. F. E. Peacock, Itasca, IL.

MacCallum, R., Wegener, D., Uchino, B., and Fabrigar, L. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychol. Bull.* **114,** 185–199.

Marcoulides, G., and Schumaker, R. (eds.) (1996). *Advanced Structural Equation Modeling: Issues and Techniques*. Erlbaum, Mahwah, NJ.

Maruyama, G. (1998). *Basics of Structural Equation Modeling*. Sage Publications, Thousand Oaks, CA.

Richardson, T., and Spirtes, P. (1999). Automated discovery of linear feedback models. In *Computation, Causation, and Discovery* (C. Glymour and G. Cooper, eds.), pp. 253–304. MIT Press, Cambridge, MA.

Structural Item Response Models

Cees A. W. Glas
University of Twente, Enschede, The Netherlands

Glossary

differential item functioning Items that have different item parameters in different groups of respondents.
item response model Stochastic measurement model with distinct person and item parameters.
Markov chain Monte Carlo estimation Estimation procedure that generates the posterior distribution of parameters via simulation.
maximum marginal likelihood estimation Estimation procedure based on a likelihood function that is marginalized with respect to nuisance parameters.
multilevel item response model Item response model where person or item parameters are nested in a hierarchical structure.
structural item response model Item response model with an additional model for person or item parameters.
testlet model Item response model where items are clustered in subtests which are usually called testlets.

Structural item response theory models are models that consist of two parts: an item response theory model with distinct person and item parameters, and an additional model that specifies the structure of either the person and item parameters (or even both). Structural item response models support modeling of differences between groups of respondents and complex response formats. Further, structural item response theory models can play a role as alternative models in tests of model fit for more basic item response models.

Introduction

Item Response Models

Item response theory (IRT) models are stochastic models for two-way data, say, the responses of persons to items. An essential feature of these models is parameter separation, that is, the influences of the items and persons on the responses are modeled by distinct sets of parameters. To illustrate parameter separation, consider the two-way data matrix in Table I. The first 3 persons responded to the first 3 items, persons 4, 5, and 6 responded to items 4, 5, and 6, and the last two persons responded to items 1, 2, and 6. Since different respondents took different tests, their total scores cannot be compared without additional assumptions. For instance, it is unclear whether the score 9 obtained by person 3 represents the same ability level as the score 9 obtained by person 5, because they might have responded to items of a different difficulty level.

In the present, highly hypothetical case, the data were constructed according to a very simple deterministic linear model given by $y_{ik} = \theta_i + b_k$, where y_{ik} stands for the response of person i to item k. The person parameter θ_i can be viewed as the ability of person i and the item parameter b_k can be viewed as the easiness item k. The values of θ_i and b_k, and the way in which they account for the data, are shown in Table II. It can now be seen that person 3 has an ability level $\theta_3 = 1$, while person 5 has an ability level $\theta_5 = 2$.

Table I Data Matrix with Observed Scores

	Item						
Respondent	1	2	3	4	5	6	*Score*
1	2	3	1				6
2	4	5	3				12
3	3	4	2				9
4				4	5	3	12
5				3	4	2	9
6				2	3	1	6
7	3	4				1	8
8	2	3				0	5

Table II Effects of Items and Persons Separated

	Item						
Respondent	1	2	3	4	5	6	θ_i
1	0 + 2	0 + 3	0 + 1				0
2	2 + 2	2 + 3	2 + 1				2
3	1 + 2	1 + 3	1 + 1				1
4				3 + 1	3 + 2	3 + 0	3
5				2 + 1	2 + 2	2 + 0	2
6				1 + 1	1 + 2	1 + 0	1
7	1 + 2	1 + 3				1 + 0	1
8	0 + 2	0 + 3				0 + 0	1
b_k	2	3	1	1	2	0	

Of course, in practice, this kind of deterministic model never fits the data, so to calibrate item and person parameters on a common scale, a statistical model with the property of parameter separation must be used. A natural choice is an IRT model, many of which are outlined in the present volume. For dichotomous items, the 1-, 2-, and 3-parameter logistic models (1PLM, 2PLM, and 3PLM, Birnbaum) are used most often. The 3-parameter logistic model is given by

$$p(y_{ik}=1\,|\,\theta_i,a_k,b_k,c_k)=c_k+(1-c_k)\Psi(a_k(\theta_i-b_k))$$
$$=c_k+(1-c_k)\frac{\exp(a_k(\theta_i-b_k))}{1+\exp(a_k(\theta_i-b_k))}, \quad (1)$$

where a_k, b_k, and c_k are the discrimination, difficulty, and guessing parameters, respectively. The 2PLM follows upon introducing the constraint $c_k=0$ and the 1PLM follows upon introducing the additional constraint $a_k=1$. Note that $\Psi(a_k(\theta-b_k))$ stands for the logistic function evaluated at $a_k(\theta-b_k)$. In an alternative, but for all practical purposes, equivalent formulation, the logistic function is replaced by the normal ogive function, $\Phi(a_k(\theta-b_k))$, which stands for the standard normal function integrated to $a_k(\theta-b_k)$. The choice between the two formulations is often determined by computational convenience. Generalizations of the 2PLM to models for responses to polytomous items include the graded response model and the nominal response model with special cases as the rating scale model and the partial credit model. Multidimensional models where the latent ability parameters are multidimensional were developed by McDonald. All these models share the feature of parameter separation and the possibility of using incomplete designs to calibrate item and person parameters on a common scale.

It should be noted that estimation of the model parameters from data in incomplete designs has its limitations. First, the design should be linked. In the example given above, the design would not be linked if only the data of the first six respondents were available, because in that case the data matrix would break down into two separate data matrices that would have no persons or items in common. The responses of the last two persons serve as a link between these two data matrices. Second, the common procedures to obtain consistent estimates of the parameters assume that the design is ignorable. In the present framework, Rubin's theory of ignorability entails that the test administration design should not depend on unobserved data. Therefore, an item administration design is ignorable in applications where the design is *a priori* fixed, but also in some applications where this is not the case, such as in multistage testing and in computerized adaptive testing.

Besides the possibility that item and person parameters can be calibrated on a common scale several other applications of IRT deserve mention. First, IRT can be used to support the construct validity of a test. If it can be empirically shown that a test is unidimensional, this can be viewed as empirical evidence that all items of the test measure the same construct. Second, the fitted IRT model implies a scoring rule for the test. For instance, if the 1PLM holds, a meaningful variable can be created by summation of the item scores for each person, and this variable is a sufficient statistic for θ. If the test proves multidimensional, multidimensional IRT models give insight in the structure of the test content. Further applications of IRT are the evaluation of differential item functioning, optimal test construction, and computerized adaptive testing. Finally, IRT provides a solution for the attenuation problem, that is, the problem that the correlation between observed scores is often attenuated by the unreliability of the measurement instruments. When properly estimated the correlation between latent variables does not suffer from this problem.

Structural Item Response Models

In this article, structural IRT models are defined as a mixture of an IRT model and a model that specifies the structure of either the person or item parameters (or both). If independence between respondents and items is assumed, a structural IRT model can be generally defined by the likelihood function

$$L(\theta,\delta,\lambda,\tau;y,x)=\prod_i^I\prod_k^K p(y_{ik}\,|\,\theta_i,\delta_k)^{d_{ik}}g(\theta_i\,|\,\lambda,x)$$
$$\times h(\delta_k|\tau,x), \quad (2)$$

where y_{ik} is the response variable and d_{ik} is a design variable assuming a value 1 if the item was responded to

and zero otherwise. Further, $p(y_{ik} \mid \theta_i, \delta_k)$ is the response probability under the IRT model (with an example given in (1)), $g(\theta_i \mid \lambda, x)$ is the density function for θ_i, and $h(\delta_k \mid \tau, x)$ is the density function of δ_k. The latter densities have parameters which are a function of parameters λ and τ, respectively, and both may depend on covariates denoted by x.

Overview

An exhaustive review of all currently used structural IRT models is beyond the scope of this article. Therefore, this article reviews some of the best-known models, as far as they are not treated elsewhere in this encyclopedia. Models for ability parameters, from a simple model to evaluate differences between groups to rather complex multilevel linear models that allow for measurement error in both the dependent and independent variables, are studied, as are models for the item parameters. A distinction is made between fixed effect models (with an application to the evaluation of differential item functioning) and random effect models (with an application to the analysis of responses to item clones). A model for testlets with a random parameter for the interaction between respondents and items is discussed. Finally, references to relevant software are mentioned.

Models for Ability Parameters

Differences in Ability Level between Groups

Suppose respondents are sampled from two populations, say males and females, and the interest is in evaluating the difference in the mean ability level of the two populations. Gender is coded as

$$x_i = \begin{cases} 1 & \text{if } i \text{ is a male} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and gender differences in ability are modeled as

$$\theta_i = \mu + \beta x_i + e_i, \tag{4}$$

where it is assumed that e_i has a normal distribution with mean zero and variance σ^2. Note that μ is the mean ability level of the females, while $\mu + \beta$ is the mean ability level of the males. So β is the effect of being male. In IRT, the assumption that the variance σ^2 is equal over groups is easily generalized to the assumption that groups have unique variances, say σ_g^2.

Maximum marginal likelihood (MML) estimation is the most used technique for parameter estimation in IRT models. In this approach, a distinction is made between structural parameters, which need to be consistently estimated and nuisance parameters, which are not of primary interest. MML estimation derives its name from maximizing a likelihood function that is marginalized with respect to the nuisance parameters. In the present case, the likelihood is marginalized with respect to the ability parameters θ, leading to the marginal likelihood

$$L(\delta, \beta, \mu, \sigma; y, x) = \prod_i^I \int_{-\infty}^{\infty} \prod_k^K p\left(y_{ik} \mid \theta_i, \delta_k\right)^{d_{ik}} \times g(\theta_i \mid \mu, \beta, \sigma, x_i)\, d\theta_i, \tag{5}$$

where $g(\theta_i \mid \mu, \beta, \sigma, x_i)$ stands for the normal density as implied by Eq. (4). The reason for maximizing the marginal rather than the joint likelihood of all parameters simultaneously is that maximizing the latter likelihood does not lead to consistent estimates.

Table III gives a small simulated example of the procedure. The data were generated with the 1PLM. The design consisted of 9 items administered to two groups. Group 1 consisted of 100 simulees who responded to the items 1 to 6. The second group consisted of 400 simulees responding to the items 4 to 9. So the items in the so-called "anchor" were responded to by 500 simulees. The true item parameters b_k are shown in the second column of Table III, the MML parameter estimates $\hat{b}_k$ and their standard errors $se(\hat{b}_k)$ are shown in the third and fourth column, respectively. Note that the standard errors are inversely proportional to the number of simulees responding to the item. The bottom lines of the table give the generating values for β, σ_g $(g = 1, 2)$, their estimates, and their standard errors. In this example, μ has been set equal to zero to identify the scale of θ. The test whether the two groups have the same mean ability

Table III Parameter Values and Estimates

Item	b_k	$\hat{b}_k$	$se(\hat{b}_k)$
1	−1.00	−0.71	0.33
2	0.00	−0.04	0.30
3	1.00	1.18	0.29
4	−1.00	−1.10	0.14
5	0.00	−0.17	0.13
6	1.00	1.09	0.14
7	−1.00	−1.09	0.35
8	0.00	0.00	0.34
9	1.00	0.94	0.35
Pop	β	$\hat{\beta}$	$se(\hat{\beta})$
1	1.00	1.07	0.22
Pop	σ_g	$\hat{\sigma}_g$	$se(\hat{\sigma}_g)$
1	1.00	1.13	0.18
2	1.50	1.45	0.10

level, that is, the test of the null hypothesis $\beta = 0$ against the alternative $\beta \neq 0$, can be based on the ratio of the estimate of β with its standard error, that is, $\hat{\beta}/se(\hat{\beta})$. In the present case, the outcome is $1.07/0.22 = 4.864$. Under the null-hypothesis, the statistic has a standard normal distribution, so the null-hypothesis is clearly rejected.

This approach can be generalized in various ways. One could introduce a second variable, say

$$x_{2i} = \begin{cases} 1 & \text{if } i \text{ lives in an urban area} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

and consider the model

$$\theta_i = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + e_i. \tag{7}$$

If x_{1i} stands for gender, then β_{12} stands for the interaction of being male and living in an urban area, and, as above, a test of the hypothesis $\beta_{12} = 0$ against the alternative $\beta_{12} \neq 0$ can be based on the parameter estimate relative to its standard error. The next section gives further generalizations of this approach.

Multilevel Regression Models on Ability

In much social research, elementary units are clustered in higher level units. A well-known example is educational research, where pupils are nested within classrooms, classrooms within schools, schools within districts, and so on. Multilevel models (ML models) have been developed to take the resulting hierarchical structure into account, mostly by using regression-type models with random coefficients. However, if variables in these multilevel models contain large measurement errors, the resulting statistical inferences can be very misleading. Measurement error can be modeled in the framework of classical test theory and IRT. In the classical framework, the variance component due to unreliability can either be imputed in the model or it can be estimated within the model, for instance by splitting test scores into subtest scores. The IRT framework is a generalization of the linear model described above. The approach entails the definition of a multilevel linear model where latent variables from IRT measurement models are entered either as dependent or as independent variables. The resulting model is the so-called multilevel IRT model. The general model is defined as follows. The dependent variables are observed item scores y_{ijk}, where the index i $(i = 1, \ldots, n_j)$ signifies the respondents, the index j $(j = 1, \ldots, J)$ signifies the Level 2 clusters, say the schools, and the index k $(k = 1, \ldots, K)$ signifies the items. The first level of the structural multilevel model is formulated as

$$\theta_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \cdots + \beta_{q'j} x_{q'ij} + \beta_{(q'+1)j} \xi_{(q'+1)ij} + \cdots + \beta_{Qj} \xi_{Qij} + e_{ij}, \tag{8}$$

where the covariates x_{qij} $(q = 1, \ldots, q')$ are manifest predictors and the covariates ξ_{qij} $(q = q' + 1, \ldots, Q)$ are latent predictors. Finally, e_{ij} are independent and normally distributed error variables with mean zero and variance σ^2. In general, it is assumed that the regression coefficients β_{qj} are random over groups, but they can also be fixed parameters. In that case, $\beta_{qj} = \beta_q$ for all j. The Level 2 model for the random coefficients is given by

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1} z_{1qj} + \cdots + \gamma_{qs'} z_{s'qj} + \gamma_{q(s'+1)} \zeta_{(s'+1)qj} + \cdots + \gamma_{qS} \zeta_{Sqj} + u_{qj}, \tag{9}$$

where z_{sqj} $(s = 1, \ldots, s')$ and ζ_{sqj} $(s = s' + 1, \ldots, S)$ are manifest and latent predictors, respectively. Further, u_{qj} are error variables which are assumed independent over j and have a Q-variate normal distribution with a mean equal to zero and a covariance matrix T.

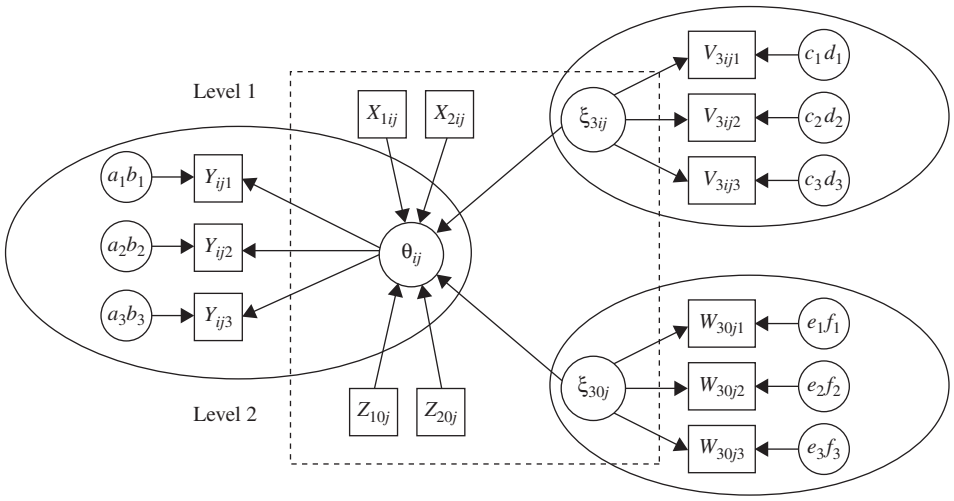


Figure 1 Path diagram of a multilevel IRT model.

An example of a MLIRT model is given in the path diagram in Fig. 1. The structural multilevel part is presented in the big square box in the middle. The structural model has two levels: the upper part of the box gives the first level (a within-schools model), and the lower part of the box gives the second level (a between-schools model). The dependent variable θ_{ij}, say math ability, is measured by three items. The responses to these items are modeled by the 2PLM with item parameters a_k and b_k, $k = 1, \ldots, 3$. Note that the measurement error models are presented by the ellipses. Both levels have three independent variables: two are observed directly, and one is a latent variable with three binary observed variables. For instance, on the first level, X_{1ij} could be gender, X_{2ij} could be age, and ξ_{3ij} could be intelligence as measured by a three item test. On the second level, Z_{10j} could be school size, Z_{20j} could be the school budget and ζ_{30j} could be a school's pedagogical climate, again measured by a three-item test. In order not to complicate the model, it is assumed that only the intercept β_{0j} is random, so the Level 2 predictors are only related to this random intercept and the slopes are fixed.

The parameters in the MLIRT model can be estimated in a Bayesian framework with a version of the Markov chain Monte Carlo (MCMC) estimation procedure: the Gibbs sampler. There are many considerations when choosing between a frequentist framework (such as MML) and Bayesian framework (such as MCMC), but the reason for adopting the Bayesian approach given by Fox and Glas is a practical one: MML involves integration over the nuisance parameters, and in the present case these integrals become quite complex. In the Bayesian approach, the interest is in the posterior distribution of the parameters, say $p(\theta, \delta, \beta, \mu, \sigma \mid y)$. In the MCMC approach samples are drawn from the posterior distribution and in this process nuisance parameters play a role as auxiliary variables. So the problem of complex multiple integrals does not arise here.

To give some idea of the output of the procedure, consider an application reported by Shalabi. The data were a cluster sample of 3384 grade 7 students in 119 schools. At student level the variables were Gender (0 = male, 1 = female), SES (with two indicators: the father's and mother's education, scores ranged from 0 to 8), and IQ (range from 0 to 80). School level variables were Leadership (measured by a scale consisting of 25 five-point Likert items, administered to the school teachers), School Climate (measured by a scale consisting of 23 five-point Likert items), and Mean-IQ (the IQ scores aggregated at school level). The items assessing Leadership and School Climate were dichotomous. The dependent variable was a mathematics achievement test consisting of 50 multiple-choice items. The 2PLM was used to model the responses to the Leadership and School Climate questionnaire and the mathematics test. The parameters were estimated with the Gibbs sampler. For a complete description of all analyses, one is referred to the work of Shalabi, here only the estimates of the final model are given as an example.

Table IV Estimates of the Effects of Leadership, Climate, and Mean IQ

	MLIRT estimates		*ML estimates*	
	Estimates	*C.I.*	*Estimates*	*C.I.*
γ_{00}	–1.096	–2.080––0.211	–0.873	–1.20––0.544
β_1	0.037	0.029–0.044	0.031	0.024–0.037
β_2	0.148	0.078–0.217	0.124	0.061–0.186
β_3	0.023	0.021–0.025	0.021	0.019–0.022
γ_{01}	0.017	0.009–0.043	0.014	0.004–0.023
γ_{02}	0.189	0.059–0.432	0.115	0.019–0.210
γ_{03}	–0.136	–0.383––0.087	–0.116	–0.236–0.004
Variance components				
τ_0^2	0.177	0.120–0.237	0.129	0.099–0.158
σ^2	0.189	0.164–0.214	0.199	0.190–0.210

The model is given by

$$\theta_{ij} = \beta_{0j} + \beta_1 \mathrm{SES}_{ij} + \beta_2 \mathrm{Gender}_{ij} + \beta_3 \mathrm{IQ}_{ij} + e_{ij}, \quad (10)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \mathrm{Mean\text{-}IQ}_j + \gamma_{02} \mathrm{Leadership}_j + \gamma_{03} \mathrm{Climate}_j + u_{0j}. \quad (11)$$

The results are given in Table IV. The estimates of the MLIRT model are compared with a traditional ML analysis where all variables were manifest. The observed Mathematics, Leadership, and School Climate scores were transformed in such a way that their scale was comparable to the scale used in the MLIRT model. Further, the parameters of the ML model were also estimated with a Bayesian approach using the Gibbs sampler.

The columns labeled C.I. give the 90% credibility intervals of the point estimates; they were derived from the posterior standard deviation. It can be seen that the magnitudes of the fixed effects in the MLIRT model were larger than the analogous estimates in the ML model. This finding is in line with the other findings, which indicates that the MLIRT model has more power to detect effects in hierarchical data where some variables are measured with error.

Models for Item Parameters

Fixed Effects Models: Differential Item Functioning

Differential item functioning (DIF) is a difference in item responses between equally proficient members of two or more groups. For instance, a dichotomous item is subject

to DIF if, conditionally on ability level, the probability of a correct response differs between groups. One might think of a test of foreign language comprehension, where items referring to football impede girls. The poor performance of the girls on the football-related items must not be attributed to their low ability level but to their lack of knowledge of football. Since DIF is highly undesirable in fair testing, several techniques for the detection of DIF have been proposed. Most of them are based on evaluation of differences in response probabilities between groups conditional on some measure of ability. The most generally used technique is based on the Mantel–Haenszel statistic, others are based on log-linear models and on IRT models.

In the Mantel–Haenszel (MH) approach, the respondent's number-correct score is used as a proxy for ability and DIF is evaluated by testing whether the response probabilities differ between the score groups. Though the MH test works quite well in practice, its application is based on the assumption that the number-correct score is a sufficient statistic for ability, that is, that the 1PLM holds. In application of the MH test in other cases, such as cases where the data follow the 2PLM or the 3PLM, the number-correct score is no longer the optimal ability measure. In an IRT model, ability is represented by a latent variable θ, and an obvious solution to the problem is to evaluate whether the same item parameters apply in subgroups that are homogeneous with respect to θ.

As an example, DIF can be investigated by introducing a more general alternative to the 3PL model as defined in Eq. (1) given by

$$p(y_{ik} = 1 \mid \theta_i, a_k, b_k, c_k, d_k) = c_k + (1 - c_k)\Psi(a_k(\theta_i - b_k - x_i d_k)), \quad (12)$$

where x_i is the background variable Gender as defined by (3), and d_k is the change in the difficulty level of item k for males. The model defined by (12) pertains to dichotomous items, but the idea of modeling DIF by introducing item parameters depending on background variables also applies to polytomous items.

Tests for DIF are usually item-oriented, that is, items are tested one at a time. In general, a test for DIF can be defined by choosing a no-DIF IRT model (say, the 3PLM) as the null hypothesis and an IRT model for DIF (say, the model given by Eq. (12)) as the alternative. The test can be based on a likelihood ratio statistic or a Wald statistic. Both statistics require maximum likelihood estimates of both the parameters under the null model and the alternative model. Therefore, Glas proposed using the Lagrange multiplier statistic, which only requires estimation of the null-model. The LM test is based on the evaluation of the first-order partial derivatives of the log-likelihood function of the alternative model evaluated using the maximum likelihood estimates of the null model. The magnitudes of these first-order partial derivatives determine the value of the statistics, that is, the closer they are to zero, the better the model fit. The LM statistic is asymptotically chi-square distributed with degrees of freedom equal to the difference in the number of parameters of the two models.

Tables V and VI give a small simulated example of the procedure. The data were generated according to

Table V Parameter Generating Values, Estimates, and the LM Statistic

Item	b_k	$\hat{b}_k$	$se(\hat{b}_k)$	*LM*	*Pr*
1	−1.00	−0.91	0.12	1.33	0.25
2	0.00	0.13	0.11	0.27	0.61
3	1.00	1.13	0.12	1.14	0.29
4	−1.00	−0.93	0.11	1.14	0.29
5	0.0/0.5	0.41	0.11	18.03	0.00
6	1.00	1.04	0.12	0.02	0.90
7	−1.00	−0.77	0.12	0.05	0.83
8	0.00	0.11	0.11	0.01	0.92
9	1.00	1.03	0.11	0.11	0.74
Pop	β	$\hat{\beta}$	$se(\hat{\beta})$		
1	1.00	1.00	0.11		
Pop	σ_g	$\hat{\sigma}_g$	$se(\hat{\sigma}_g)$		
1	1.00	1.01	0.07		
2	1.50	1.41	0.08		

Table VI Parameter Generating Values, Estimates, and the LM Statistic after Splitting the DIF Item into Two Virtual Items

Item	b_k	$\hat{b}_k$	$se(\hat{b}_k)$	*LM*	*Pr*
1	−1.00	−0.88	0.12	0.32	0.57
2	0.00	0.18	0.11	1.27	0.26
3	1.00	1.18	0.12	0.23	0.63
4	−1.00	−0.90	0.12	0.23	0.63
5	0.50	0.81	0.15	—	—
6	1.00	1.10	0.12	0.22	0.63
7	−1.00	−0.73	0.12	0.11	0.74
8	0.00	0.16	0.11	0.23	0.63
9	1.00	1.09	0.11	0.90	0.34
10	0.00	0.08	0.15	—	—
Pop	β	$\hat{\beta}$	$se(\hat{\beta})$		
1	1.00	1.00	0.11		
Pop	σ_g	$\hat{\sigma}_g$	$se(\hat{\sigma}_g)$		
1	1.00	1.01	0.07		
2	1.50	1.41	0.08		

the same setup as the example of Table III, but with some differences. First, both groups now consist of 400 simulees, and both groups respond to all 9 items. However, to simulate DIF, for the first group the parameter of item 5 was changed from 0.00 to 0.50. Table V gives the generating values of the parameters, the estimates and the standard errors. The last two columns of Table V give the value of the LM statistic and the associated significance probability. In the present case, the LM statistic has an asymptotic chi-square distribution with one degree of freedom. The test is highly significant for item 5.

For the analysis of Table VI, item 5 has been split into two virtual items: item 5 was assumed to be administered to group 1, and item 10 was assumed to be administered to group 2. So the data are now analyzed assuming an incomplete item administration design, where group 1 responded to the items 1 to 9 and group 2 responded to the items 1 to 4, 10, and 6 to 9 (in that order). As a consequence, one group only responded to the virtual items 5 and 10, and the LM test for DIF cannot be performed for these items. It can be seen in Table VI that the values of the LM statistics for the other items are not significant, which gives an indication that the model fit now fits.

Random Effect Models for Variability in Item Parameters

In the previous section, variability of item parameters was treated as a fixed effect, that is, the item parameters were a finite number of unique entities. In the present section, the focus is on item parameters as random effects, that is, the item parameters are seen as exchangeable draws from a distribution. Interest in item sampling is related to the introduction of computer-generated items in educational measurement. Using item-cloning techniques, items can be generated by a computer from a smaller set of "parent items" through the use of transformation rules. An example is the "replacement set procedure," where elements of the parent item (e.g., key terms, relations, numbers, and distractors) are randomly chosen from a well-defined set of alternatives. Because this introduces (slight) random variation between items derived from the same parent, it is possible to model the item parameters as random and shift the interest to the hyper-parameters that describe the distributions of the item parameters within parents.

To define the model, consider a set of item populations $p = 1, \ldots, P$ of size $K_1, \ldots, K_P$, respectively. The items in population p will be labeled $k_p = 1, \ldots, K_p$. The first-level model is the 3PLM which describes the probability of a correct response as $p(y_{ik_p} \mid \theta_i\ a_{k_p},\ b_{k_p},\ c_{k_p})$, as in Eq. (1) but with the subscript changed from k to k_p. In the Level 2 model, the values of the item parameters a_{k_p}, b_{k_p}, c_{k_p} are considered as realizations of a random vector. It is assumed that the item parameters, say ξ_{k_p}, have a 3-variate normal distribution with mean μ_p and a covariance matrix Σ_p. To support the assumption of normality, the item parameters are transformed as $\xi_{k_p} = (a_{k_p},\ b_{k_p},\ \text{logit}\, c_{k_p})$ or as $\xi_{k_p} = (\log a_{k_p},\ b_{k_p},\ \text{logit}\, c_{k_p})$. The logit transformation is a standard way to map a probability, such as c_{k_p}, to the real numbers, and taking the logarithm of a_{k_p} assures that a_{k_p} is positive. The model can be estimated by Bayesian methods based on the MCMC procedure or by MML.

The Testlet Model

A testlet is a subset of items related to some common context. Usually these sets take the form of a number of multiple choice items organized under or within some text. When a test consists of a number of testlets, both the within and between dependence between the items play a role. One approach is to ignore this hierarchical dependence structure and analyze the test as a set of atomistic items. This generally leads to an overestimate of measurement precision and bias in the item parameter estimates. Another approach is to aggregate the item scores within the testlet to a testlet score and analyze the testlet scores using an IRT model for polytomously scored items. This approach discards part of the information in the item responses, which will lead to loss of measurement precision. The rigorous way to solve the problem is to model the within and between dependence explicitly. Wainer, Bradlow, and Du introduce a generalization of the 2PLM given by

$$p(y_{ik} \mid \theta_i, a_k, b_k, c_k, \gamma_{id(k)}) = c_k + (1 - c_k)\Psi(a_k(\theta_i - b_k + \gamma_{id(k)})),$$

where $d(k)$ is the testlet to which item k belongs and $\gamma_{id(k)}$ a person-specific testlet effect. It is assumed that $\gamma_{id(k)}$ has a normal distribution with a mean equal to zero and variance σ^2_γ. Further, it is assumed that θ has a standard normal distribution.

The parameters in the model can be estimated in a Bayesian framework using MCMC or in a frequentist framework using MML. Glas, Wainer, and Bradlow report a number of simulation studies performed to assess the effect of ignoring the testlet structure on the precision of item calibration. Some of their results are reported here. Every simulee responded to 40 items. The item discrimination parameters a_i were drawn from a uniform distribution on the interval $[0.8, \ldots, 1.2]$, the item difficulty parameters b_i were drawn from a uniform distribution on $[-1, \ldots, 1]$, and all item guessing parameters c_i were equal to 0.25. The ability parameters θ were drawn from a standard normal distribution. Fixing the guessing parameter to its true value was sufficient to obtain MML estimates

Table VII Mean Absolute Error of Item Parameter Estimates for 3PLM and Testlet Model

$\gamma_{id(k)}$	*Number of items in testlet*	*Number of testlets*	N	*3PLM*		*Testlet Model*	
				MAE(a)	*MAE(b)*	*MAE(a)*	*MAE(b)*
0.25	10	4	2000	0.083	0.070	0.083	0.071
			5000	0.046	0.046	0.046	0.046
	5	8	2000	0.083	0.068	0.081	0.068
			5000	0.046	0.036	0.048	0.037
1.00	10	4	2000	0.092	0.064	0.072	0.062
			5000	0.025	0.060	0.038	0.035
	5	8	2000	0.112	0.082	0.072	0.066
			5000	0.087	0.075	0.039	0.041

(without priors on the other parameters). In the simulation study, three factors were varied: the number of testlets (4 or 8, and, hence, 10 and 5 items per testlet), the number of simulees (2000 or 5000), and the testlet effect size $\gamma_{id(k)}$ (0.25 or 1.00). Table VII gives results averaged over 10 replications. In the columns labeled *MAE*(*a*) and *MAE*(*b*), the mean absolute errors of the estimates of the discrimination and difficulty parameters are presented, computed ignoring and including the testlet parameters. The difference between the parameter estimates was negligible for the cases where $\gamma_{id(k)} = 0.25$, while moderate effects appear for the more substantial within-persons standard deviation $\gamma_{id(k)} = 1.00$.

Conclusion

A final remark concerns the software for estimation and testing of the models discussed above. First, simple linear models for θ can be directly computed using standard IRT software that can handle multiple groups, such as Bilog, Multilog, Parscale, Testfact (products of Scientific Software International), ConQuest (developed in part to meet the needs of large scale educational surveys as the TIMSS project), or OPLM (developed by Cito, the National Institute for Educational Measurement in the Netherlands). The latter program has an appendix called Saul that can estimate more complex linear models. The MLIRT model can be estimated by the MLIRT program (available via the Web). The testlet model can be estimated using Scoright (a product available through Educational Testing Service).

See Also the Following Articles

Item Response Theory • Maximum Likelihood Estimation • Multidimensional Item Response Models

Further Reading

Birnbaum, A. (1968). Some latent trait models. In *Statistical Theories of Mental Test Scores* (F. M. Lord and M. R. Novick, eds.), Addison-Wesley, Reading, MA.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* **37,** 29–51.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika* **46,** 443–459.

Fox, J. P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66,** 271–288.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Stat. Sinica* **8,** 647–667.

Glas, C. A. W., and van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Appl. Psychol. Meas.* **27,** 249–263.

Glas, C. A. W., Wainer, H., and Bradlow, (2000). MML and EAP estimates for the testlet response model. In *Computer Adaptive Testing: Theory and Practice* (W. J. van der Linden and C. A. W. Glas, eds.), pp. 271–287. Kluwer–Nijhoff, Boston.

Holland, P. W., and Thayer, D. T. (1988). Differential item functioning and the Mantel–Haenszel procedure. In *Test Validity* (H. Wainer and H. I. Braun, eds.). Erlbaum, Hillsdale, NJ.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika* **54,** 681–697.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psych. Monographs* **15**.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Suppl.* **17**.

Shalabi, F. (2002). *Effective Schooling in the West Bank.* Doctoral thesis, Twente University.

Wainer, H., Bradlow, E. T., and Du, Z. (2000). Testlet response theory: An Analogue for the 3-PL useful in testlet-based adaptive testing. In *Computer Adaptive Testing: Theory and Practice* (W. J. van der Linden and C. A. W. Glas, eds.), pp. 245–269. Kluwer–Nijhoff, Boston.

Structural Models in Anthropology

David B. Kronenfeld
University of California, Riverside, Riverside, California, USA

Glossary

cognitive structure A structure consisting of concepts (and/or categories of concepts) showing a meaningful set of relationships among them. "Concepts" here can refer to knowledge of factual matters, of conceptual relations, of behavioral or emotional concomitants of entities, actions or events, of action consistent with goals or values, and so forth.

collective representations Knowledge—whether abstract or embodied in action—that is constructed and held by the membership of a community, that entails some sort of coordination among community members, and that is ascribed to the community as a whole.

componential structure A semantic structure formed by a set of contrasting terms that share a root defining semantic attribute and that are distinguished from one another by contrasting values on one or more out of a set of intersecting semantic dimensions. "Mother," "father," "sister," etc. are all "blood" kin. "Mother" differs from "father" in sex; "mother differs from "sister" in generation; "father" differs from "sister" in generation and sex; and so forth.

cultural grammar A modeling of relations among form classes for some cultural entity—on the model of a linguistic grammar.

cultural structures The collective representations for some domain of knowledge or activity held by (or ascribed to) members of some community. For example, kinterm systems, ethnobotanical systems, the system of roles in a university, or how to play football.

kinship terminological system The kinship terms of a language (e.g., "father," "mother," "uncle," and "cousin") organized and analyzed as a distinct system.

section system A system of kinship and marriage in which the membership of a society is divided into opposed groups (usually seen as the defined by the intersection of cross-cutting moieties), and in which husbands and wives must come from one set of opposed groups, and in which children and their spouses belong to a contrasting set of opposed groups.

social structure Modeling of interactive relations among groups of people, where group membership is based on (or recognized in terms of) some set of collective representations.

taxonomic structure A tree- or dendrogram-shaped semantic structure formed by a hierarchy of inclusion relations (a dog is a kind of mammal, a mammal is a kind of animal, etc.) and contrast (Boxers, Bassetts, and Collies are contrasting (or opposed) kinds of dog; dogs, cows, and raccoons are contrasting kinds of mammals, etc.).

unilineal descent group Corporate (holding some property in common) kingroup consisting of the descendants of some apical ancestor in either the male line (a patrilineal descent group) or in the female line (a matrilineal descent group). The system is "segmentary" if successively wider groups are defined around a genealogically based "taxonomy" of apical ancestors.

After a discussion of what anthropologists understand by "structural models," the range of such models is described and exemplified. Different kinds of structural models vary in the topics they address, in their degree of formality, in the questions that are asked of them, and in the sources they are based on. Sources include structural linguistics, mathematics (graph theory and algebra), and computer programming. Topics include social structure in Radcliffe-Brownian Structural-Functionalism, social and cognitive structures in Levi-Straussean Structuralism, componential/paradigmatic and taxonomic structures of linguistic anthropology, the structure of marking hierarchies and implicational chains of concepts, the structure of culturally standardized decisions, and the conceptual structure of culture itself.

Meaning of "Structural Model" as Commonly Understood in Anthropology

A model reflects some literal representation (or construction) of some piece or aspect of something else—in anthropology usually some presumed reality. In this usage, a model differs from a theory since the latter is a set of formal propositions about underlying or generating entities and relations (cf. axioms) from which outcomes can be deduced—and represents a stab at an analytic understanding. Model airplanes can be literal replications (to any desired degree of exactitude) of real planes. As such, they may actually fly, and they can be used for experiments relating to many aspects of aircraft design (e.g., air flow, lift, and drag). At the same time, they are not any kind of theory of flight, and, indeed, can be made successfully in the absence of any such theory. As "models," they are simplifications, and, hence, cannot represent the full reality of the modeled aircraft. For instance, our little flight models will not tell us much about metal fatigue nor about when the wings will be too heavy for the plane (or vice versa).

Simplifying can be good when it allows us to see (or examine) things that we would not otherwise see (or be able to look at). Models are helpful to us as simplifications that we can experiment with—adding whatever complexity is necessary to address the issues we choose to address. As models become more closely and effectively tuned to "systems" they do thereby come more closely to approximate instantiations of "theories" of the systems in question. Simplified models can also be used by people to avoid having to think too hard about stuff they cannot readily articulate or do not really believe; thereby, they can serve a variety of ideological roles by artificially justifying extreme actions, making people feel good or important, allaying fears and low self esteem, and so forth.

Structure represents the systematic relationships among the entities that make up a system, that is, the relationships that govern interactions among the entities. Often "structure" is seen as a kind of skeleton on which the meat of actual behavior is mounted. In anthropology, these entities are often grouped into sets that make up the subsystems of "culture" known as a cultural domains (e.g., kinship, politics, ethnobotany). In such a set, the entities represent the effective operative units (whether of people, groups of people, concepts, or whatever) in the domain. The label "structural model" is less widely used (except in certain particular schools of work—see below), but is usually understood as referring to a representation of the structure of a domain. Typically (and perhaps prototypically) this is presented as a diagrammatic model (though algebraic representations sometimes appear). A structural model differs from a simple diagram or picture in that it represents the underlying relations among some sort of parts that are seen to govern behavior in the given domain. However, we should note that neither structure nor structural models has any well-defined and generally shared meaning within anthropology.

Sources and Kinds of Structural Models

Radcliffe-Brownian Structural Functionalism

The initial significant anthropological appeal to structure was in Radcliffe-Brownian (British) Structural Functionalism. The central idea was that a set of institutions (kinship, politics, economics, religion, ecology, etc.) had a structural organization—a system of analytic or systemic units and relations among them—around which social functioning was organized. Structure was constituted by the functionally significant units and relations among them. Social structure, then, referred to the sets of functionally significant groups of people (kin groups, communities, neighborhoods, sodalities, etc.) and the understood regularities in relations among them. These relations could include obligations for mutual assistance, gifts, or support; residential rights and obligations; or competitive access to resources, and they could derive from the nature of the group or from obligations (such as marriage) assumed by group members.

The Radcliffe-Brownian emphasis, however, was on function. There was no general move to pull out anything like what might be called structural models. Illustrative diagrams—of, say, ties among kingroups or of relations among political entities—were often used to show or explicate structural relations Malinowski, while at least as important to the British Structural Functional School and the students who came out of it, was even less concerned with structure, and was even, in some significant ways, opposed to the idea of structure.

In the first student generation, Evans-Pritchard and Fortes did offer something more like explicit models in their discussions of segmentary systems of balanced opposition (linking kinship and politics). For important examples see the work of Evans-Pritchard and Fortes.

For societies with such systems, a genealogy based on unilineal descent relations served as a political framework that defined relative degrees of affiliation among territorial/organizational units and the political and military concomitants of those affiliations. For instance, Evans-Pritchard described the patrilineal Nuer as having a system in which ego, whether male or female, belonged to his or her father's immediate family corporation and to the wider kin groups to which it was affiliated via chains of ancestral males (apical ancestors of various segments).

Inheritance, bridewealth, and homicide payments were among the activities tied to various levels of relationship. That male-line genealogy was continued up to the point at which all Nuer were related via a single genealogy. It was much like the Biblical genealogy of Abraham, Isaac, Jacob/Israel, and the apical ancestors of the various Hebrew tribes. With the genealogy came a feud system in which closer relatives unite against more distant relatives. The genealogy became a political charter via the association of territorial segments (immediate communities and the larger groups which they in turn made up) with the descendants of various apical ancestors—i.e., with segments of the encompassing Nuer genealogy. The military and juridical obligations of community members in any given local conflict were normally defined by the genealogical relations of the lineage segments with which the communities were associated, even though only a minority of any local community actually belonged to the lineage segment with which the community was associated. A system of justice was compounded out of a combination of feud obligations, places of sanctuary, and negotiations for compensatory payments. The successful operation of the system depended on opposed (intra-Nuer) segments (segments associated with the sons of a given ancestral father) being of roughly equal population and power. Such balance was maintained through a system of selective pruning of more distant ancestors from the genealogy (what has sometimes been called "structural amnesia"); the pruning was the unself-conscious effect of the fact that Nuer learned their genealogies only via their participation in activities tied to genealogical units; they only learned the names and positions of ancestors who were relevant to actual events (i.e., structurally relevant). (See Fig. 1.)

Within the Radcliffe-Brownian tradition the conception of social or political "structure" ranged from Evans-Pritchard's mental conception of it to Fortes' behavioral. That is, Evans-Pritchard (in his Nuer books) saw social structure as a mental template in terms of which the somewhat chaotic assemblages of everyday life were recognized, categorized, and understood. Fortes, on the other hand (in his Tallensi books), saw social structure as the actual recurrent groupings of actual people on the ground (a view very similar to the classic Bloomfieldian view of linguistic structure). Thus, in Evans-Pritchard's Nuer ethnography, if two men were on opposite sides of a fight, their presence there would be credited to their own descent group relations, the descent group relations of the local lineages associated with their residences (commonly different from their own lineages), or other social ties—depending on what possibilities were factually available and on how those doing the crediting saw the reason for the fight. In Fortes' Tallensi accounts, worshipping together at certain kinds of shrines was based on shared descent group membership, and a careful mapping of who

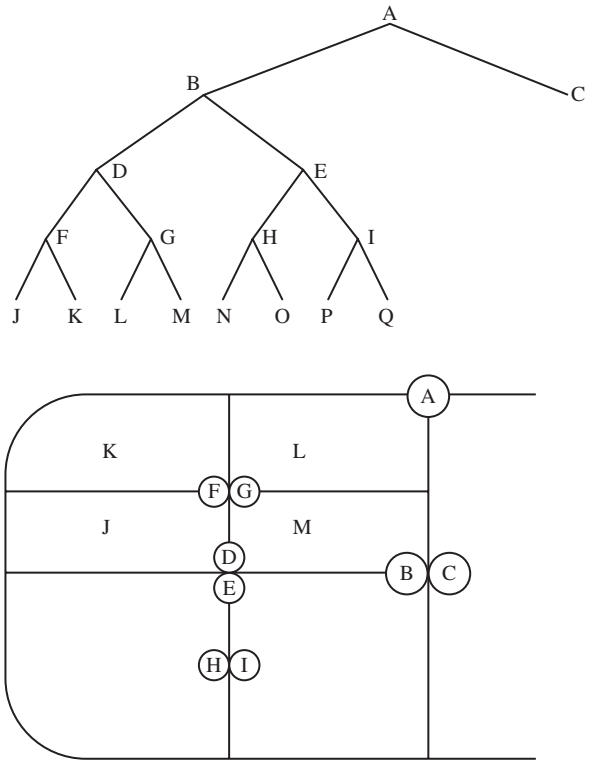


Figure 1 Nuer political structure: Evans-Pritchard's linked genealogical and territorial units (adapted from Evans-Pritchard, 1940). (Top) Geneology: A–Q are ancestors of living people; (bottom) Territory: A–Q are regions or communities associated with the descent groups descended from the given ancestors.

worshipped with whom at such shrines would directly reveal the descent group structure. Changes in such co-worshipping—and in the descent group structures apparently implied—from one year to the next accounts for most of the confusing complexities of Fortes' Tallensi accounts (vs. the great clarity with which he often wrote).

Gregory Bateson, in *Naven*, suggested the existence in different cultures of cognitive (as well as social) structures (which he modeled), even if his conception of them was somewhat abstract and thin by today's standards, and even if his conception never really caught on. Additionally, and importantly, with his brilliant development of "schismogenesis," he pioneered development of the concept of feedback, and our understanding of its role in the change over time of the shared cognitive and behavioral entities that make up culture.

Edmund Leach, before he moved in more Levi-Straussian directions, offered a new and exciting approach to the structural analysis of kinship terminologies. In his 1945 paper on Jinghpaw kinship terms (reprinted in his *Rethinking Anthropology*), he explored what would be the significant set of kinship roles and definitions, given normative rules of lineage affiliation and postmarital residence. Even though it failed to fully account for the regularities of

terminological systems, Leach's effort was a useful contribution. It in some ways prefigured later prototype-extension approaches to terminological categories, and it offered a kind of structural frame in which terminological categories, kinship roles, and kin groups could be brought together in a common treatment.

Section Systems

Another area of kinship in which structural-functionalists associated with Radcliffe-Brown produced early structural models was the "section systems" of aboriginal Australian societies. These were systems of affiliation based on aspects of descent and linked to kinterm categories. These systems constrained marriages and served as the basis of wide regional systems of putative kin (systems by which nonkin or people whose actual kin status was unknown were assimilated into kin categories for various, but not all, social, political, and economic purposes). Four-section systems eventually came to be described as, and are often pictured as, an intersection either of patri- and matri-moieties or of a moiety system with an alternating generation rule. Eight-sections systems were then seen as based on subdivisions of a basic 4 (see Fig. 2).

Structural Linguistics

An important source for contemporary anthropological understandings of structure and structural models has been structural linguistics in its Saussurean roots and in its developed Bloomfieldian/Yale/American and Prague versions.

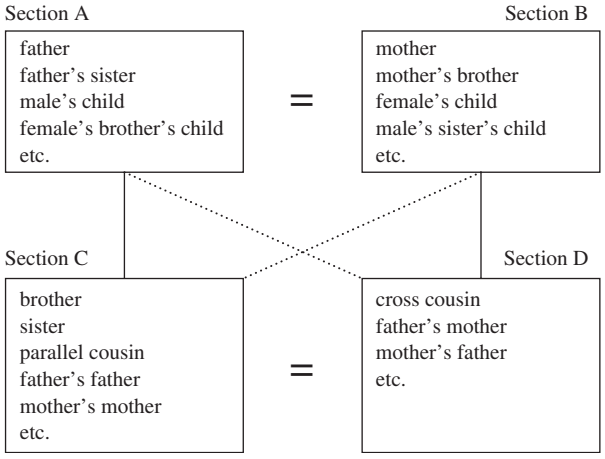


Figure 2 A Four-Section System. Each box is a section and in it are kinds of relatives of an ego (who happens to be in Section C) who fall in that section. Vertical lines between boxes (in either direction) link fathers with their children. Lines of dots between boxes (in either direction) link mothers with their children. An equal sign between boxes indicates a marriage relationship; that is, a person in the one box must take a spouse from the other. (Adapted from Gould, citation in Kronenfeld.)

Saussure's early programmatic version laid out the -analytic concepts, including relations among entities of opposition (and inclusion), a "sign" defined by the joining of a "signifier" (a "sound image," made up of cognized phonological units) to a "signified" (a concept, defined in opposition to other concepts), paradigmatic structures of opposition vs. syntagmatic structures of co-occurrence. These structures were cognitive (i.e., mental) as was his notion of "linguistic value" (function or communicative effect of linguistic entities) which was dependent on structures of opposition. But his only essay in actual model building came much earlier in his analysis of the Proto-Indo-European (PIE) vowel system ("Memoire sur le systeme primitif des voyelles dans les langues indo-europeenes," 1878), where he used an analysis of the various phonological structures of daughter languages to suggest for PIE systematic phonological elements (laryngeals). At the time no known Indo-European language had such phonological elements, but they were later discovered in ancient Hittite.

Structure in American linguistics had to do with systematic and recurring relationships (in speech) among form classes, whether in phonology, morphology, or syntax (e.g., what could be seen as stops vs. fricatives vs. etc., as first person vs. second person vs. etc., as nouns, count nouns, adjectives, noun phrases, etc.). There was some concern with the functions of different structures, but (unlike British structural functional anthropology) the focus was on the accurate and efficient elicitation, description, and presentation of the structures. Structure represented the constant framework of relationships within which variable constructions acquired their interpretability.

In the Prague School version of phonology, structure was more a matter of the formal defining elements of phonemes, their paradigmatic possibilities for opposition, and their syntagmatic possibilities for combination in one or another language. Phonological structure was tightly tied to phonological function. Within the wider range of structural linguistic approaches, the Prague School of Trubetzkoy and Jakobson was relatively heavily concerned with the structural effects of phonetic content (i.e., the physical or phenomenal world) while others such as Hjelmslev's Glossematic School saw structure more as a system of essentially arbitrary combinations or arbitrary entities.

Two major anthropological approaches to cultural structures developed out of structural linguistics, the "structuralism" of Levi-Strauss and those influenced by him and the "ethnoscience" approach (later to become "cognitive anthropology") associated initially with Ward Goodenough, Floyd Lounsbury, Harold Conklin, and Charles Frake.

Levi-Strauss

His early analytic work dealt with kinship (see Kronenfeld and Decker's 1979 article for an overview). He built on British social anthropology, but especially went back to its sources (and the locus of his own training)—Emile Durkheim, Marcel Mauss, and their colleagues. This led Levi-Strauss to a concern for solidarity of various kinds and for the systemic effects of repeated individual decisions. But, in distinction from the Radcliffe-Brownians, Levi-Strauss foregrounded structural relations, showing, or example, the systemic effects of different marriage rule and descent rule combinations and (based on the work of Audrey Richards) the cumulative social effects of different combinations of power and affect relations within a basic kinship unit (comprising a man, his sister, her husband, and the couple's child). Levi-Strauss did bring mathematics, and the idea of mathematical structure, into the picture via Andre Weil's appendix to the French edition of Elementary Structures of Kinship, but he never much followed up on mathematical aspects beyond his use of mathematical appearing formulas.

Levi-Strauss's later work on myth shifts from social anthropology to a Boas-based approach to cultural content, though with much influence from a Jakobsonian version of structural linguistics. The change came after a series of discussions with Boas and Jakobson in New York in 1941 just before Boas's death. Structure here referred to patterns or series of analogous oppositions wherein successive versions were more increasingly contained or contextualized in a way that enabled their apparent transcendence. The oppositions were found in culture content. The structures were not part of the overt surface content, but were analytically revealed as the underlying entities and relations that generated the surface content. The Levi-Straussian approach was a little like Bateson's in looking for cumulative (systemic) effects of repeated instances of a given opposition. In his mythological studies, he often found the structure of myths to be a counterfactual version of normal human events in which the negative myth outcomes were seen a reinforcing (or justifying) the social rules of the society in question.

Levi-Strauss introduced and/or underlined some basic useful distinctions among types of structural models, including that between mechanical and statistical models (of behavior), that between "native" and "anthropological" models, and that between conscious (or explicit) and unconscious (or implicit) models. He also went on to suggest linkages among the former elements vs. the latter of each of the oppositions—linkages that seem more problematic.

In the Structuralism that grew up around Levi-Strauss, structure, taken from linguistics, was taken in much more of a Praguean or Jakobsonian sense than in an American or Bloomfieldian sense. While supposedly based on Saussure, the approach significantly misconstrued the original Saussurean model—particularly in its misunderstanding of the nature of the sign (including the signifier–signified relationship, the role of speech (*parole*), and the relationship between synchrony and diachrony. "Post-structuralist" writers (for example, Bourdieu in Outline of a Theory of Practice) corrected many of the structuralist anthropology excesses, but without recognizing either Structuralism's misconstrual of its Saussurean roots or the modernity and analytic power of those roots.

A different approach to structure was that of the cultural grammar. In part modeled on linguistic grammars, but more directly deriving from Vladimir Propp's 1958 (orig. 1929) *Morphology of the Folktale*, Colby pioneered a computer analysis of Eskimo folktales in 1973, and more recently has followed up with an analytic parsing of Ixil Maya Divination (*The Daykeeper* in 1981). Frake relied more directly on American descriptive linguistic traditions in his *Structural Description of Subanun "Religious Behavior"* which, along with his article on *Notes on Queries in Ethnography*, pioneered the examination and analysis of culturally standardized cognitive structure in terms of native speaker/actor categories. Frake's approach matches well with the simulations of cognitive systems coming out of Cognitive Sciences as seen in Edwin Hutchins 1980 analysis of Trobriand Litigation in *Culture and Inference* and in Schank and Abelson's 1977 simulations of conversations about restaurants in their *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*.

Kinship Terminologies

The following discussion is based on Kronenfeld's introductory overview of kinship terminology studies, and full citations for the works discussed here can be found in the Special Issue of *Anthropological Theory* (2001). Kinship terminological systems have long been described via idealized genealogical diagrams. Such diagrams focused on an "ego" and traced down from ego through a son and a daughter, and a son and daughter of each to ego's various kinds of grandchildren, then up through ego's parents and grandparents (and maybe higher) and then down through a male and a female child (not on the chain of parents) of each ancestor to the range of collateral descendants of the apical ancestors (usually in ego's grandchildren's generation, but sometimes shallower or deeper). These idealized genealogies were used in anthropology as early as the mid-19th Century, e.g., by Dorsey in 1884, and formed part of the basis of Rivers' classic 1910 article on the "genealogical method." But such diagrams, as also the lists of genealogical positions that Morgan used in his survey schedules and Kroeber's 1909 list of significant analytic features, fell short of the idea of getting at the

underlying shapers of behavior that is implicit in the idea of a "structural model." Tax's 1955 Fox analysis came closer, but did not really contain any model, and Leach's Jinghpaw approach never caught on.

Componential or Paradigmatic Structures

It was with Goodenough and Lounsbury's adaptation in 1956 of "componential analysis" from structuralist phonology that something like genuine or full-fledged structural models were introduced into studies of kinship terminological systems. These were distinctive feature analyses in which the goal was to find the minimal set of features that were necessary and sufficient to distinguish the referents of kinterms in a given system from one another. These were attempts to model the semantic structure of kinship terminologies, and existed within a wider context of concern with analyzing semantic structure in general; the conception of structure and how it related to behavior was taken from structural linguistic analyses of phonological systems (especially Prague, but also see Zellig Harris's work within the Bloomfieldian tradition).

One result of componential work in kinship was the realization that semantics differs from phonology in important ways having to do with the function of the systemic entities, the role of features, and the degree of constraint of the relevant universe. A second result emerged when it was realized that native speaker assignments of relatives to kin categories did not depend on distinctive features (unlike the phonological case where features do govern assignments), but on a relative product "calculus" of the "he's my mother's brother, so that makes him my uncle" sort. The distinctive features found in a componential analysis of kinterms are, in fact, dependent on prior knowledge of how the relatives are related (genealogically, say); that is, they are defined in terms of kinterms, rather than vice versa. But, there exists considerable evidence that people use such distinctive features in sorting kinfolk, behaving toward them, and so forth. There still exists some question concerning the degree to which the distinction in kinship between the means by which entities are defined and the features by which those entities are associated with related thought and behavior is normal for some wider set of semantic systems—or is unique to the special domain via which people receive their basic social locations or identities.

These realizations posed basic questions about the nature of semantic structure and its modeling. Componential analyses of a densely populated and complex domain such as kinship yielded empirically powerful structural models, as shown in the classic studies of Wallace and Atkins in 1962 and Romney and D'Andrade in 1964. But, attempts to describe and formally model the system used by native speakers in their definitions led to very different kinds of structures (see Kronenfeld's 1980 analysis of Fanti for an early version, and the articles by Read and Lehman in Kronenfeld's 2001 edited collection as well as Gould's 2000 *A New System for the Formal Analysis of Kinship*, for sophisticated algebraic versions). These structures have the formal properties of algebraic structures and lend themselves to rigorous and informative graphic representations.

Taxonomic Structures

Work on folk taxonomic systems (see Berlin's 1992 book for one overview) has led to the development of tree or dendrogram models that represent successive subdivisions of a superordinate category by two or more subordinate categories. These structural models are based on the empirical delineation and concatenation of Saussurean relations of contrast and inclusion. The contrast between opposed categories sometimes seems based on contrasting values on distinctive features—making these models a kind of variant on the componential one (which is based on intersecting, cross-cutting, features); but at other times the contrasts seem to be between complex gestalts that cannot be easily represented by features.

Componential (Saussure's paradigmatic) and taxonomic structures represent two ways of building a larger structure out of relations of contrast and inclusion. A problem is that both kinds of structures—at least as developed, consistently held wholes—seem rare; much of cultural semantic classification seems looser and more *ad hoc*: cars contrast with trucks and motor bikes as kinds of motor vehicles, and all of those with bicycles, tricycles, and kids' wagons as nonmotor vehicles, but alternatively we can group our vehicles by number of wheels, by who drives/rides them, by who makes them, by the surfaces on which they are used, and so on. Berlin distinguished a "basic" taxonomy from "special purpose taxonomies (e.g., the set of answers to questions such as "what is that fish" vs. answers to questions like "what fish are caught in nets" or "what fish are most expensive"). Atran, among others, has argued that "natural kinds" (natural entities and their groupings—part of the context of our evolutionary history—such as fish and plants) are classified differently than are our cultural constructions (such as vehicles, occupational roles, and so forth, cf. potential answers to "what is that car?" [a Ford, a sedan, . . . ?]), while Kronenfeld among others has not been so sure of the usefulness or analytic effectiveness of that distinction (as opposed to one having to do with the kinds and frequency of people's interaction with items and categories in the given domain.

Algebraic Structures and Pragmatic Issues

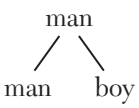
The kinds of complex algebraic structures based on relative products found to generate kinship terminologies so far seem unique to the kinship domain, but all domains seem to exhibit special properties and constraints deriving from the pragmatic natures of their construction, interaction, use, and so forth. One solution seems to be to separate narrowly semantic structures of inclusion and opposition (however they may cumulate into larger structures) from pragmatic models of the worlds to which they pertain. From this perspective, algebraic representations of kin categories pertain not to the terms' semantics but to systematic features of the conceptual world they refer to (and the social and biological worlds variously linked to it)—what categories of people get linked in what ways to what other categories. The comparable information for vehicles perhaps concerns how they are powered, how many wheels they have, what they are used for, where they are used, etc.; these constraints are looser and more flexible than are those of kinship, but certainly there exist technological areas in which the constraints are tight (for instance, rockets to space and fast submarines). In the taxonomic realm, ethnobiological classifications must accommodate both the nature of the biological world (structured as it is by Darwinian evolution) and the commonalities and contrasts associated with people's interaction with that world.

Piaget, among others, considers the ways in which the nature of the pragmatic world constrains and shapes the cognitive structures we form to represent parts of that world, and how our push to get productive mental control over wide classes of situations leads to mathematically tight representations (especially, in Piaget's view, group and lattice structures).

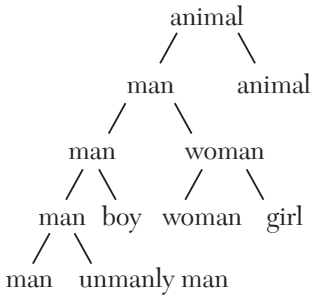
An important issue raised by Piaget's work, and posed in anthropology from a variety of theoretical perspectives, e.g., by Giddens (as "structuration"), is the degree to which structure is interactively constructed and reconstructed in an ongoing manner (vs. more rarely and maybe accidentally constructed), and then mostly more passively received. Fredrik Barth (again, among others), e.g., in his article in A. Kuper's *Conceptualizing Society* has raised the question of the degree to which cultural or social structure (systems of shared and distributed knowledge or the systems of interactions produced by that knowledge) is epiphenomenal (i.e., simply the patterns of activity produced by repeated and similar experiences and feedback experiences of separate individuals) vs. something that has some real cognitive force (an active shaper of behavior). If active, then one has to worry about how it is constructed or transmitted, that is, how it comes to be shared and collective. Constructivist approaches to cognition, such as Piaget's, are important here, but with the general addendum that the push for finding or creating structure applies also to representations of the social world and of putatively collective knowledge in addition to the individual knowledge structures mostly studied by psychologists. The cognitive structure of putative "collective representations" of the sort brought to our attention by Durkheim and Saussure is at issue here, though no good general account or theoretical model of such apparently collectively held structures yet has been proposed.

Marking Structures

The cognitive structures discussed so far all deal with intercategory relations. One insight that has come out of work on kinship, color, and plant classifications is that the relationship of exemplars to categories in ordinary usage (whether we be talking about semantic, pragmatic, or other cultural categories) is that the relationship of specific items (situations, exemplars, actions, or whatever) to the categories that we use to label them, think about them, etc. is often (even usually) not a simple matter of directly fitting the set of defining features that define or structure the intercategory relations. Instead, our cultural and linguistic categories often come with presumed (imagined, understood) prototypic exemplars, and then our recognition of an instance of the category involves assessing the relative similarity of the given instance to the prototypes of possible categories. The similarity assessment takes account of various aspects of context, including the relative plausibilities of competing categories, what is at issue in the choice, the categorizer's goals, and stake, and so forth; different kinds of classification (semantic, behavioral, etc.) will foreground different aspects of context and different kinds of purposes (communicative, instrumental, aesthetic, etc.). In one approach, the prototype represents a default referent or instance of the category; with more contextual information a different referent or instance may become a secondary default, and so forth. In linguistics Trubetzkoy developed and then Greenberg generalized marking theory—in which among a set of alternatives the default value was considered the "unmarked" option as opposed to "marked" options (which were marked by their need for additional specification). The unmarked option could represent either the generic category (ignoring the opposition at issue) or the default value of the opposition. Thus, the word "man" or "men" can be both adult and young men, but the default expectation is adult men vs. the marked alternative, "boy." A structural diagram of this relationship would be the following:



In this example, the term "man" is said to appear at two levels of contrast—one in which "man" (vs., say, "woman") includes "boy" and a more specified one in which "man" contrasts with "boy." Marking relations can be concatenated to form hierarchies, such as



Such structural models have been used by Berlin to illustrate the development of vocabulary, by Randall in his 1977 Ph.D. dissertation *Change and Variation in Samal Fishing: Making Plans to "Make a Living" in the Southern Philippines* to define the structure of activity choices by Samal fisherman, and by myself (in current work) in attempts to specify one kind of interrelation among cultural models based on their degrees of specificity. Hage (see, for example, his 1999 article on "Linguistic Evidence for Primogeniture and Ranking in Proto-Oceanic Society") has used such models to describe the historical development within language families of kinterm categories.

Mathematical Models

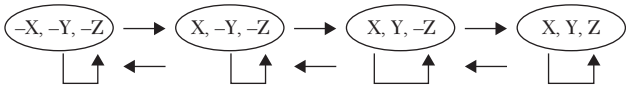
Algebraic models of kinship structures have already been discussed above. In many parts of the world outside of anthropology "structure" or "structural model" refers typically to a mathematically described (though commonly not quantitative) regularities. Mathematical models of structure have played a role in anthropology, even if not yet any really central role. An early example was Weil's previously mentioned 1949 appendix to Levi-Strauss's *Elementary Structures of Kinship* in which he used the theory of permutation groups to construct an algebraic analysis of some types of marriage laws (particularly of the sort related to Australian section systems. Harrison White's *An Anatomy of Kinship: Mathematical Models for Structures of Cumulated Roles* in 1963 dealt with a much wider range of what have been called "prescriptive marriage systems," that is, systems in which a person's spouse is supposed to come from a particular category of kin. These are mostly systems that have been analyzed (especially since Levi-Strauss's work) as being built on marriage alliances among unilineal descent groups—Levi-Strauss's "elementary" structures of kinship. Mathematical models of such structures continue to be a topic of research (see, e.g, Gould's *A New System for the Formal Analysis of Kinship*, and various works of Tjon Sie Fat's, both cited in the introduction to Kronenfeld's 2001 edited collection). There has been no comparably broad interest in mathematically modeling Levi-Strauss's analytic approach to the cultural content of cognitive systems (i.e., his analyses of myth and other symbolic systems), though El Guindi and Read's 1979 article on "Mathematics in Structural Theory" in *Current Anthropology* represents one early attempt at such modeling.

Graph Theory

Graph theory is a branch of mathematics. In both the simple sense of using graphs to represent theoretical or structural relations and in the more sophisticated sense of making use of the axiomatic analytic machinery, graph theory has provided a useful tool for forming, viewing, and analyzing a variety of kinds structural models in anthropology. Hage and Harary's 1996 *Island Networks: Communication, Kinship and Classification Structures in Oceania* provides a particularly useful overview of its anthropological relevance, and Flament's overview of applications remains good. Directed graphs ("digraphs," graphs in which the points are connected with lines that go in only one direction) were used by Greenberg in his work on substantive universals in language to describe the state transitions (e.g., addition or deletion of distinctive features in a phonological system) that were possible, given a set of interrelated implicational universals. An implicational universal was recognized in a large-scale cross-language comparison when the 2 by 2 table formed by the comparison of the presence or absence of X with the presence or absence of Y had a "zero," i.e., empty, cell (given numbers in the other cells adequate for statistical significance). For example

		Y: +	Y: −
X	+	cases	cases
	−	no cases	cases

means that Y can only occur when X occurs, or, in logical terms, $Y \rightarrow X$ (or "the presence of X is a necessary but not sufficient condition for the presence of Y"). A chain of implicational universals among defining features of some linguistic or cultural system—as, e.g., $Z \rightarrow Y \rightarrow X$ could be graphically represented as follows—where the universe is systems of the given sort and the different states are different bundles of defining features.



The arrows represent potential changes from the state "behind the arrow" to the state at the arrow's head; changes can (logically) consist in the addition of a feature, the loss of a feature, or no change. Note that −X,−Y,Z −X,Y,Z X,−Y,Z −X,−Y,Z are all logically possible states, but unreachable under the observed empirical constraints.

Greenberg's linking of zero cells in empirical codistributions to implicational relations and then representing chains of those relations in a directed graph has been taken up by several anthropologists to form structural models of aspects of particular cultures or of intercultural comparisons. Among important examples are the following. Berlin and Kay use the approach in their classic 1964 cross-language study *Basic Color Terms* of the evolution of color terminologies. D'Andrade in 1976 used the technique on data representing the cooccurrence of various attributes in informant characterizations across a variety of illnesses within each of two speech communities. That representation nicely summarized a wide set of diagnostic and disease-development sequences even if it did not speak to the folk theories that underlay them, and the comparison between the diagrams revealed interesting differences between the two communities. Burton, White, and Brudner (in their 1977 *American Ethnologist* article on "A Model of the Sexual Division of Labor") applied the approach in a cross-cultural study of features that shape the sexual division of labor—comparing their findings with earlier ones by Murdock *et al*. using correlational measures. Their implication findings were powerful and provided substantial support for theoretical assumptions relating cognitive economy to extreme case differences in cultures' exposure of men vs. women to risk. Hage (e.g., in his 1998 "Proto-Polynesian Kin Terms and Descent Groups") has applied the approach to the study of historical changes in the defining features of kinterminological systems within a single language family.

Network Structures

Graph theory has provided a particularly powerful and useful way of modeling networks and network related phenomena (see Wasserman and Faust for an overview). White (with Jorion in their 1992 "Representing and Analyzing Kinship: A Network Approach" in *Current Anthropology*, with Houseman in Houseman and White's 1998 "Taking Sides: Marriage Networks and Dravidian Kinship in Lowland South America," and with Denham) has made creative and effective use of network models to investigate and solve some classic problems concerning the empirical fit of marriages to a proposed marriage rule and the empirical interrelationship of descent, marriage, and demographic variables in kinterm categories.

Computational Models

Our categories overlap, and most of the algebraic, implicational, and network models described above have significant computational (in its now common sense of computer implemented) components, as in Read's kinship terminology analysis. In many cases, the computational component allows the model implicit in a body of empirical data to be pulled out inductively via a computer program designed to recognize and cumulate the relevant kinds of regularities, as in White's work on marriage patterns. We have also seen some directly computational structural models of, e.g., the fundamentals of society itself as an emergent system (Kronenfeld and Kaus's 1993 "Starlings and Other Critters: Simulating Society" in the *Journal of Quantitative Anthropology*), of urban produce markets (in Plattner's contribution "Economic Decision Making of Marketplace Merchants: An Ethnographic Model"), kinship and demography (Read's 1998 "Kinship Based Demographic Simulation of Societal Processes" in the on-line *Journal of Artificial Societies and Social Simulation*), and cultural systems for ecological management (Lansing *et al*.'s 1998 "System-Dependent Selection, Ecological Feedback and the Emergence of Functional Structure in Ecosystems" in the *Journal of Theoretical Biology*).

The topic of computational models of structure in anthropology extends considerably beyond sociocultural anthropology to a moderate but significant set of such models in archaeology (particularly regarding the developmental of regional economic, social, and political systems in relation to ecological and demographic conditions) and a rich and large set in biological anthropology (for instance, of patterns of migration, demographic change, and gene flow). A classic overview of models in human population biology is found in the work of Harrison and Boyce, which includes chapters such as "Migration, Exchange, and the Genetic Structure of Populations" by the editors and "Genetic Implications of Population Breeding Structure" by W. J. Schull. More recently, Fix has reviewed population genetics models in relation to human migration and shown how more complex evolutionary models may be constructed and evaluated using computer simulation. In archaeology, one interesting approach is that of Reynolds *et al*.; the Proceedings volume in which it appears contains much else that is relevant to computational models in anthropology and related disciplines.

The increasing use of formal models of structural relations in anthropology (both computer program based and directly mathematical) and the increasing sophistication and empirical relevance of these models suggest a significant and important role for them in the future.

See Also the Following Articles

Aggregation • Cross-Cultural Data Applicability and Comparisons • Cultural Consensus Model • Graph Theory • Qualitative Analysis, Anthropology

Further Reading

Berlin, B. (1992). *Ethnobiological Classification.* Princeton University Press, Princeton, NJ.

Bloomfield, L. (1933). *Language.* Holt, New York.

Evans-Pritchard, E. E. (1940). *The Nuer.* Oxford University Press, London.

Fix, A. G. (1999). *Migration and Colonization in Human Microevolution.* Cambridge University Press, Cambridge, UK.

Flavell, J. (1963). *The Developmental Psychology of Jean Piaget.* Van Nostrand, New York.

Fortes, M. (1953). The structure of unilineal descent groups. *Am. Anthropol.* **55,** 17–41.

Gladwin, C. H. (ed.) (1984). Special issue: Frontiers in hierarchical decision modeling. *Human Org.* **43**(3).

Harrison, G. A., and Boyce, A. J. (1972). *The Structure of Human Populations.* Clarendon, Oxford.

Kronenfeld, D. B. (ed.) (2001). Special issue: Kinship. *Anthropol. Theory* **1**(2).

Kronenfeld, D. B., and Decker, H. (1979). Structuralism. *Ann. Rev. Anthropol.* **8,** 503–541.

Levi-Strauss, C. (1949). *Les Structures Elementaire de la Parente.* Universitaires de France, Paris.

[See Needham, R. (ed.) (1969). *The Elementary Structures of Kinship*. Translated by J. H. Bell and J. R. von Sturmer (transl.), revised. Beacon Press, Boston.]

Levi-Strauss, C. (1967). *Structural Anthropology.* C. Jacobson and B. G. Schoepf (transl.) Doubleday, New York.

Radcliffe-Brown, A. R. (1952). *Structure and Function in Primitive Society.* The Free Press, Glencoe, IL.

Reynolds, R. G., Lazar, A., and Kim, S. (2002). The agent based simulation of archaic states. In *Social Agents: Ecology, Exchange, and Evolution—Proceedings of the Agent 2002 Conference on Social Agents: Ecology, Exchange, and Evolution*. Argonne National Laboratory, Argonne, IL.

Saussure, F. (1916/1959). *Cours de Linguistique Generale—A Course in General Linguistics* (C. Bally, A. Sechehaye, and A. Reidlinger, eds.) (W. Baskin, transl.). Philosophical Library, New York.

Trubetzkoy, N. S. (1969). *Principles of Phonology.* University of California Press, Berkeley, CA. [Translated from (1958) *Grundzüge der Phonologie*. Vandenhoeck & Ruprecht, Göttingen.]

Wasserman, S., and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, UK.

Survey Design

Theodore C. Wagenaar
Miami University, Oxford, Ohio, USA

Glossary

anonymity When a respondent's identity is not known.
case–control designs Designs used to compare two groups, one of which is involved with the issue of interest.
confidentiality When the respondent's identity is known, but the researcher promises not to reveal it.
contingency question A question that is asked only of those who have given a particular response to a prior question.
cross-sectional designs Survey designs that are completed at one point in time.
exhaustive Item responses that provide an appropriate response for every respondent.
generalizability The extent to which survey results can be applied to the larger population from which a sample was taken.
longitudinal designs Survey designs in which data are collected at multiple points in time.
mutually exclusive Item responses that allow respondents to fit in only one response category.
random digit dialing A strategy for doing telephone interviews that involves dialing computer-generated telephone numbers.
reliability The feature a measure has if repeated measures yield the same results.
respondent Someone who responds to a questionnaire or interview.
response set The tendency for respondents to mark the same response to a series of items.
specifications Instructions for interviewers regarding their responses to ambiguous situations.
validity The feature a measure has when it accurately measures what it is intended to measure.

Survey design helps researchers reach large numbers of respondents easily. Variations of survey design enable researchers to use various random sampling designs, which enhance the generalizability of the results. Questionnaires and interviews comprise the two basic approaches, each with its own advantages and disadvantages. Questionnaires tend to be cheaper and easier to complete, whereas interviews require trained interviewers and take longer to complete. Items in surveys can be structured to give respondents specific responses, or they can be unstructured to elicit responses in the respondents' own words. Researchers have developed specific strategies for conducting effective questionnaires and interviews that have been found to enhance the reliability and validity of the results. The various questionnaire and interview designs each have their own advantages and disadvantages. Researchers are increasingly using online surveys to simplify the data-gathering process.

Utility of Survey Design

Survey design has distinct benefits when compared with other designs commonly used by social scientists, such as experimental and observation designs. Surveys enable researchers to use large random samples to gather data on many variables from many respondents. Surveys can be completed in relatively short periods of time. They can be used to describe something in a population, to explore a topic that may be studied in more depth later, and to develop causal models of attitudes and behaviors. Surveys can be used to solicit individuals' responses about themselves, other people, organizations, neighborhoods, and other units of analysis.

Surveys can be cross-sectional or longitudinal. Cross-sectional studies are done at one point in time, whereas longitudinal studies are done at multiple points in time. Both types of designs can be used to describe, explore, or explain concepts of interest. Longitudinal studies are better for establishing causality because the time order of variables is more clear. Longitudinal studies can be

accomplished with three approaches. The trend study simply compares populations at multiple points in time. For example, support for abortion can be compared over time. The cohort study follows a specific group of people over time, although the same individuals are not necessarily studied. For example, a random sample of first-year students can be studied and 1 year later a random sample of sophomores can be studied, with similar designs for juniors and seniors at subsequent time points. Finally, the panel study follows the same individuals over time. This design is superior to the cohort design because the researcher can identify changes, and hence causality, more definitively. The disadvantage of the panel design is panel attrition, which occurs when individuals withdraw from the study over time. The cohort design addresses this weakness by drawing samples of the same size over the 4 years of college, for example. Researchers can approximate longitudinal approaches when using cross-sectional studies by asking retrospective questions. For example, respondents could be asked how often they participated in high school extracurricular activities as well as how many community groups they belong to currently if the goal is to assess a possible causal link between these two variables.

Surveys are particularly useful with case–control designs. Researchers sometimes wish to learn why one group has a particular experience and another does not. For example, a survey could be designed to ask retrospective and other questions to learn why some new mothers experienced postpartum depression while others did not. The case–control design requires that the researcher be able to identify ahead of time those with the experience and those without.

Self-Administered Questionnaires vs Interviews

Both self-administered questionnaires and interviews are used in survey research. Interviews can be conducted in person or with the telephone. Self-administered questionnaires are generally cheaper because interviews typically involve training costs and in-person interviews involve travel time. It also takes longer to execute an interview than it does for a respondent to complete a self-administered questionnaire. Response rates are typically higher with interviews, often approximately 70–80%, whereas self-administered questionnaires often have response rates less than 50%. It is more difficult for respondents to refuse an interview because of the personal attention given to the respondents, whereas it is easy for respondents to discard surveys received in the mail. The presence of an interviewer helps reduce the uncertainty that respondents may experience with some items, thereby increasing the validity of the results. This feature also makes interviews more flexible. An easily misunderstood item, for example, can be reworded before completing additional interviews. After such an item appears in a self-administered questionnaire, it cannot be fixed. Some people argue that telephone interviews yield more honest answers because respondents are less inhibited when they do not have to look directly at an interviewer. Others argue that people are more suspicious of telephone interviewers and may give less honest answers. Interview studies usually involve many interviewers, who may each introduce their own biases into the data-gathering process. Reliability is therefore greater with self-administered questionnaires.

Interviews require more skill because the interviewers must be trained; it takes less skill to stuff questionnaires into envelopes and score the results. Self-administered questionnaires can be sent to hundreds or thousands of people quickly, but interview studies generally involve smaller samples because each respondent must be interviewed. Telephone interviews can reach more people than do in-person interviews, but issues of interviewer and respondent safety often affect response rates. Self-administered questionnaires can be sent easily via the mail, whereas face-to-face interviews limit the geographic area of samples. Anonymity is easily maintained with self-administered questionnaires, but it is more difficult to maintain with interviews. Instead, researchers typically offer respondents confidentiality when doing interviews—a promise that the respondent's identity will not be revealed to others. Interviews can be done in cases in which respondents cannot read or write, such as when interviewing preschoolers. Because self-administered questionnaires require literacy, their use is limited in populations with limited reading ability (approximately one-fifth of the U.S. population is functionally illiterate). The atmosphere is critical when doing interviews. For example, the presence of a spouse may affect a respondent's responses to questions about marital happiness. Such distractions are less intrusive for respondents completing self-administered questionnaires, either by mail or via the Internet.

Structured vs Unstructured Formats

Items in surveys can be structured or unstructured, also known as closed-ended and open-ended formats, respectively. The structured approach includes fixed responses to an item. For example, "woman" and "man" are the fixed responses to the question "What is your gender?" The unstructured approach lacks such fixed responses. "What do you see as the three major

problems facing our country today?" is an example of an unstructured question.

The structured approach has several advantages. It is more standardized and hence easier to execute. The results are easier to analyze. Such items provide a frame of reference. Instead of asking an open-ended question about frequency of church attendance, for example, a closed-ended item provides a framework for responding by providing such responses as "once a year or less" and "several times a year." The structured approach reduces the likelihood that interviewers will introduce their own biases. The structured approach also has several disadvantages. Closed-ended questions sometimes force respondents into stating an opinion when they really have none on a particular issue. Such items may overlook possible responses. An item asking why students dropped a course, for example, may list various common reasons pertaining to the professor or the workload but may overlook the possibility that the student simply switched sections of the course.

The unstructured approach also has advantages and disadvantages. Open-ended items may help the researcher determine if the respondent is telling the truth or knows what he or she is talking about. Such items may be more appropriate when the goal is an intensive study of attitudes. Researchers often employ an unstructured approach in the exploratory phase of a research study to help formulate relevant hypotheses and measures. The unstructured approach can help identify relevant possible response alternatives for a question that may later be used as a structured item. Some researchers follow a grounded theory strategy by letting data emerge for subsequent theoretical analysis; the unstructured approach is central to this strategy. The unstructured approach also has difficulties. Perhaps most salient is the difficulty researchers experience with data analysis. Pages and pages of responses to open-ended questions pose unique analysis difficulties, even with the help of computer programs. This method also requires considerably more time than the structured approach. If using an unstructured self-administered questionnaire, respondents must be able to write.

Effective Self-Administered Questionnaires

Effective self-administered questionnaires are attractive, easy to complete, arranged in a logical manner, and include specific directions. Postage-paid envelopes should be included if a mail survey is used, and specific response instructions should be included if the survey is completed on the Internet. A well-designed cover letter should be included that explains the purpose of the study, reflects human subjects guidelines, and informs the respondent if the results will be confidential or anonymous. One way to maintain anonymity but still keep track of who responds is to include a separate postcard addressed to the researcher indicating that the survey has been completed. A good cover letter also underscores the importance of participation, provides a deadline and instructions for returning the questionnaire, tells respondents that there are no right or wrong answers, and thanks the respondents. Effective cover letters also indicate who is doing the survey and include a general statement about how the respondent was selected. The researcher may wish to offer respondents a summary of the results. Respondents should be encouraged to leave blank those items that they believe they cannot complete, to contact the researcher with questions, and to offer comments if they wish.

In order to improve your questionnaire, show a draft to experts in the field as well as to people who are like those who will receive the survey. The first strategy will help enhance the validity of your measures, and the second strategy will help identify troublesome items. Be sure to ask pretest respondents to comment on the difficulties they see in the survey. Such strategies will help identify the problems with the following type of item: "To which social class do you belong?" This item could be interpreted to mean "Which social class best characterizes you?" as well as "You may not be in it, but in which class do you belong?" Also, avoid double-barreled questions, which combine two (or more) questions into one question, such as "How satisfied are you with your working conditions and wages?" If someone gives a low satisfaction score, you will not know if that person is dissatisfied with working conditions, wages, or both. Develop items that respondents are able to answer. For example, respondents are not likely to remember if their first day at kindergarten was happy or traumatic. Avoid negatives in items to help minimize confusion. For example, the item "I am not satisfied with my working conditions" may inadvertently lead respondents to overlook the word "not" and select "strongly agree" to indicate high satisfaction.

Experts generally recommend that items in a questionnaire do not all reflect a "positive" or "negative" view because this situation can lead to response set, the tendency for respondents to not read items carefully and to mark the same response for a series of items. Consider the following two items: "Women who work, either full-time or part-time, outside the home should have help from their husbands in doing the housework" and "It is more important for a woman to help her husband in his career than to develop her own." A response of "strongly agree" to the first item would generally parallel a response of "strongly disagree" to the second item. The use of items reflecting alternating views of an issue can help identify those respondents who did not take the survey seriously and simply marked the same responses to all items.

Sometimes, one may wish to ask questions only if a particular response is given to a prior question; these are known as contingency questions. For example, one may first ask if the respondent graduated from college. If a "yes" response is given, the respondent can then be directed to additional questions on major, grade point average, and the like.

Avoid biasing items by eliminating words and phrases that may lead the respondent to answer in a particular way. For example, asking "Do you agree with the president that welfare support should be limited?" may bias respondents by associating the issue with the president. Leading questions and phrases should also be avoided, such as "don't you agree with the president that welfare support should be limited?" Emotional words, such as "absurd" or "completely wrong," should be avoided. Where items are placed may affect responses. For example, placing items on how overpopulated the world is before items on birth control usage may affect respondents' feelings about the latter issue. Socially desirable items should be used with caution. Such items generally yield very high levels of agreement or disagreement and, therefore, contribute little to explanatory analyses. Examples include "Are you basically a warm and loving person?" and "A world at peace is desirable."

Item responses should be mutually exclusive and exhaustive. Mutually exclusive responses enable respondents to fit in only one response. For example, responses of "Protestant," "Catholic," "Jewish," "Lutheran," "other," and "none" are not mutually exclusive because someone who is Lutheran could fit in two categories. Exhaustive responses enable all respondents to fit into a response. For example, those with no religion would have no response to select if the response of "none" were missing from the responses in the previous example. Use questions asking for raw numbers sparingly. Respondents are not likely to know the exact response to a question such as "How much did you earn last year?" or "How many movies did you see last year?" Instead, provide responses with ranges, such as "less than $10,000" and "$10,000 through $20,000." Avoid the use of responses such as "rarely," "sometimes," and "often" because respondents will interpret these words differently. If you believe there is a need for a question asking for raw numbers for a response, insert the word "approximately" in the item so that respondents do not feel obligated to remember the exact number.

Follow-ups can increase the response rate notably. Without follow-ups, the typical mail questionnaire will generate a response rate of approximately 40%. The rate increases to approximately 55% with one follow-up and approximately 60% with two follow-ups. It is critical to include new copies of the questionnaire and new postage-paid response envelopes with follow-ups if doing the survey by mail. It is best to send the first follow-up approximately 2 weeks after the deadline noted in the cover letter of the original survey and to send the second follow-up approximately 2 weeks after the deadline noted in the cover letter of the first follow-up survey.

Several factors may affect response rates. Inducements can help increase response rates, although amounts less than $1 seem to have little effect. One sociologist taped two pennies to the top of the cover letter and noted in large letters, "We want your two cents worth, and are willing to pay you for it." The disadvantage of this approach is that respondents may feel insulted because they may think that their time is worth more than 2¢. Other social scientists offer to contribute $1 to one of several charities; the respondent can select the charity. The nature of the respondents may also affect response rates. Generally, higher response rates are obtained from more educated, middle-class populations. Sponsorship is also important; a self-administered questionnaire that has a cover letter on a university letterhead will be seen as more legitimate than one that has a post office box return address. In addition, length is important; self-administered questionnaires with three or fewer pages generate a higher response rate than do longer questionnaires. An effective cover letter helps generate interest. Questionnaires on interesting topics will generate higher response rates than those on more mundane topics, and questionnaires on topics of interest to the respondents will generate even higher response rates.

It may be relevant to compare respondents with nonrespondents to help estimate response bias. This can be done only on factors for which one has prior knowledge. Nonrespondents in a survey of students, for example, could be compared with respondents on residency (on campus or off) and class level but not on how much they study. A response rate graph plotting responses received per day may help estimate any biasing effect of historical events. For example, a campus survey on racial attitudes may be affected by a particular racial incident that occurred on campus. Comparing those who responded before and after the incident will help estimate this effect.

Various mailing options exist for sending out self-administered questionnaires and receiving them back. There is debate about whether to use bulk mail or postage stamps for the outgoing survey. Some believe that a brightly colored commemorative stamp will help distinguish the survey from junk mail sent by bulk mail. On the other hand, it is much cheaper to use bulk mail if the sample size is relatively large. Postage-paid reply envelopes should be used. Respondents may remove and use postage stamps on reply envelopes for personal use. Business-reply mail helps solve this problem and is generally cheaper (even with the postage surcharge) due to survey response rates. When a high response rate for a survey is anticipated, it is less expensive to use first-class

postage, and it is less expensive for low-response surveys to use business-reply mail.

Effective Interviews

Conducting effective interviews requires comprehensive training and practice. The process is more than just reading items on the interview schedule. Interviewers should do several practice interviews, and these should be videotaped. The tapes should then be reviewed with a supervisor to elicit strengths and weaknesses. The researcher can help by providing specifications for the interview, a set of clarifying comments embedded within the instrument to help the interviewer respond to difficult situations. For example, if an item asks, "Has support for the police in your area increased or decreased?" the specifications can make it clear that a volunteered response of "stayed the same" is acceptable.

Interviews should be scheduled. Respondents are more likely to decline the interview if an interviewer shows up at the door or calls without prior notification. It also helps to send a letter on official letterhead informing the prospective respondent of the study, the sampling procedures used, the importance of the respondent's participation, and names of contact persons in case of questions. Interviewers should present appropriate identification from the sponsoring institution and should be coached in how to deal with refusals. Those doing in-person interviews should dress appropriately. Generally, this means that they should be dressed as well as or better than the potential respondents. Research shows that appearance affects credibility. When interviewing strangers, well-dressed interviewers enhance the likelihood that respondents will agree to the interview.

Interviewers should be very familiar with the instrument so that they can do the introduction and the first few questions almost from memory. This practice will also enhance rapport and credibility, and it will enable interviewers to respond quickly and effectively to questions that the respondents may ask. This practice will also help interviewers skip inappropriate questions for particular respondents. Interviews must be administered in a consistent fashion. This is particularly important because multiple interviewers may otherwise each introduce their own biases into the process. Consistent interviews also yield greater accuracy of results.

Interviewers should be trained to look and listen for nonverbal cues. Such cues can help assess the honesty of responses and can indicate the need for item clarification. Probing skills can help generate more accurate and thorough responses. Various probing techniques can be used. A brief assertion of understanding and interest will make respondents feel more comfortable. Sometimes, simply waiting a moment for the respondent to formulate a response is necessary, particularly for open-ended questions. Allowing insufficient time to respond is a common mistake made by new interviewers, and they should be reminded to not let silence bother them. Occasionally, it helps to repeat the question, which gives the respondent more time to respond and reminds the respondent of the question (many respondents are reluctant to ask that items be repeated). Another probe is to simply repeat the respondent's words, particularly if the response was very brief. Doing so encourages the respondent to expand on the response. Finally, the interviewer can simply ask for clarification when needed. It is important to record responses exactly and to not simply assume that one understands a particular response. For example, the response, "All politicians are crooked," may carry different meanings, and the interviewer should clarify if this means that politicians are bribed, do not represent their constituency adequately, or something else.

The police should be informed when doing in-person interviews. This practice may help reduce the likelihood of being interrogated when an interviewer is sitting in a car jotting down notes on an interview. It may also help convince a respondent to participate in the survey. I once helped conduct a market interview on banking institutions and knocked on the door of a bank vice president. He assumed that I was gathering competitive information, so I encouraged him to call the police department to verify my identity. He did, and the interview was completed. In-person interviews should be done only during daylight hours because respondents are less likely to participate after dark.

Interviewers should remain neutral so that responses are not biased. Interviewers should be reminded that respondents may respond according to the image that they perceive that the interviewer has of the respondent. Hence, neutrality is important. Interviewers should retain the upper hand and not provide advice, act judgmentally, or become emotionally involved. Respondents will occasionally ask interviewers about their own views on the survey items. It is best if interviewers not do so, but if they believe that they must to facilitate the interview, the interviewers' opinions should not be shared until the interview has been completed. Social scientists may feel a need to be helpful when conducting interviews, particularly if the topic under study is a sensitive one. One sociologist who interviewed gay men in-depth felt the need to offer some form of reciprocity and compiled a list of gay-friendly professionals in various fields to offer respondents.

Respondents should be isolated from others, which may be difficult with telephone interviews. The presence of others may influence responses. With telephone interviews, the interviewer may ask if another time would be better. With in-person interviews, perhaps another room could be used that would minimize interference.

Respondents should be assured at the outset that the survey has no right or wrong answers, that the results will be held confidential, and that they are free to omit items or to withdraw from the interview at any time. It may be helpful to engage in a few icebreakers as a transition to the interview, but one should avoid anything that may bias the survey responses.

Comparison of Survey Approaches

Surveys can be completed using mailed or hand-distributed questionnaires, telephone interviews, face-to-face interviews, or online. Data collection for mailed questionnaires requires approximately 10 weeks, whereas telephone interviews can be completed in a relatively short period of time if enough interviewers are used. Because of travel time, in-person interviews take longer and limit the geographical coverage. Mailed questionnaires and telephone interviews theoretically have no geographic boundaries, an advantage when considering generalizability. People seem more willing to respond to longer surveys when administered as an interview. Mailed surveys should generally be simple in design, whereas interviews can be more complex. Question ordering is highly constrained in questionnaires, whereas interviews allow for more variability in question order. Open-ended questions generally receive more complete responses when used in interviews. The greater rapport afforded by in-person interviews yields opportunities for more extensive questioning.

One advantage of mailed questionnaires is that respondents can take the time to consult their personal records. For example, a questionnaire may ask about the cost of property taxes or whether children have been vaccinated. Questionnaires may also be better for asking about sensitive or embarrassing topics, such as trouble with the law. The lower response rates to questionnaires generate greater response biases, such that those surveyed are less representative of the population, whereas in-person interviews generate higher response rates and lower response biases. In-person interviews allow the researcher to present charts listing alternative responses that are less easily communicated in telephone interviews. Both types of interviews have the distinct advantage of application to less literate populations.

Online surveys pose several advantages and disadvantages. Perhaps the strongest advantage over the questionnaire design is that the survey can be tailored to each respondent. If a respondent indicates that he or she has no children, for example, the online survey will simply skip subsequent questions about children. A questionnaire, on the other hand, will take up space with questions about children and ask the respondent to skip them. The skip patterns embedded in online surveys can also be much more complicated than those employed by a human interviewer. Perhaps the weakest aspect of online surveys is limited generalizability. Many people do not have online access, and those who do tend to be more educated and of a higher social class than those who do not.

The choice of survey method is largely dictated by the nature of the population to be sampled and the nature of the research instrument. Time and resources available are also factors. For example, random digit dialing may be employed to survey a specific population without the need to consult a telephone directory. This procedure employs computers to dial randomly selected phone numbers, including both listed and unlisted numbers.

Conclusion

Survey design lies at the heart of social science research. It has been used successfully for decades in such notable studies as the General Social Survey. Survey research yields the broadest reach in terms of sampling and generalizability, particularly when compared to experimental and observational research. Methodologists have made considerable headway in operationalizing concepts, and the survey design has contributed to this progress. Survey research affords comprehensive and efficient measurement of many concepts and has laid the groundwork for the well-developed multivariate causal analyses so popular in the past few decades. Survey research has also been the basis for numerous policy decisions, ranging from the organizational to the national level.

See Also the Following Articles

Confidentiality and Disclosure Limitation • Interviews • Longitudinal Cohort Designs • Reliability • Sample Design • Surveys • Validity Assessment

Further Reading

Babbie, E. (2004). *The Practice of Social Research.* 10th Ed. Wadsworth, Belmont, CA.

Czaja, R., and Blair, J. (1996). *Designing Surveys: A Guide to Decisions and Procedures.* Pine Forge Press, Thousand Oaks, CA.

de Vaus, D. (2002). *Surveys in Social Research.* 5th Ed. Routledge, London.

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method.* Wiley, New York.

Fink, A. (2001). *How to Design Survey Studies.* 2nd Ed. Sage, Thousand Oaks, CA.

Fink, A. (2003). *The Survey Handbook.* 2nd Ed. Sage, Thousand Oaks, CA.
Fowler, F. J., Jr. (1995). *Improving Survey Questions: Design and Evaluation.* Sage, Thousand Oaks, CA.
Gillham, B. (2000). *Developing a Questionnaire.* Continuum, New York.
Nesbary, D. K. (2000). *Survey Research and the World Wide Web.* Allyn & Bacon, Boston.
Newman, I., and McNeil, K. (1998). *Conducting Survey Research in the Social Sciences.* University Press of America, Lanham, MD.

Survey Questionnaire Construction

Elizabeth Martin
U.S. Census Bureau, Washington, DC, USA

Glossary

closed question A survey question that offers response categories.
context effects The effects that prior questions have on subsequent responses.
open question A survey question that does not offer response categories.
recency effect Overreporting events in the most recent portion of a reference period, or a tendency to select the last presented response alternative in a list.
reference period The period of time for which a respondent is asked to report.
response effects The effects of variations in question wording, order, instructions, format, etc. on responses.
retention interval The time between an event to be remembered and a recall attempt.
screening questions Questions designed to identify specific conditions or events.
split sample An experimental method in which a sample is divided into random subsamples and a different version of a questionnaire is assigned to each.
standardized questionnaire The wording and order of questions and response choices are scripted in advance and administered as worded by interviewers.

Questionnaires are used in sample surveys or censuses to elicit reports of facts, attitudes, and other subjective states. Questionnaires may be administered by interviewers in person or by telephone, or they may be self-administered on paper or another medium, such as audiocassette or the Internet. Respondents may be asked to report about themselves, others in their household, or other entities, such as businesses. This article focuses on construction of standardized survey questionnaires.

The utility of asking the same questions across a broad group of people in order to obtain comparable information from them has been appreciated at least since 1086, when William the Conqueror surveyed the wealth and landholdings of England using a standard set of inquiries and compiled the results in the "Domesday Book." Sophistication about survey techniques has increased vastly since then, but fundamental insights about questionnaires advanced less during the millennium than might have been hoped. For the most part, questionnaire construction has remained more an art than a science. In recent decades, there have been infusions of theory from relevant disciplines (such as cognitive psychology and linguistic pragmatics), testing and evaluation techniques have grown more comprehensive and informative, and knowledge about questionnaire design effects and their causes has cumulated. These developments are beginning to transform survey questionnaire construction from an art to a science.

Theoretical Perspectives on Asking and Answering Questions

Three theoretical perspectives point toward different issues that must be considered in constructing a questionnaire.

The Model of the Standardized Survey Interview

From this perspective, the questionnaire consists of standardized questions that operationalize the measurement constructs. The goal is to present a uniform stimulus to respondents so that their responses are comparable. Research showing that small changes in question wording or order can substantially affect responses has reinforced the assumption that questions must be asked

exactly as worded, and in the same order, to produce comparable data.

Question Answering as a Sequence of Cognitive Tasks

A second theoretical perspective was stimulated by efforts to apply cognitive psychology to understand and perhaps solve recall and reporting errors in surveys of health and crime. A respondent must perform a series of cognitive tasks in order to answer a survey question. He or she must comprehend and interpret the question, retrieve relevant information from memory, integrate the information, and respond in the terms of the question. At each stage, errors may be introduced. Dividing the response process into components has provided a framework for exploring response effects, and it has led to new strategies for questioning. However, there has been little research demonstrating that respondents actually engage in the hypothesized sequence of cognitive operations when they answer questions, and the problems of retrieval that stimulated the application of cognitive psychology to survey methodology remain nearly as difficult as ever.

The Interview as Conversation

Respondents do not necessarily respond to the literal meaning of a question but rather to what they infer to be its intended meaning. A survey questionnaire serves as a script performed as part of an interaction between respondent and interviewer. The interaction affects how the script is enacted and interpreted. Thus, the construction of meaning is a social process, and it is not carried by question wording alone. Participants in a conversation assume it has a purpose, and they rely on implicit rules in a cooperative effort to understand and achieve it. They take common knowledge for granted and assume that each participant will make his or her contribution relevant and as informative as required, but no more informative than necessary. (These conversational maxims were developed by Paul Grice, a philosopher.) The resulting implications for the interview process are as follows:

1. Asking a question communicates that a respondent should be able to answer it.
2. Respondents interpret questions to make them relevant to the perceived intent.
3. Respondents interpret questions in ways that are relevant to their own situations.
4. Respondents answer the question they think an interviewer intended to ask.
5. Respondents do not report what they believe an interviewer already knows.
6. Respondents avoid providing redundant information.
7. If response categories are provided, at least one is true.

These implications help us understand a number of well-established questionnaire phenomena. Consistent with item 1, many people will answer survey questions about unfamiliar objects using the question wording and context to construct a plausible meaning. As implied by items 2 and 3, interpretations of questions vary greatly among respondents. Consistent with item 4, postinterview studies show that respondents do not believe the interviewer "really" wants to know everything that might be reported, even when a question asks for complete reports. Consistent with items 5 and 6, respondents reinterpret questions to avoid redundancy. As implied by item 7, respondents are unlikely to volunteer a response that is not offered in a closed question.

The conversational perspective has been the source of an important critique of standardization, which is seen as interfering with the conversational resources that participants would ordinarily employ to reach a common understanding, and it has led some researchers to advocate flexible rather than standardized questioning. A conversational perspective naturally leads to a consideration of the influences that one question may have on interpretations of subsequent ones and also the influence of the interview context—what respondents are told and what they infer about the purposes for asking the questions—on their interpretations and responses.

Constructing Questionnaires

Constructing a questionnaire involves many decisions about the wording and ordering of questions, selection and wording of response categories, formatting and mode of administration of the questionnaire, and introducing and explaining the survey. Although designing a questionnaire remains an art, there is increasing knowledge available to inform these decisions.

Question Wording

Although respondents often seem to pay scant attention to survey questions or instructions, they are often exquisitely sensitive to subtle changes in words and syntax. Question wording effects speak to the power and complexity of language processing, even when respondents are only half paying attention.

A famous experiment illustrates the powerful effect that changing just one word can have in rare cases. In a national sample, respondents were randomly assigned to

be asked one of two questions:

1. "Do you think the United States should allow public speeches against democracy?"
2. "Do you think the United States should forbid public speeches against democracy?"

Support for free speech is greater—by more than 20 percentage points—if respondents answer question 2 rather than question 1. That is, more people answer "no" to question 2 than answer "yes" to question 1; "not allowing" speeches is not the same as "forbidding" them, even though it might seem to be the same. The effect was first found by Rugg in 1941 and later replicated by Schuman and Presser in the United States and by Schwarz in Germany, so it replicates in two languages and has endured more than 50 years—even as support for freedom of speech has increased, according to both versions.

Terminology

"Avoid ambiguity" is a truism of questionnaire design. However, language is inherently ambiguous, and seemingly simple words may have multiple meanings. Research by Belson and others demonstrates that ordinary words and phrases, such as "you," "children," and "work," are interpreted very differently by different respondents.

Complexity and Ambiguity

Both cognitive and linguistic factors may impede respondents' ability to understand a question at all, as well as give rise to variable or erroneous interpretations. Questionnaire designers often intend a survey question to be interpreted literally. For example,

"During the past 12 months, since January 1, 1987, how many times have you seen or talked with a doctor or assistant about your health? Do not count any times you might have seen a doctor while you were a patient in a hospital, but count all other times you actually saw or talked to a medical doctor of any kind about your health."

Such questions challenge respondents, who must parse the question, interpret its key referents (i.e., "doctor or assistant" and "medical doctor of any kind"), infer the events to be included (i.e., visits to discuss respondent's health in person or by telephone during the past 12 months) and excluded (i.e., visits while in a hospital), and keep in mind all these elements while formulating an answer. Apart from a formidable task of recall, parsing such a complex question may overwhelm available mental resources so that a respondent does not understand the question fully or at all. Processing demands are increased by embedded clauses or sentences (e.g., "while you were a patient in a hospital") and by syntactic ambiguity. An example of syntactic ambiguity appears in an instruction on a U.S. census questionnaire to include "People living here most of the time while working, even if they have another place to live." The scope of the quantifier "most" is ambiguous and consistent with two possible interpretations: (i) "... [most of the time] [while working] ..." and (ii) "... [most of the [time while working]]...."

Ambiguity also can arise from contradictory grammatical and semantic elements. For example, it is unclear whether the following question asks respondents to report just one race: "I am going to read you a list of race categories. Please choose one or more categories that best indicate your race." "One or more" is contradicted by the singular reference to "race" and by "best indicate," which is interpretable as a request to select one.

Cognitive overload due to complexity or ambiguity may result in portions of a question being lost, leading to partial or variable interpretations and misinterpretations. Although the negative effects of excessive burden on working memory are generally acknowledged, the practical limits for survey questions have not been determined, nor is there much research on the linguistic determinants of survey question comprehension.

Presupposition

A presupposition is true regardless of whether the statement is true or false; that is, it is constant under negation. (For example, the sentences "I am proud of my career as a survey methodologist" and "I am not proud of my career as a survey methodologist" both presuppose I have a career as a survey methodologist.) A question generally shares the presuppositions of its assertions. "What are your usual hours of work?" presupposes that a respondent works, and that his or her hours of work are regular. Answering a question implies accepting its presuppositions, and a respondent may be led to provide an answer even if its presuppositions are false. Consider an experiment by Loftus in which subjects who viewed accident films were asked, "Did you see *a* broken headlight?" or "Did you see *the* broken headlight?" Use of the definite article triggers the presupposition that there was a broken headlight, and people asked the latter question were more likely to say "yes," irrespective of whether the film showed a broken headlight.

As described by Levinson, linguists have isolated a number of words and sentence constructions that trigger presuppositions, such as change of state verbs (e.g., "Have you *stopped* attending church?") and factive verbs (e.g., "regret," "realize," and "know"). (For example, "If you knew that the AMA is opposed to Measure H, would you change your opinion from *for* Measure H to *against* it?" presupposes the AMA is opposed to Measure H.) Forced choice questions such as "Are you a Republican or a Democrat?" presuppose that one of the alternatives is true.

Fortunately for questionnaire designers, presuppositions may be cancelled. "What are your usual hours of

work?" might be reworded to ask, "What are your usual hours of work, or do you not have usual hours?" Filter questions [e.g., "Do you work?" and (if yes) "Do you work regular hours?"] can be used to test and thereby avoid unwarranted presuppositions.

Question Context and Order

Question order changes the context in which a particular question is asked. Prior questions can influence answers to subsequent questions through several mechanisms. First, the semantic content of a question can influence interpretations of subsequent questions, especially when the subsequent questions are ambiguous. For example, an obscure "monetary control bill" was more likely to be supported when a question about it appeared after questions on inflation, which presumably led respondents to infer that the bill was an anti-inflation measure.

Second, the thoughts or feelings brought to mind while answering a question may influence answers to subsequent questions. This is especially likely when an answer to a question creates expectations for how a subsequent one should be answered. A famous experiment manipulated the order of a pair of questions:

"Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?"

"Do you think a Communist country like Russia should let American newspaper reporters come in and send back to America the news as they see it?"

Respondents were much more likely to think Communist reporters should be allowed in the United States if they answered that question second. Respondents apparently answered whichever question was asked first in terms of pro-American or anti-Communist sentiments. The second question activated a norm of reciprocity. Since many respondents felt constrained to treat reporters from both countries equally, they gave an answer to the second question that was consistent with the first.

Third, following conversational maxims, respondents may interpret questions so they are not redundant with prior questions. When a specific question precedes a general question, respondents "subtract" their answer to the specific question from their answer to the general one in certain circumstances. Respondents asked questions about marital satisfaction and general life satisfaction reinterpret the general question to exclude the specific one: "Aside from your marriage, which you already told us about, how satisfied are you with other aspects of your life?"

This type of context effect, called a part–whole effect by Schuman and Presser, can occur for factual as well as attitudinal questions. For example, race and Hispanic origin items on the U.S. census form are perceived as redundant by many respondents, although they are officially defined as different. When race (the more general item) appears first, many Hispanic respondents fail to find a race category with which they identify, so they check "other" and write in "Hispanic." When Hispanic origin is placed first so that such respondents first have a chance to report their Hispanic identity, they are less likely to report their Hispanic origin in the race item. Thus, when the specific item comes first, many respondents reinterpret race to exclude the category Hispanic. In this case, manipulating the context leads to reporting that is more consistent with measurement objectives.

One might wonder why a prior question about marital satisfaction would lead respondents to exclude, rather than include, their feelings about their marriages in their answers to a general life satisfaction question. Accounts of when information primed by a prior question will be subtracted rather than assimilated into later answers or interpretations have been offered by Schwarz and colleagues and by Tourangeau *et al.*

The argument is that when people are asked to form a judgment they must retrieve some cognitive representation of the target stimulus, and they must also determine a standard of comparison to evaluate it. Some of what they call to mind is influenced by preceding questions and answers, and this temporarily accessible information may lead to context effects. It may be added to (or subtracted from) the representation of the target stimulus. The questionnaire format and the content of prior questions may provide cues or instructions that favor inclusion or exclusion. For example, Schwarz and colleagues induced either an assimilation or a contrast effect in German respondents' evaluations of the Christian Democratic party by manipulating a prior knowledge question about a highly respected member (X) of the party. By asking "Do you happen to know which party X has been a member of for more than 20 years?" respondents were led to add their feelings about X to their evaluation of the party in a subsequent question, resulting in an assimilation effect. Asking "Do you happen to know which office X holds, setting him aside from party politics?" led them to exclude X from their evaluation of the party, resulting in a contrast effect.

Alternatively, the information brought to mind may influence the standard of comparison used to judge the target stimulus and result in more general context effects on a set of items, not just the target. For example, including Mother Teresa in a list of public figures whose moral qualities were to be evaluated would probably lower the ratings for everyone else on the list. Respondents anchor a scale to accommodate the range of stimuli presented to them, and an extreme (and relevant) example in effect

shifts the meaning of the scale. This argues for explicitly anchoring the scale to incorporate the full range of values in order to reduce such contextual influences.

Response Categories and Scales

The choice and design of response categories are among the most critical decisions about a questionnaire. As noted, a question that offers a choice among alternatives presupposes that one of them is true. This means that respondents are unlikely to volunteer a response option that is not offered, even if it might seem an obvious choice.

Open versus Closed Questions

An experiment by Schuman and Presser compared open and closed versions of the question, "What do you think is the most important problem facing this country at present?" The closed alternatives were developed using responses to the open-ended version from an earlier survey. Just as the survey went in the field, a prolonged cold spell raised public fears of energy shortages. The open version registered the event: "food and energy shortages" responses were given as the most important problem by one in five respondents. The closed question did not register the energy crisis because the category was not offered in the closed question, and only one respondent volunteered it.

This example illustrates an advantage of open questions: their ability to capture answers unanticipated by questionnaire designers. They can provide detailed responses in respondents' own words, which may be a rich source of data. They avoid tipping off respondents as to what response is normative, so they may obtain more complete reports of socially undesirable behaviors. On the other hand, responses to open questions are often too vague or general to meet question objectives. Closed questions are easier to code and analyze and compare across surveys.

Types of Closed-Response Formats

The previous example illustrates that response alternatives must be meaningful and capture the intended range of responses. When respondents are asked to select only one response, response alternatives must also be mutually exclusive.

The following are common response formats:

Agree–disagree: Many survey questions do not specify response alternatives but invite a "yes" or "no" response. Often, respondents are offered an assertion to which they are asked to respond; for example, "Do you agree or disagree?—Money is the most important thing in life." Possibly because they state only one side of an issue, such items encourage acquiescence, or a tendency to agree regardless of content, especially among less educated respondents.

Forced choice: In order to avoid the effects of acquiescence, some methodologists advocate explicitly mentioning the alternative responses. In a stronger form, this also involves providing substantive counterarguments for an opposing view:

"If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law?"

"If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law because it would be too difficult to enforce?"

Formal balance, as in the first question, does not appear to affect response distributions, but providing counterarguments does consistently move responses in the direction of the counterarguments, according to Schuman and Presser's experiments. Devising response options with counterarguments may not be feasible if there are many plausible reasons for opposition, since the counterargument can usually only capture one.

Ordered response categories or scales: Respondents may be asked to report in terms of absolute frequencies (e.g., "Up to $\frac{1}{2}$ hour, $\frac{1}{2}$ to 1 hour, 1 to $1\frac{1}{2}$ hours, $1\frac{1}{2}$ to 2 hours, 2 to $2\frac{1}{2}$ hours, more than $2\frac{1}{2}$ hours"), relative frequencies (e.g., "All of the time, most of the time, a good bit of the time, some of the time, a little bit of the time, none of the time"), evaluative ratings (e.g., "Excellent, pretty good, only fair, or poor"), and numerical scales (e.g., "1 to 10" and "−5 to +5").

Response scales provide a frame of reference that may be used by respondents to infer a normative response. For example, Schwarz and colleagues compared the absolute frequencies scale presented in the previous paragraph with another that ranged from "Up to $2\frac{1}{2}$ hours" to "More than $4\frac{1}{2}$ hours" in a question asking how many hours a day the respondent watched television. The higher scale led to much higher frequency reports, presumably because many respondents were influenced by what they perceived to be the normative or average (middle) response in the scale. If there is a strong normative expectation, an open-ended question may avoid this source of bias. Frequently, ordered categories are intended to measure where a respondent belongs on an underlying dimension (scale points may be further assumed to be equidistant). Careful grouping and labeling of categories is required to ensure they discriminate. Statistical tools are available to evaluate how well response categories perform. For example, an analysis by Reeve and Mâsse (see Presser *et al.*) applied item response theory to show that "a good bit of the time" in the relative frequencies scale presented previously was not discriminating or informative in a mental health scale.

Rating scales are more reliable when all points are labeled and when a branching structure is used, with an initial question (e.g., "Do you agree or disagree . . . ?")

followed up by a question inviting finer distinctions ("Do you strongly agree/disagree, or somewhat agree/disagree?"), according to research by Krosnick and colleagues and others. The recommended number of categories in a scale is 7, plus or minus 2. Numbers assigned to scale points may influence responses, apart from the verbal labels. Response order may influence responses, although the basis for primacy effects (i.e., selecting the first category) or recency effects (i.e., selecting the last category) is not fully understood. Primacy effects are more likely with response options presented visually (in a self-administered questionnaire or by use of a show card) and recency effects with aural presentation (as in telephone surveys).

Offering an Explicit "Don't Know" Response Option

Should "don't know" be offered as an explicit response option? On the one hand, this has been advocated as a way of filtering out respondents who do not have an opinion and whose responses might therefore be meaningless. On the other hand, it increases the number of respondents who say "don't know," resulting in loss of data. Schuman and Presser find that the relative proportions choosing the substantive categories are unaffected by the presence of a "don't know" category, and research by Krosnick and others suggests that offering "don't know" does not improve data quality or reliability. Apparently, many respondents who take the easy out by saying "don't know" when given the opportunity are capable of providing meaningful and valid responses to opinion questions. Thus, "don't know" responses are best discouraged.

Communicating Response Categories and the Response Task

Visual aids, such as show cards, are useful for communicating response categories to respondents in personal interviews. In self-administered questionnaires, the categories are printed on the questionnaire. In either mode, the respondent does not have to remember the categories while formulating a response but can refer to a printed list. Telephone interviews, on the other hand, place more serious constraints on the number of response categories; an overload on working memory probably contributes to the recency effects that can result from auditory presentation of response options. Redesigning questions to branch, so each part involves a smaller number of options, reduces the difficulty. Different formats for presenting response alternatives in different modes may cause mode biases; on the other hand, the identical question may result in different response biases (e.g., recency or primacy effects) in different modes. Research is needed on this issue, especially as it affects mixed mode surveys.

The same general point applies to communicating the response task. For example, in developmental work conducted for implementation of a new census race question that allowed reports of more than one race, it proved difficult to get respondents to notice the "one or more" option. One design solution was to introduce redundancy so respondents had more than one chance to absorb it.

Addressing Problems of Recall and Retrieval

Psychological theory and evidence support several core principles about memory that are relevant to survey questionnaire construction:

1. Autobiographical memory is reconstructive and associative.
2. Autobiographical memory is organized hierarchically. (Studies of free recall suggest the organization is chronological, with memories for specific events embedded in higher order event sequences or periods of life.)
3. Events that were never encoded (i.e., noticed, comprehended, and stored in memory) cannot be recalled.
4. Cues that reinstate the context in which an event was encoded aid memory retrieval.
5. Retrieval is effortful and takes time.
6. Forgetting increases with the passage of time due to decay of memory traces and to interference from new, similar events.
7. The characteristics of events influence their memorability: Salient, consequential events are more likely to be recalled than inconsequential or trivial ones.
8. Over time, memories become less idiosyncratic and detailed and more schematic and less distinguishable from memories for other similar events.
9. The date an event occurred is usually one of its least accurately recalled features.

Principle 6 is consistent with evidence of an increase in failure to report events, such as hospitalizations or consumer purchases, as the time between the event and the interview—the retention interval—increases. Hospitalizations of short duration are more likely to be forgotten than those of long duration, illustrating principle 7. A second cause of error is telescoping. A respondent who recalls that an event occurred may not recall when. On balance, events tend to be recalled as happening more recently than they actually did—that is, there is forward telescoping, or events are brought forward in time. Forward telescoping is more common for serious or consequential events (e.g., major purchases and crimes that were reported to police). Backward telescoping, or recalling events as having happened longer ago than they did, also occurs. The aggregate effect of telescoping and forgetting is a pronounced recency bias, or piling up of reported events in the most recent portion of a reference

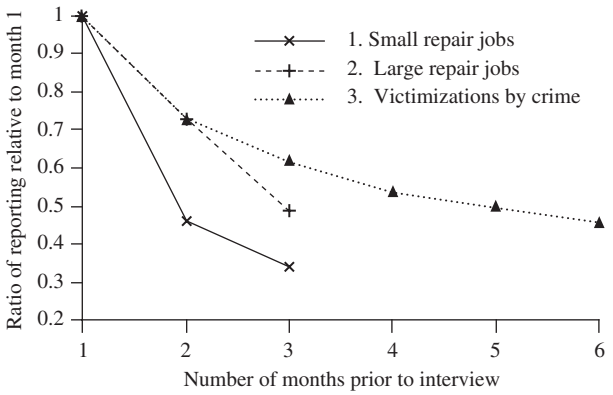


Figure 1 Recency bias for two surveys. *Sources:* Neter, J., and Waxberg, J. (1964). A study of response errors in expenditures data from household interviews. *J. Am. Stat. Assoc.* **59**, 18–55; and Biderman, A. D., and Lynch, J. P. (1981). Recency bias in data on self-reported victimization. *Proc. Social Stat. Section (Am. Stat. Assoc.)*, 31–40.

period. Figure 1 illustrates the effect for two surveys. The rate for the month prior to the interview is taken as a base, and the rates for other months are calculated relative to it. Line 3 shows that monthly victimization rates decline monotonically each month of a 6-month reference period. Lines 1 and 2 show the same for household repairs over a 3-month reference period; note the steeper decline for minor repairs. Recent theories explain telescoping in terms of an increase in uncertainty about the timing of older events. Uncertainty only partially explains telescoping, however, since it predicts more telescoping of minor events than of major ones, but in fact the opposite occurs.

Because of the serious distortions introduced by failure to recall and by telescoping, survey methodologists are generally wary of "Have you ever . . . ?"-type questions that ask respondents to recall experiences over a lifetime. Instead, they have developed various questioning strategies to try to improve respondents' recall.

Strategies to Improve Temporal Accuracy

In order to improve recall accuracy, questions are usually framed to ask respondents to recall events that occurred during a reference period of definite duration. Another procedure is to bound an interview with a prior interview in order to prevent respondents from telescoping in events that happened before the reference period. Results of the bounding interview are not included in survey estimates. Another method attempts to make the boundary of the reference period more vivid by associating it with personal or historical landmark events. This can reduce telescoping, especially if the landmark is relevant to the types of events a respondent is asked to recall. A more elaborate procedure, the event history calendar, attempts to structure flexible questions in a way that reflects the organization of memory, and it has proved promising in research by Belli and associates.

For many survey questions, respondents may rely on a combination of memory and judgment to come up with answers. When the number of events exceeds 10, very few respondents actually attempt to recall and enumerate each one. Instead, they employ other strategies, such as recalling a few events and extrapolating a rate over the reference period, retrieving information about a benchmark or standard rate and adjusting upward or downward, or guessing. By shortening the reference period, giving respondents more time, or decomposing a question into more specific questions, questionnaire designers can encourage respondents to enumerate episodes if that is the goal.

Aided and Unaided Recall

In general, unaided (or free) recall produces less complete reporting than aided recall. It may also produce fewer erroneous reports. Cues and reminders serve to define the scope of eligible events and stimulate recall of relevant instances. A cuing approach was employed to improve victimization reporting in a 1980s redesign of the U.S. crime victimization survey. Redesigned screening questions were structured around multiple frames of reference (acts, locales, activities, weapons, and things stolen) and included numerous cues to stimulate recall, including recall for underreported, sensitive, and nonstereotypical crimes. The result was much higher rates of reporting.

Although cuing improves recall, it can also introduce error because it leads to an increase in reporting of ineligible incidents as well as eligible ones. In addition, the specific cues can influence the kinds of events that are reported. The crime survey redesign again is illustrative. Several crime screener formats were tested experimentally. The cues in different screeners emphasized different domains of experience, with one including more reminders of street crimes and another placing more emphasis on activities around the home. Although the screeners produced the same overall rates of victimization, there were large differences in the characteristics of crime incidents reported. More street crimes and many more incidents involving strangers as offenders were elicited by the first screener.

Dramatic cuing effects such as this may result from the effects of two kinds of retrieval interference. Part-set cuing occurs when specific cues interfere with recall of noncued items in the same category. For example, giving "knife" as a weapons cue would make respondents less likely to think of "poison" or "bomb" and (by inference) less likely to recall incidents in which these noncued items were used as weapons. The effect would be doubly biasing if (as is true in experimental studies of learning) retrieval

in surveys is enhanced for cued items and depressed for noncued items.

A second type of interference is a retrieval block that occurs when cues remind respondents of details of events already mentioned rather than triggering recall of new events. Recalling one incident may block retrieval of others because a respondent in effect keeps recalling the same incident. Retrieval blocks imply underreporting of multiple incidents. Early cues influence which event is recalled first, and once an event is recalled, it inhibits recall for additional events. Therefore, screen questions or cues asked first may unduly influence the character of events reported in a survey.

Another illustration of cuing or example effects comes from the ancestry question in the U.S. census. "English" appeared first in the list of examples following the ancestry question in 1980 but was dropped in 1990. There was a corresponding decrease from 1980 to 1990 of approximately 17 million persons reporting English ancestry. There were also large increases in the numbers reporting German, Acadian/Cajun, or French-Canadian ancestry, apparently due to the listing of these ancestries as examples in 1990 but not 1980, or their greater prominence in the 1990 list. These effects of examples, and their order, may occur because respondents write in the first ancestry listed that applies to them. In a related question, examples did not have the same effect. Providing examples in the Hispanic origin item increased reporting of specific Hispanic origin groups, both of example groups and of groups not listed as examples, apparently because examples helped communicate the intent of the question.

Tools for Pretesting and Evaluating Questions

It has always been considered good survey practice to pretest survey questions to ensure they can be administered by interviewers and understood and answered by respondents. Historically, such pretests involved interviewers completing a small number of interviews and being debriefed. Problems were identified based on interview results, such as a large number of "don't know" responses, or on interviewers' reports of their own or respondents' difficulties with the questions. This type of pretest is still valuable and likely to reveal unanticipated problems. (For automated instruments, it is also essential to test the instrument programming.) However, survey researchers have come to appreciate that many questionnaire problems are likely to go undetected in a conventional pretest, and in recent decades the number and sophistication of pretesting methods have expanded. The new methods have led to greater awareness that survey questions are neither asked nor understood in a uniform way, and revisions based on pretest results appear to lead to improvements. However, questions remain about the validity and reliability of the methods and also the relationship between the problems they identify and measurement errors in surveys. Because the methods appear better able to identify problems than solutions, an iterative approach involving pretesting, revision, and further pretesting is advisable. (A largely unmet need concerns pretesting of translated questionnaires. For cross-national surveys, and increasingly for intranational ones, it is critical to establish that a questionnaire works and produces comparable responses in multiple languages.)

Expert Appraisal and Review

Review of a questionnaire by experts in questionnaire design, cognitive psychology, and/or the relevant subject matter is relatively cost-effective and productive in terms of problems identified. Nonexpert coders may also conduct a systematic review using the questionnaire appraisal scheme devised by Lessler and Forsyth (see Schwarz and Sudman) to identify and code cognitive problems of comprehension, retrieval, judgment, and response generation. Automated approaches advanced by Graesser and colleagues apply computational linguistics and artificial intelligence to build computer programs that identify interpretive problems with survey questions (see Schwarz and Sudman).

Think-Aloud or Cognitive Interviews

This method was introduced to survey researchers from cognitive psychology, where it was used by Herbert Simon and colleagues to study the cognitive processes involved in problem solving. The procedure as applied in surveys is to ask laboratory subjects to verbalize their thoughts—to think out loud—as they answer survey questions (or, if the task involves filling out a self-administered questionnaire, to think aloud as they work their way through the questionnaire). Targeted probes may also be administered (e.g., "What period of time are you thinking of here?"). Tapes, transcripts, or summaries of respondents' verbal reports are reviewed to reveal both general strategies for answering survey questions and difficulties with particular questions. Cognitive interviews may be concurrent or retrospective, depending on whether respondents are asked to report their thoughts and respond to probes while they answer a question, or after an interview is concluded. Practitioners vary considerably in how they conduct, summarize, and analyze cognitive interviews, and the effects of such procedural differences are being explored. The verbal reports elicited in cognitive interviews are veridical if they represent information available in working memory at the time a report is verbalized, if the respondent is not asked to explain and interpret his or her own thought processes, and if the

social interaction between cognitive interviewer and subject does not alter a respondent's thought process, according to Willis (see Presser *et al.*). Cognitive interviewing has proved to be a highly useful tool for identifying problems with questions, although research is needed to assess the extent to which problems it identifies translate into difficulties in the field and errors in data.

Behavior Coding

This method was originally introduced by Cannell and colleagues to evaluate interviewer performance, but it has come to be used more frequently to pretest questionnaires. Interviews are monitored (and usually tape recorded), and interviewer behaviors (e.g., "Reads question exactly as worded" and "Reads with major change in question wording, or did not complete question reading") and respondent behaviors (e.g., "Requests clarification" and "Provides inadequate answer") are coded and tabulated for each question. Questions with a rate of problem behaviors above a threshold are regarded as needing revision. Behavior coding is more systematic and reveals many problems missed in conventional pretests. The method does not necessarily reveal the source of a problem, which often requires additional information to diagnose. Nor does it reveal problems that are not manifested in behavior. If respondents and interviewers are both unaware that respondents misinterpret a question, it is unlikely to be identified by behavior coding. Importantly, behavior coding is the only method that permits systematic evaluation of the assumption that interviewers administer questions exactly as worded.

Respondent Debriefing or Special Probes

Respondents may be asked directly how they answered or interpreted specific questions or reacted to other aspects of the interview. Survey participants in effect are asked to assume the role of informant, rather than respondent. Probes to test interpretations of terminology or question intent are the most common form of debriefing question, and their usefulness for detecting misunderstandings is well documented by Belson, Cannell, and others. For example, the following probes were asked following the previously discussed question about doctor visits: "We're interested in who people include as doctors or assistants. When you think of a doctor or assistant, would you include a dentist or not? Would you include a laboratory or X-ray technician or not?... Did you see any of those kinds of people during the last year?" Specific probes targeted to suspected misunderstandings have proved more fruitful than general probes or questions about respondents' confidence in their answers. (Respondents tend to be overconfident, and there is no consistent evidence of a correlation between confidence and accuracy.) Debriefing questions or special probes have also proved useful for assessing question sensitivity ("Were there any questions in this interview that you felt uncomfortable answering?"), other subjective reactions ("Did you feel bored or impatient?"), question comprehension ("Could you tell me in your own words what that question means to you?"), and unreported or misreported information ("Was there an incident you thought of that you didn't mention during the interview? I don't need details."). Their particular strength is that they reveal misunderstandings and misinterpretations of which both respondents and interviewers are unaware.

Vignettes

Vignettes are brief scenarios that describe hypothetical characters or situations. Because they portray hypothetical situations, they offer a less threatening way to explore sensitive subjects. Instead of asking respondents to report directly how they understand a word or complex concept ("What does the term *crime* mean to you?"), which has not proved to be generally productive, vignettes pose situations that respondents are asked to judge. For instance,

"I'll describe several incidents that could have happened. We would like to know for each, whether you think it is the kind of crime we are interested in, in this survey. ... Jean and her husband got into an argument. He slapped her hard across the face and chipped her tooth. Do you think we would want Jean to mention this incident to us when we asked her about crimes that happened to her?"

The results reveal how respondents interpret the scope of survey concepts (such as crime) as well as the factors influencing their judgments. Research suggests that vignettes provide robust measures of context and question wording effects on respondents' interpretations.

Split-Sample Experiments

Ultimately, the only way to evaluate the effects of variations in question wording, context, etc. on responses is to conduct an experiment in which samples are randomly assigned to receive the different versions. It is essential to ensure that all versions are administered under comparable conditions, and that data are coded and processed in the same way, so that differences between treatments can be unambiguously attributed to the effects of questionnaire variations. Comparison of univariate response distributions shows gross effects, whereas analysis of subgroups reveals conditional or interaction effects. Field experiments can be designed factorially to evaluate the effects of a large number of questionnaire variables

on responses, either for research purposes or to select those that produce the best measurements. When a survey is part of a time series and data must be comparable from one survey to the next, this technique can be used to calibrate a new questionnaire to the old.

Conclusion

Survey questionnaire designers aim to develop standardized questions and response options that are understood as intended by respondents and that produce comparable and meaningful responses. In the past, the extent to which these goals were met in practice was rarely assessed. In recent decades, better tools for providing feedback on how well survey questions perform have been introduced or refined, including expert appraisal, cognitive interviewing, behavior coding, respondent debriefing, vignettes, and split-sample experiments. Another advance is new theoretical perspectives that help make sense of the effects of question wording and context. One perspective examines the cognitive tasks in which a respondent must engage to answer a survey question. Another examines the pragmatics of communication in a survey interview. Both have shed light on the response process, although difficult problems remain unsolved. In addition, both perspectives suggest limits on the ability to fully achieve standardization in surveys. New theory and pretesting tools provide a scientific basis for decisions about construction of survey questionnaires.

See Also the Following Articles

Interviews • Mail Surveys • Survey Design • Surveys • Total Survey Error • Web-Based Survey

Further Reading

Belson, W. A. (1981). *The Design and Understanding of Survey Questions.* Gower, London.

Biderman, A. D., Cantor, D., Lynch, J. P., and Martin, E. (1986). *Final Report of the National Crime Survey Redesign Program.* Bureau of Social Science Research, Washington, DC.

Fowler, F. J. (1995). *Improving Survey Questions: Design and Evaluation.* Sage, Thousand Oaks, CA.

Levinson, S. C. (1983). *Pragmatics.* Cambridge University Press, Cambridge, UK.

Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., and Singer, E. (2004). *Methods for Testing and Evaluating Survey Questionnaires.* Wiley, New York.

Schaeffer, N. C., and Presser, S. (2003). The science of asking questions. *Annu. Rev. Sociol.* **29,** 65–88.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* Academic Press, New York.

Schwarz, N., and Sudman, S. (eds.) (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research.* Jossey-Bass, San Francisco.

Sudman, S., Bradburn, N. M., and Schwarz, N. (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology.* Jossey-Bass, San Francisco.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press, Cambridge, UK.

Surveys

Kenneth A. Rasinski
NORC, University of Chicago, Chicago, Illinois, USA

Glossary

anchor points Values that serve as bases for judgments. They are usually a consideration in response formats.
bounded recall A technique used in panel surveys whereby respondents are reminded of their responses from the previous wave. It is used to combat seam effects.
clustered sampling A sampling strategy that requires the selection of a sample of higher order groups within which desired elements are organized in order to survey the desired elements; for example, selecting a sample of schools in order to survey the children attending those schools.
cognitive pretesting A method of pretesting survey questions that considers respondents' understanding of the question, their ability to recall information, the way they arrive at judgments, and the accuracy with which they report information.
estimation strategies Systematic methods that survey respondents use to make reasonable guesses about behavioral frequency information of which they are unsure.
interrogatives, space of uncertainty, and propositions Linguistic terms describing the question asking and answering process. An interrogative is a question statement. The space of uncertainty is the set of possible answers. Propositions are suggestions made by the question wording that may lead respondents to a conclusion, either intended or unintended.
narrative chains The idea that events are linked in memory as narratives, or stories, rather than by topic.
non-sampling error The difference between the population estimate and the true population value due to factors not related to the sample design, such as undercoverage, nonresponse, and reporting inaccuracies.
panel survey A survey in which a sample of respondents is interviewed more than once.
population estimate A statistic derived from a sample that is used as a proxy for the population value.
repeated cross section A series of cross-sectional surveys—that is, surveys in which individual samples are selected at different points in time—in which the same questions are repeated.
sample frame List of elements (units to be studied in a survey) in a population from which a sample will be drawn. In many surveys, elements are individuals. The sample frame sometimes has to be constructed from multiple sources and may contain duplicate or missing elements.
sampling error The difference between the population estimate, derived from the sample, and the true population value. The difference is due to the fact that not all elements in the population are selected into the sample.
seam effect Bias in reporting that occurs in panel surveys in which the respondent reports about several months during each wave of the panel. More change in status is noted across the seam of any two waves—that is, from the most recent reporting month of the last wave to the most remote reporting month of the subsequent wave—than between any months within waves.
stratified sampling Organizing the sample frame by mutually exclusive and exhaustive categories (e.g., region, sex, race, and age) and selecting samples within each of the categories, sometimes using different sampling rates for each category.
telescoping The tendency of respondents to recall events as happening earlier than they actually occurred (forward telescoping) or, less frequently, later than they actually occurred (backward telescoping).
time line An aid used in surveys to help respondents recall events that happened in the past.

Surveys are powerful research tools used by social scientists to study social phenomena. The process of conducting a survey includes defining a population, selecting a sample, developing and administering a questionnaire, and collecting and analyzing data. Survey sampling permits the researcher to make inferences from

samples to populations. Questionnaire design has been advanced by studies of the psychology of survey responding, primarily examining the role of cognition in the survey response process. Advances in questionnaire design have also emerged from attention to the conversational nature of a survey interview and linguistic aspects of survey questions. Survey researchers have always taken advantage of new methodologies and they continue to do so. They exploit new technologies as opportunities for collecting better data, and they conduct methodological research to overcome obstacles posed by the new technologies.

Introduction

Survey research is a popular and powerful means by which to study people and organizations in society. It consists of a rich set of techniques used to obtain information about individual attitudes, values, behaviors, opinions, knowledge, and circumstances. Surveys are also used to study organizations and institutions, for example, assessing their culture, policies, and finances. This article discusses both of these uses of surveys but emphasizes surveys conducted on individuals. The goal of this article is to inform the consumer of survey information about survey techniques and their impact on the interpretation of results. Most of the discussion about individual-level surveys applies to interpretation of surveys about organizations.

A social survey is a standardized and systematic method for obtaining information about a population by using a questionnaire to measure elements sampled from that population. Standardization and systemization facilitate replication, which is a hallmark of the scientific method. Thus, surveys are an important tool in advancing social science. The sampling component—that is, the ability to study a sample and make projections to a population—makes the sample survey an efficient tool for studying the characteristics of populations. The questionnaire is the main method for extracting information, and careful questionnaire construction is important in order to obtain high-quality data. Social, cognitive, and linguistic elements are at play in questionnaire construction. The means for administering the questionnaire encompasses both the training of interviewers, who are the information collection agents, and the selection of a modality of the interview. Modalities include face-to-face administration, administration by telephone, mail and computerized self-administered questionnaires, and, recently, surveys via the Internet and the World Wide Web.

Most surveys involve the following steps, although not necessarily in the order presented. First, a population one wishes to study is determined and a list of the population elements, called a sample frame, is obtained. Second, a topic or topics of interest are determined. The first and second steps combined may be thought of as defining a research question (or set of questions). For example, studying the employment patterns of high school dropouts necessitates both selecting topics relevant to employment and obtaining a sample of high school dropouts. The third step in the survey process is designing a method for sampling elements from the population. The fourth step is designing a questionnaire that reflects the topical areas of interest. The fifth step is deciding on a modality of administration. The fourth and fifth steps are important to consider together because the mode of administration will affect the design of the questionnaire. The sixth step consists of training interviewers in the administration of the questionnaire (which is sometimes, but not always, necessary for self-administered surveys), and the seventh step involves devising a method for compiling and aggregating the survey information. The analysis of the information is the eighth step. This article focuses on the first six steps.

Before the steps are discussed in detail, it is worthwhile to examine the types of surveys. There are three general types of surveys. A cross-sectional survey represents a population at a given point in time. A well-known example of a national cross-sectional survey is the General Social Survey (GSS) conducted by the National Opinion Research Center at the University of Chicago. In the GSS, which has been conducted nearly every year on independent samples of Americans for the past 30 years, some questions appear in only 1 year, whereas others are repeated year after year. Repeating questions in cross-sectional surveys is useful for assessing trends in opinions, attitudes, values, knowledge, or behavior. When cross-sectional samples include the same items year after year, the entire set of surveys is called a repeated cross-section design. Trend data for repeated cross-section designs over an extended period represent a combination of change (or stability) in responses to the questions and in the population demographics.

A second type of survey is the longitudinal survey. In this type of survey, the same respondents are interviewed repeatedly over time. Some longitudinal surveys are of an age cohort; the age group is followed over time, sometimes for as long as 20 years. For example, the National Longitudinal Survey of Youth consists of a sample of the senior class of 1978. This group has been reinterviewed nearly every year since that time. Another type of longitudinal survey is the panel survey. An example of this is the Panel Study of Income Dynamics (PSID) conducted by the Institute for Survey Research at the University of Michigan. In this type of survey, a sample is selected and interviewed. Sample members are then reinterviewed, sometimes at fixed intervals and other times to track the effects of some important social event. The sample is not an age cohort but a cross section of the population at the time the panel was created. Some panel surveys consist of only two rounds of interviews. Others, like the

PSID, have many rounds. More complicated designs include combinations of cross sections and panels.

Populations and Sampling

As mentioned previously, defining a population is one of the early steps in the survey process. A population is a total set of elements. It may be small, such as the total number of cars in a given parking lot, or large, such as the total number of households in the United States. Typically, surveys are concerned with collecting information from large populations. One way of collecting information about populations is to survey all the elements. This is called a census. A very famous example of a census done on a large population is that conducted in the United States every 10 years. More typically, surveys use scientific sampling methods and collect information about a subset of the total elements. If this is done properly, estimates about the characteristics of the entire population can be derived. The goal of a sample survey is to obtain unbiased estimates of population information without having to collect information from all the elements of the population. At this point, it is worthwhile to examine properties of information obtained from sample surveys compared to the information as it exists in the population. Because sample surveys collect information from only a portion of a population, statistics derived from the information obtained from a sample survey will not likely match exactly the statistics in the populations. For example, a scientifically designed sample survey of the household income of residents in a city may indicate that the median income is $35,500. If household incomes from all the households in that city were obtained, the true value might be $36,200. The survey value and the population value do not match.

However, if the sample is constructed according to scientific principles it is possible to determine a range within which the true population household income probably lies. This range will be centered around the population statistic. The range is called the margin of error; it is expressed as the estimate plus or minus a value. That value is a function of two factors, the standard deviation of the estimate and the sample size. Figure 1 shows sample sizes needed for margins of error of different magnitudes for a variable with standard deviation 0.5 and for another one with standard deviation 0.26. These numbers are the standard deviations of sample proportions of 0.5 and 0.93, respectively. To make these numbers more concrete, they can be thought of in terms of the proportion of a sample engaging in some behavior, for example, saying they will vote for candidate Jones in the upcoming mayoral election. If 50% of the sample say they will vote for Jones, the sample proportion voting for Jones is 0.5 and the standard variance of that proportion is $(0.5^* (1 - 0.5))^{0.5}$, or 0.5. On the other hand, if 93% of the sample say they will vote for Jones, the sample proportion is 0.93 and the standard deviation is $(0.93^* (1 - 0.93))^{0.5}$, or approximately 0.26.

A standard benchmark is to be 95% certain that the true value lies within the range determined by the margin of error, and the sample sizes in Fig. 1 were constructed using this benchmark. According to Fig. 1, if 50% of the sample say they will vote for Jones, and the survey was conducted using responses from 246 scientifically selected voters in the city for which Jones is running for mayor, then one can be 95% certain that the proportion of the population saying they will vote for Jones is

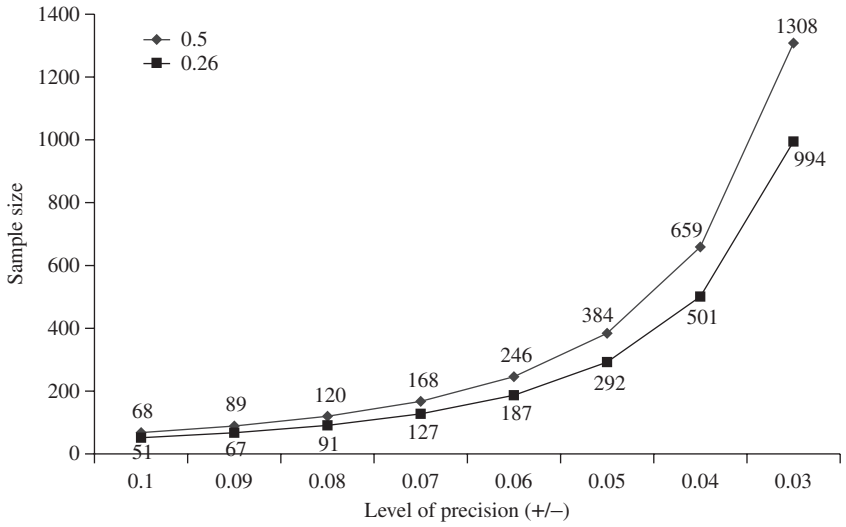


Figure 1 Approximate sample sizes needed for different levels of precision in a simple random sample survey assuming variances of 0.5 and 0.26.

between 44 and 56%. Figure 1 illustrates three points. First, for the same standard deviation, as the sample size increases the margin of error becomes smaller. Second, the relationship is not linear. To obtain increasingly smaller margins of error, the sample size begins to double or triple. When the standard deviation is 0.5, to go from a margin of error of ± 0.04 to one of ± 0.03 the sample size has to be doubled. To go from 0.03 to 0.02, the sample size has to almost be tripled. Although it is not shown in the figure, the survey would need to have a sample size of more than 16,000 cases in order to go from a margin of error of 0.02 to one of 0.01. The third point illustrated in the figure is that for variables with smaller standard deviations the sample size needed to obtain a given margin of error is less and the increase in sample size to obtain smaller standard errors is also less. Of course, it makes no sense to talk about a ± 0.1 margin of error for a sample proportion of 0.93 because $0.93 + 0.1$ gives a proportion greater than 1, which is clearly impossible. When proportions are very close to 0 or very close to 1, special statistical considerations have to be made for estimating some of the standard errors. The example here was used for illustration of the three points made previously, which hold despite this anomaly. Although this example used a sample proportion as the statistic, the same principles apply to other statistics, such as averages or totals. Of course, the standard deviations are calculated differently than for the proportion.

An additional point about sample sizes is that when dealing with populations of 10,000 or more, the size of the population does not affect the margin of error appreciably. Therefore, a sample size of 1000 will give roughly the same margin of error for a population of 10,000 as it will for a population of 100,000 or 184 million, the current estimate of the number of people in the United States. If a population is small such that it is practical to draw a sample of 10% or more, then a factor called the finite population correction can be used to reduce the standard deviation of an estimate and thus reduce the sample size needed for a given margin of error. The factor is calculated as $((N-n)/N))^{0.5}$, where N is the size of the population, and n is the size of the sample.

Sample Designs

Reference has been made to scientifically designed samples. Some basic elements of such samples are now discussed. A scientifically designed sample, at a minimum, specifies a method for choosing elements from a population with known selection probability. A simple random sample is the most elementary form of a scientifically designed sample. More complex designs (discussed later) use principles of the simple random sample as their basis. An example will illustrate a number of features of a simple random sample. Imagine that there are 5423 businesses with annual revenue of less than $50,000 in a certain geographic region and that the Internal Revenue Service wants to conduct an audit using a sample survey. Also imagine that you own one of these businesses. Using a simple random sample design, each business would be assigned an equal probability of selection, which would be 1/5423. In a sample of size 1, your business would have a 1/5423 or 0.000184% chance of being selected. If 10 businesses were selected simultaneously at random, then your business would have a 10/5423 chance of being in the sample. The larger the sample, the greater the probability that your business would be selected into it, but in each case the selection probability is known. For any population, the probability of any element within that population to be selected into a sample is the sample size divided by the population size.

Simple random sampling may work in certain, limited situations, but surveys usually call for more complicated designs. One level of complexity is stratification. In a stratified design, elements in the sample frame are grouped according to one or more characteristics and a simple random sample is selected from each stratum. Sometimes, elements are selected at different rates from each stratum. It is important to note that statistics derived from the sample must take into account the sampling rate in each stratum or erroneous population estimates will be calculated.

Another level of complexity is called clustering. Sometimes, it is impossible or impractical to obtain a good sample frame of the elements one wishes to study. If these elements are grouped into higher order elements, as is, for example, the population of eighth graders in public schools in a given county, then one can start with a list of all the public schools with eighth grades within the county and work down. This is done by selecting a sample of public schools from the list of schools and either surveying all the eighth graders in the sampled schools or selecting a sample of the eighth graders in each of the sampled schools. The higher order elements (schools) are called primary sampling units, and the elements contained within the primary units (students) are called secondary sampling units. The full-sample, or complex, probability of selection of a secondary element into the sample is the product of two selection probabilities. The first is the probability of the selection of the primary sampling unit, and the second is the probability of selection of the secondary sampling unit.

To make this concrete, suppose a researcher was interested in studying the characteristics of eighth graders in public schools within your county. Also suppose that there were 150 such schools and the researcher randomly selected a sample of 15. The selection probability for any given school would be 15/150 or 0.1. Now suppose that in 1 sampled school there were 100 eight graders and the

researcher sampled 30 of them using simple random sampling. The within-class selection probability for each student in the sample for that class would be 30/100 or 0.33. The full survey or complex selection probability for each sampled student from that class would be the selection probability for the school (0.1) multiplied by the selection probability of the student within the class in the school (0.33) or 0.033.

In a survey design, clustering can be extended to three levels (e.g., counties sampled within a region, schools sampled within counties, and students sampled within schools) or more. In addition, stratification and clustering can be combined to create complex survey designs. For example, the country could be divided into mutually exclusive quadrants of approximately the same geographic size. An equal number of schools could be selected within each quadrant, ensuring that the sample is not dominated by areas of the country that are heavily populated. This is useful especially if it is desirable to make comparisons between schools in densely populated areas and schools in sparsely populated areas. The calculation of the margin of error for stratified and clustered designs requires the use of specialized techniques. Making the assumption that these complex designs are like simple random samples, and analyzing them accordingly, is likely to result in the overestimation of the margin of error (making it too large) with a stratified design and the underestimation (making it too small) in a clustered design. The ratio of the sample design-based standard error to the standard error based on the assumption that the design was a simple random sample is called the sample's design effect, and it is a useful statistic to compare the efficiency of different sample designs.

Two other types of survey designs that are worth mentioning are convenience sampling and quota sampling. These are not scientific designs because there is no way of knowing the selection probability of each case in the sample. Although they can be useful under certain limited conditions, margins of error cannot be calculated. Convenience sampling involves collecting information from respondents who are available. If the population is small (e.g., the number of people who work in a small office building) and the convenience sample is large (e.g., 90% of the workers complete the survey), it is not unreasonable to assume that the results of the convenience sample represent results for the population.

However, convenience samples of large populations may be grossly misleading. Consider, for example, surveys that popular magazines regularly conduct. Their readership may be in the hundreds of thousands and 10,000 readers may mail in a completed questionnaire on a particular topic. However, it is likely that only those respondents who were interested in the topic chose to participate in it, causing a large amount of selection bias in the sample. Also, there is no way to determine whether a single respondent completed two or more surveys. Although information from these respondents may provide entertainment value, it should not be considered representative of any population because the sample of respondents was not selected in a scientific manner.

A second type of nonprobability sampling is quota sampling. If one knows something about the distribution of a population on key demographic characteristics, one can construct a convenience sample that reflects that distribution. Consider the following example. A certain town is known to have a population that consists of 30% African Americans, 20% Hispanics, and 50% whites. Furthermore, it is known that males and females are distributed evenly across the race/ethnicity groups. A quota sampling strategy would be to construct a convenience sample of town residents that consisted of 15% African American males, 15% African American females, 10% Hispanic males, 10% Hispanic females, 25% white males, and 25% white females. Quota sampling was frequently used by survey researchers into the early 1970s, and it is used occasionally today. There are cost advantages to this method, but, strictly speaking, margins of error cannot be determined because selection probabilities are not known.

Error and Bias in Surveys

Probability samples are the mainstay of modern survey research because they allow the researcher to control the margin of error. Probability samples, perfectly constructed and implemented, all other things being equal, produce unbiased estimates. However, rarely is it the case that construction and implementation are perfect, and when they are not the potential for bias arises. Bias is the difference between the population value that is supposed to be estimated and the population value that is actually estimated. There are two common sources of bias that are usually discussed with regard to sampling. One common source of bias emerges from the use of a sample frame that does not fully cover the units in the population (the undercoverage problem). The second common source is produced from the failure to obtain data from each of the selected elements in the sample (the nonresponse problem).

Samples are typically selected from lists of units that encompass the entire population (or by using an approximation technique if no such list is available). This list is called a sample frame, and it often only imperfectly represents the population. In part, this is because populations are dynamic, whereas sample frames are static. For example, a list of all the small businesses in the country that one might construct today could be out of date tomorrow if a new business started or an existing one ended. There are methods to update sample frames, but even in

the best circumstances a sample frame is unlikely to be a complete listing of all the elements in a population. To the extent that a sample frame covers the population completely or nearly completely, there is hope for obtaining unbiased estimates of population values from a properly designed sample. To the extent that there is substantial undercoverage—that is, a substantial number of elements of the population are missing in the sample frame—there may be bias in the survey estimates.

One example of sample frame undercoverage is seen in telephone survey sampling. Techniques are available to obtain a good approximation of all the possible telephone numbers in a geographical area or even in the entire country. However, if telephone numbers are to be used as a sample frame, households without telephones will have no chance of being represented in a sample selected from that sample frame. The survey estimates will be of the population of households that have telephones. To the extent that this population differs from the population of households without telephones, the estimate will be biased. Undercoverage is a potential source of bias in survey estimates but it need not be a serious source. If the undercoverage is small, for example, in a medium to large community in which 98% of households have telephones, the amount of bias in information is likely to be small. If only 50% of households in an area have telephones, the bias in information collected from a telephone survey has the potential to be large, and telephone survey methodology would not be recommended. If it is the case that elements are not represented in the sample frame for random reasons, then undercoverage is unlikely to produce bias; however, this is usually not the case.

Another common failing leading to bias is the inability to collect information from some of the selected units. For example, in a confidential survey conducted in a corporation some employees selected into the sample may choose not to participate. The failure to collect information from units selected into the survey is expressed as the rate of nonresponse. The nonresponse rate is the number of individuals (or schools or other organizations) who refused to participate in the survey (or could not be located) divided by the total number of individuals (or schools/organizations) selected into the sample. The response rate is the number of completed surveys divided by the total number in the sample. High nonresponse rates can lead to significant bias in survey estimates. Therefore, it is important to know the response rate of the survey when evaluating the quality of its results.

Questionnaire Design

The sample design determines who will be measured. Questionnaire design determines what is being measured. Social surveys use questionnaires to obtain many different types of information—values, attitudes, opinions, knowledge, behaviors, characteristics, and circumstances. Questionnaire design is a multifaceted activity. It involves first deciding what information is desired and then crafting a questionnaire that is likely to elicit accurate responses. Attention has to be paid to instructions, question format (close-ended or open-ended), question wording, response categories, question order, and the mode of administration. Each one of these characteristics of a questionnaire can affect results. Poor question construction or poor questionnaire design can result in nonsampling error, which is the technical term for error or bias due to respondent reports.

Elements of Good Questionnaire Design

The prototypical survey question has a stem and response format. The stem is the question. For example, a stem for a question about educational attainment might read as follows: "What is the highest level of education you have completed?" The response format may consist of a list of typical education levels. The choice of words in the question stem is important, particularly for attitude and opinion questions. A question that consists of words indicating extreme positions may elicit more disagreement than a question that is semantically equivalent but uses less strong language. In a classic example from the 1940s, fewer respondents agreed that the United States should "forbid" reporters from a communist country to come into this country to report the news back home than agreed that the United States should "not allow" the foreign reporters. A recent study showed that far fewer respondents endorsed spending more government money on "welfare" than on "assistance to the poor." The general recommendation is to avoid asking questions with strong or "loaded" wording.

Questions should be limited to one topic. A question that has a conditional phrase may be "double-barreled"; that is, it may be asking about two things. For example, the question, "Do you think that Congress should increase the income tax to support increased civil defense?" asks about tax increases and increased civil defense. A respondent who favors increasing civil defense but who thinks taxes are high enough already may have difficulty answering this question. Similarly, an analyst may have difficulty interpreting the results. A better strategy would be to ask if the respondent favors increasing civil defense in one question and then ask whether it should be done through tax increases (or perhaps through other means) in a second question.

Questions may be close-ended or open-ended. A close-ended question contains a stem and a set of preselected response categories. An open-ended question has the stem but no response categories; verbatim responses are recorded. There is an intermediate type, the semi-structured question, which has a number of response options but allows the respondent to offer a response that is not one of the options. Most survey questions are of the close-ended type because these types of questions do not require the expensive categorization and coding of responses that completely open-ended questions require. However, in close-ended questions, response selections are limited to those generated by the researcher and may not reflect those that are important to respondents. Either format is legitimate, but when interpreting the results, it is important to know which format was used; results from the same question asked in different formats may not be comparable.

One area in which open-ended questions are particularly useful is in the assessment of behavioral frequency. For example, a question may ask, "How many hours of television do you watch during an average week?" or "How many alcoholic drinks did you have last week?" It is common practice to include frequency ranges, such as "none, one to five, six to ten, eleven to fifteen, more than fifteen," as response categories. However, research has shown that the range of response categories can influence the category that is selected. Thus, estimates of frequencies will differ depending on how the categories are constructed.

The placement of questions in the questionnaire is important to consider. For attitude or opinion questions, related prior questions can affect responses to questions about obscure topics or to questions about which the respondent has mixed feelings or no strong opinion. For example, responses to a question about whether the United States should invade Iraq, a topic over which there was public debate and uncertainty, may be swayed in a positive direction if it is preceded by questions about the largely successful U.S. military activities in Afghanistan or in a negative way if it is preceded by a question about the flagging U.S. economy. These effects are not limited to attitude questions. A question asking respondents how often they go to the dentist followed by a vaguely worded question such as "Does your employer provide insurance benefits?" will almost certainly have respondents thinking about dental insurance. If the researcher is interested in health insurance, the data obtained will be very inaccurate.

Whether a question includes an explicit "don't know" option or not may affect responses to topics that are obscure. With the absence of an explicit "don't know" option, and motivated by a desire to please the interviewer or by the fear of appearing uninformed, respondents will sometimes give opinions when they know very little about a topic. If the survey context suggests an interpretation, then biased responses may be given. Providing an explicit "don't know" category as a response option will mitigate the tendency for uninformed respondents to give substantive responses. Another technique to eliminate asking questions to respondents who know nothing about an issue is to provide a screening question asking whether the respondent has heard about the issue and only eliciting opinions from those who answer affirmatively.

For sensitive topics respondents may use an explicit "don't know" category as a way to avoid giving an answer. This can lead to response bias. For example, if a question that asks smokers how many cigarettes they smoke in a day has an explicit "don't know" option and enough heavy smokers choose this option rather than admitting that they smoke two packs of cigarettes a day, an underestimate of the number of cigarettes smoked will be obtained. A screening question before a sensitive question may also give respondents an easy way to avoid answering the question by simply saying no to the screener. Asking respondents whether they smoke and then asking only those who say yes how much they smoke may lead to an underestimate of smokers compared to a single question that asks how many cigarettes a respondent usually smokes in a day. Those who do not smoke will simply say "none" or "I don't smoke cigarettes." If a topic is sensitive, it may be better to assume that the behavior occurs and allow the respondent to offer a "never" response rather than explicitly providing that option.

Questionnaire Design and Human Cognition

The design of questionnaires has been improved by paying attention to the cognitive processes of the respondent. From the 1940s through the mid-1980s, research on different techniques of asking questions suggested that the way questions were asked affected responses. Research on survey response effects (i.e., finding that different response formats result in different patterns of response), question wording effects (i.e., finding that apparently synonymous terms in the question stem result in very different responses), and question order effects (i.e., finding that response patterns differ depending on preceding questions) suggested that survey questions were interacting with psychological processes of the respondent. However, with few exceptions, very little attention was given to systematically exploring the psychology of the survey respondent. A breakthrough conference in the mid-1980s, the Conference on Cognitive Aspects of Survey Methodology, changed this to a large extent and became a catalyst for research on cognition and survey responding that continues today.

A popular framework that describes the tasks that the survey respondent confronts emerged from this conference. The framework suggests that respondents engage in question interpretation, retrieval of information from memory, judgment formation, and editing for social desirability when they give their answer to a survey question. Consideration of these tasks has been useful in providing a greater understanding of the influence of question type, question wording, response format, and question order on responding. The first three tasks—interpretation, retrieval, and judgment—are, by and large, performed unconsciously. The fourth task, editing, is thought to be strategic, driven for the most part by respondents' tendencies to wish to avoid presenting themselves in an unfavorable way. One of the more enduring fruits of this framework has been a technique for pretesting survey questions called cognitive pretesting in which test respondents are administered carefully developed probe questions. This technique has helped survey researchers create questions that minimize bias introduced by the elements of questionnaire construction discussed previously.

As mentioned previously, questions can be close-ended or open-ended. Attention to cognitive processes reveals that each type of question involves a different cognitive task. The open-ended format clearly involves a recall task. For some types of questions, such as the standard questions used to measure occupation and industry (e.g., "What kind of work do you?" and "What business is it in?") in which the respondent is largely asked to report factual information, recall is straightforward. Other types of open-ended questions require more complicated cognitive processing.

For example, the question, "What is the most important issue facing the country today?" requires respondents first to search memory for candidate issues and then to judge the importance of each issue. The thoroughness of this search and evaluate process may depend on the respondent's motivation or the amount of distraction in the interview setting. A nonthoughtful response may be generated based on some superficial source, such as what he or she had read in the newspaper that morning or had seen on the news the previous night. A completely thoughtful response might require an amount of time that neither the respondent nor the interviewer care to spend. If the question is presented in close-ended format in which a list of response options is provided, the task for respondents is simplified to that of judgment and the respondents can put more effort into it. Studies have shown that very different patterns can be produced under the two question types, even when pretesting has been conducted to try to include in the close-ended format the range of responses that are likely to emerge from an open-ended format. Perhaps the differences are due to the types of cognitive tasks that respondents have to perform under the different formats.

It was previously mentioned that open-ended questions about behavioral frequency are preferable to those that are close-ended because of the influence of the response ranges on the response. Studies have shown that the range of response options may affect the respondent's interpretation of the question. For example, if a set of response options has a range of frequencies that indicate experiences are high in number and the question is "How often have you been criticized in the last month?" a respondent may interpret criticism as involving small incidents. The range of response options may also provide respondents with one or more anchor points from which to estimate their own behavioral frequency. Also, if the question topic is sensitive (e.g., asking how many sexual partners a respondent had in the pastyear) or if there is a social desirability component (e.g., asking how many hours per week one exercises), a respondent may view the middle category as a normative response and choose that to appear neither over nor under the norm.

It was previously mentioned that the use of strong words may bias the respondent into choosing a "lenient" response, for example, agreeing to "not allow" foreign reporters from a hostile nation into the United States but not agreeing to "forbid" them. There are other examples of question wording effects in which similar wording may indicate completely different concepts, perhaps depending on characteristics of the respondent. One study demonstrated a remarkably strong reaction against supporting "welfare" compared to supporting "assistance to the poor." Although these words appear to refer to the same phenomenon, the term welfare may have brought to mind government welfare programs that were unpopular with the public not because they provided assistance to the poor but because they were associated with wasteful bureaucracies and popular notions that they are often abused. Other studies have shown similar results in the area of crime fighting and dealing with the problem of illegal drugs. Sometimes, the effects interact with individual differences, such as political orientation, suggesting that different groups read different things into these sometimes controversial social issues. These results justify the injunction made earlier of keeping the wording of survey questions as neutral and objective as possible.

Studies have shown that prior questions can affect responses to target questions by influencing the retrieval process—that is, the material from memory that respondents bring to mind when answering the target questions. For example, in a public opinion survey administered during a time when the United States was considering military involvement in Nicaragua, questions about the Vietnam War that preceded questions about U.S. military involvement in Nicaragua resulted in more opposition to

U.S. involvement in Nicaragua and in more mentions of the Vietnam War as a reason for the response in a probe question that followed the Nicaragua question. The example about how prior questions about dental care may affect the interpretation of subsequent questions about insurance indicates that the interpretative and retrieval effects of prior questions are not limited to attitude questions.

Memory and Survey Reporting

Many surveys ask respondents to report on life events, such as what activities they have done or products they have consumed, when they have done them, and how frequently. A wide variety of activities are assessed. Recreation surveys might be interested in frequency and dates for visiting a national park or a museum. Surveys of consumer behavior ask about types of products purchased, when they were purchased, and in what quantity. Other surveys ask about life course questions, such as changes in residence, education level, and employment. Health surveys ask respondents to report on health care utilization, expenditures on health care, and courses of illness. Because all these questions rely heavily on the respondent's memory, much attention has been paid to the role of memory in survey responding and to recall or estimation strategies. The goal of this attention is to construct survey questions that help the respondent give accurate and complete reports.

One of the fundamental principles of memory is that accuracy of recall fades with the passage of time. Therefore, if respondents are asked to report on events that occurred a long time ago, it behooves the survey researcher to provide a structure to aid recall or run the risk of obtaining grossly inaccurate information. A number of structures have been developed around the idea of providing a time line. With a time line, respondents are given a physical representation of the recall period, month by month, and are asked to place salient markers such as their high school graduation, when they got married, birth dates of their children, or any relevant personal or public event that they can place accurately in time. Respondents are then asked to use these events as cues to prompt recall of events of interest to the researcher, such as periods of illness, changes in insurance status, changes in employment, and even personal events such as the number of sexual partners during a 5-year period. The literature is not clear as to whether it is better to begin with the present and work backward, begin with the most remote period and work forward, or let the respondent chose the direction.

Accurately placing events in time is difficult for respondents. Events that were salient or memorable seem to have occurred more recently and events that are more difficult to recall may seem to have happened longer ago than was actually the case. The difficulty of placing events in time poses a particular problem for longitudinal surveys. A phenomenon seen in panel surveys in which respondents are asked to report quarterly on changes in status in areas such as marital status, employment, earnings, and welfare payments is that, across the sample, the change in status is greater across the most recent month of the last reporting period and the furthest month of the current reporting period than across any two adjacent months. Since these two months represent the areas where the two waves of the panel survey join, the phenomenon is called the seam effect.

Imagine a situation in which respondents were interviewed in January, April, July, and October for a given year (i.e., once every quarter). During each interview, they were asked to report their total household income. Across respondents, the percentage reporting a change in income from month to month can be calculated within reporting periods (i.e., nonseam months such as January to February, February to March, April to May, and May to June) and across the seam months (in this case, March to April, June to July, and September to October). On average, the calculated change in income across the nonseam months might be 4 or 5%. However, the calculated income across the seam months may be as much as 9 or 10%.

Although the reasons for the seam effect are not well understood, a remedy called bounded recall has been developed. The bounded recall technique involves collecting information for the first reporting period and using the values about the last month of the first period as a bound for respondents. At the beginning of the second wave of interviews, and for all subsequent interviews, respondents are reminded of their answers from the prior interview and are asked to verify the answer. They are then asked to report on their current status. Information collected during the first wave serves only as a bound and is not used in analysis. This technique has proven effective in reducing reporting bias due to the seam effect.

The study of how respondents answer questions requiring event dating and behavioral frequency has led to insights about human cognitive functioning. One phenomenon noted long ago was that respondents tended to date events that happened outside of a recall period as happening within the period. This phenomenon, called telescoping, seemed to increase as the recall period extended. Therefore, respondents were more likely to engage in telescoping if asked to report on events that happened 1 year ago than if asked to report on those happening within the past 4 months. Early attempts to explain this phenomenon used the notion of forgetting curves derived from psychology.

Table I Example of Telescoping Due to Use of Prototypical Values in Estimating Elapsed Time

Event occurred (days)	*Event reported*	*Midpoint col. 1*	*Event occurred (days)*	*Net bias*
	7		8	−1
9–10	10	(8.5)	11–12	−1
13–14	14	(12)	15–17	−2
18–21	21	(17.5)	22–25	−1
26–30	30	(25.5)	31–44	−10
45–60	60	(45)	61–74	1
75–90	90	(75)	91–104	1
105–180	180	(135)	181–364	−109

Another explanation is provided by a fascinating study of reporting bias by Huttenlocher *et al.*, who explain the phenomenon in part by the way we structure time. Table I provides an illustration about how this affects reporting and can account for telescoping. The first and third columns represent the number of days ago an event occurred. The second column indicates when the event is likely to be reported to have occurred. Note that the entries in the second column are prototypical reporting periods: 1 week, 10 days, 2 weeks, 3 weeks, 1 month, 2 months, and 3 months. When making estimates, we tend to use these prototypical values. The numbers in parentheses in the second column are the midpoints between two adjacent reporting dates. The fourth column shows the net bias—that is, the difference between when the event occurred and when it was reported to have occurred. The negative numbers indicate that because we tend to round to prototypical values, events are reported, on average, sooner than they occurred.

To understand how this works, assume that events that occurred between 1 and 7 days ago are reported accurately, but that events that occur more than 1 week ago tend to be rounded to one of the prototypical values. Values in the first column in Table I will be reported at the prototypical value to the right because they are at or greater than the midpoint between that prototypical value and the preceding one. Values in the third column will be reported as prototypical values to the left because they are between that prototypical value and the midpoint for that prototypical value and the following one. Therefore, if an event occurred 8 days ago it is likely to be reported as occurring 1 week ago. Assuming events that occurred between 1 and 7 days ago are reported accurately, the result is a net bias of −1 days for reporting of events occurring between 1 and 8 days ago. An event that occurred 9 or 10 days ago is likely to be reported as occurring 10 days ago (because 10 is another prototypical number in our culture), producing a positive bias that will be no greater than 10−9 or 1 day. However, events occurring 11 and 12 days ago are also reported as occurring 10 days ago, giving a negative bias of no less than 10−12 or −2 days. So far, events occurring between 1 and 12 days ago have a cumulative negative bias of −3 days. If one does the exercise through all the categories in the chart, it is easy to see how telescoping can be explained at least in part by the use of prototypical values when respondents do not remember exact dates.

Survey respondents are asked to report on a wide variety of events, purchases, and behaviors. Since it is highly unlikely that we carry around counts of all these events in our heads, our responses are often based on the use of estimation strategies. A number of estimation strategies have been identified. On a survey of dental behavior, respondents who do not use dental floss regularly may make a wild guess when they are asked to report on the number of times they have flossed in the past month. Respondents who floss with some regularity still may not know exactly how many times they have flossed in the past month, but they may calculate a rate and then extrapolate across the reporting period. For example, a respondent asked to report on the number of fruits and vegetables consumed during a 2-week period might approach the task by reasoning that he or she usually eats five servings a day. Multiplying that by 14 days, the estimated total number of servings might be reported as 70. Another type of estimation strategy begins with a rate and multiplication, as in the previous case, but then employs addition or subtraction for exceptions. Using this strategy, the respondent may begin with an estimate of 70 but realize that she was out of town one weekend visiting friends and did not eat her usual 5 servings on those days. The respondent would make an allowance by subtracting the number of servings (or an estimate of the number of servings) less than 5 that she had on the days that she was out of town. The researcher, knowing that this is an estimated rather than an actual number, can take this into account when judging the quality of the data. In addition, if it is likely that most respondents will estimate their answers, survey instructions can be included to help them estimate more realistically. For the example, asking the respondent to consider whether eating habits on the weekend are the same as during the week might help the respondent adjust the answer to one that more closely matches the true value.

Modes of Data Collection

Social surveys are conducted face to face, by telephone, or by use of a self-administered questionnaire. Many surveys today use multiple modes in order to obtain high response rates and honest reporting of sensitive topics. The social survey began as a face-to-face interview conducted in a respondent's home, and this modality still plays an important role in contemporary survey research. There are many advantages to this modality. Typically, higher response rates are possible with face-to-face interviews. In addition, it is possible for the interviewer to pick up on nonverbal cues from the respondent indicating confusion about the question. Face-to-face interviews allow the interviewer to use visual aids, such as show cards with response categories, time lines (also called event-history calendars) to aid recall, and pictures or other props necessary to enhance respondent understanding. The interviewer is better able to control the pace of the interview and may be able to help the respondent to minimize distractions.

Nonetheless, face-to-face interviews are expensive. When telephone coverage became nearly universal in the United States, telephone surveys became increasingly more popular. They have proliferated at an astounding rate but, although providing great opportunities for researchers, have opened the door to abuse by sales promotions or political campaign solicitations that pretend to be surveys. In general, this has made the public somewhat distrustful of telephone surveys. Telephone survey response rates have been declining slowly but steadily during the past two decades.

Still, telephone surveys can offer an inexpensive alternative to a face-to-face interview survey and can provide useful data when conducted carefully. To increase response rates, many attempts must be made to call sampled households during different times of the day and days of the week. When possible, letters explaining the importance of the survey should be sent to the household in advance of the interview. In questionnaire construction, attention must be paid to the fact that there is no visual contact between the interviewer and that the respondent must process all information sequentially. Certain survey tasks, such as rank-ordering long lists of items, cannot be accomplished readily in a telephone survey because of the limitations imposed by the inability to use visual displays.

Typically, telephone survey questionnaires should not be longer than 30 minutes, whereas face-to-face interviews as long as 3 hours have been conducted. Undercoverage is greater among population members with low incomes, and technologies such as answering machines, caller-ID, and call blocking have made it more difficult and expensive to obtain high response rates. Sometimes, the anonymity associated with a telephone survey in which the respondent is aware that his or her household was selected completely at random can lead to candor on topics that are not highly sensitive but might have a social desirability component leading to a biased response if asked in a face-to-face format. However, telephone survey respondents can more easily misrepresent characteristics of themselves because of the interviewer's inability to directly observe them or their surroundings. Nonetheless, telephone surveys, and short telephone polls, continue to represent a viable data collection methodology.

Self-administration has been employed in three ways in survey research and a fourth method of self-administration is rapidly gaining popularity. First, a substantial amount of survey research of individuals and institutions is conducted by mail, in which a targeted respondent or household is sent a self-administered questionnaire. There are established techniques for obtaining high response rates in mail surveys, and high-quality data can be collected inexpensively if one follows the procedures, but mail surveys have limitations. One limitation is that it is difficult to know whether the targeted respondent actually completed the survey. Another limitation is that it is easy for respondents to leave questions unanswered and they may, deliberately or inadvertently, skip entire pages or sections of the questionnaire. Carelessly conducted mail surveys can result in very low response rates (as can carelessly conducted telephone surveys), so it is important to have information about response rates when assessing the quality of information from mail surveys. However, great success has been obtained with mail surveys using attractive envelopes and carefully laid-out and short questionnaires and by using a carefully crafted introduction, endorsement, and appeal letters sent at carefully determined intervals.

Self-administered questionnaires have been used to good effect when surveys are administered in group settings. A number of national surveys of elementary and secondary school students have been conducted in this way. If the students are well supervised by trained interviewers, the method can yield high-quality data. The interviewer can check questionnaires to determine if students have omitted critical questions and can ask them to give an answer (even if that answer is a "refused" response). Self-administered questionnaires have also been used in surveys in the workplace. Social norms, incentives, and the promise of anonymity can result in high response rates and high-quality data, provided that the questionnaire is crafted properly.

Self-administered questionnaires are used in the collection of sensitive information, but this topic is discussed in the next section. The most recent innovation in self-administered surveys involves the Internet. Internet and Web-based surveys have proliferated during the past

few years. Using the Internet to collect survey data has exciting possibilities. Short, simple questionnaires in the text of electronic mail, or attached as a document, can be distributed quickly and cheaply to a sample or an entire population of a group with universal electronic mail coverage, such as the faculty on all of the campuses of a large state university or employees of a large multinational corporation. Follow-up reminders to nonrespondents can also be done quickly and cheaply. Responses still have to be coded and compiled, but the potential for quickly collecting information of high quality on important but nonsensitive topics is great.

Surveys conducted via the World Wide Web (WWW) also offer exciting possibilities. As with Internet surveys, coverage is important. Unbiased sample estimates from populations with near universal access to the WWW are possible if good lists of unique Internet addresses are available. Techniques can be used to make it likely that the targeted respondent is the one actually completing the survey and that the responses are confidential and secure. For both Internet and Web-based surveys, coverage is important. Conducting such surveys on a general population, where access to the Internet is not universal, is likely to lead to biased estimates. Web-based surveys offer all of the advantages of computer-assisted surveys. Skip patterns are easily built-in so that the questionnaire can accommodate some level of complexity. Respondents can be prevented from answering questions out of sequence and from skipping a question without providing at least a "don't know" or "refused" response. Visual stimuli, such as pictures and even film clips, can be utilized and the time it takes respondents to answer a question (known as response latency) can be surreptitiously collected along with the responses.

Sensitive Questions

Sometimes it is necessary to obtain information about sensitive topics in surveys. For example, one result of the AIDS epidemic of the 1980s is that today it is not uncommon for surveys to contain explicit questions about sexual behavior. Surveys on topics such as smoking, drinking, and consuming illegal drugs are also common. One can imagine a whole host of behaviors to ask of the general population or of subpopulations that would be sensitive because they are embarrassing to talk about (e.g., sexual practices or problems, health problems, and gambling behavior) or illegal (e.g., owning a handgun that is unregistered, cheating on income tax, and employing illegal immigrants), yet it may be important, for policy reasons, to have accurate estimates of the numbers of people who engage in the behavior.

Special survey procedures have been developed to reduce the underreporting of sensitive behaviors that may result from fear of embarrassment or fear of self-incrimination. One early procedure, which is still in use, is called the randomized response technique. This method was designed to obtain aggregate estimates of sensitive behaviors while maintaining respondent privacy by combining responses to questions about sensitive behaviors for which rates are not known with questions about events for which rates are known. This is illustrated in the following example. The respondent is shown a card with two questions. Question A is the sensitive question for which the population proportion is not known (e.g., "Have you watched a pornographic movie in the past month?"). Question B is a nonsensitive question for which the population proportion is known (e.g., "Does your birthday fall in the month of July?"). Next, the respondent is asked to select a bead from a box containing 50 beads while the interviewer looks away. Seventy-percent of the beads (35 beads) are red and 30% (15 beads) are blue. The respondent is told to give a truthful answer to Question A if a red bead is selected and a truthful answer to Question B if the blue bead is selected. Because the interviewer does not know the color of the bead, he or she does not know which question the interviewer answered truthfully.

Assuming respondents do as they are instructed, the proportion engaging in the sensitive behavior can be estimated using simple algebra. In this example, the estimate of the percentage admitting to having rented a pornographic movie in the past month is $100 \times [(P(\text{Yes}) - (30/12))/70]$. Say 20% of the sample gave a yes response across the two questions. If we assume that birthdays are distributed evenly across months, then we expect that the 1/12 (2.5%) of the 30% who received the birthday question were born in July. The percentage that said yes to the pornography question is estimated by first subtracting 2.5% from 20%, giving 17.5%. This percentage is divided by the percentage that received the pornography question (i.e., 70%), and the result is multiplied by 100. The final result is the estimate that 25% of the population from which the sample was drawn viewed a pornographic movie in the past month. It should be noted that this technique will work best with large samples of 1000 or more because it is only in large samples that the distribution of birthdays and beads across respondents will conform to the expected values. An elegant simplification of this method, with examples of how to use it to study the rate of illegal immigration, has been developed by the Government Accounting Office.

A drawback of this technique is its inability to link the sensitive information to individual respondent characteristics. Thus, although it has utility in providing aggregate estimates of sensitive behaviors, it is of limited use in modeling the sensitive behaviors. Other methods for eliciting truthful responses to sensitive behaviors that do increase the likelihood of reporting and also permit

modeling behavior at the individual level have been developed. A popular, and obvious, method is to introduce privacy into the interview through the use of a self-administrated questionnaire. For example, in a household survey in which an interviewer is reading questions to a respondent in the respondent's home, the survey may be designed so that the part of the survey with the sensitive questions is completed using a self-administered questionnaire. The respondent can then put the questionnaire in a sealed envelope and either return it to the interviewer or, in some cases, mail it. This gives the respondent both privacy in responding and, if the questionnaire is mailed, the assurance that the interviewer will not "peek" at the responses after the interview.

Currently, most large sample household surveys use a computerized questionnaire. In the case in which it is desirable to ask about sensitive topics, a computerized self-administered section of the interview can be utilized. Research has indicated that respondents are more likely to report sensitive behaviors on a self-administered questionnaire than directly to an interviewer, and they are even more likely to report sensitive behaviors when the self-administered questionnaire is computerized. With current technology, it is possible to add an audio component to the computerized self-administered questionnaire. The respondent sits at the keyboard, sees the questions on the screen, and also hears a recording of each question through a headset. The addition of the audio component appears to elicit the most sensitive behavior. Why this is so is not well understood.

Telephone surveys are common because they are less expensive to conduct than household surveys. Traditionally, the telephone survey has not been a good medium for asking sensitive questions in part because of the difficulty of allowing the respondent to answer questions privately. However, promising new technology may change this. Devices that allow respondents to key in their answer using the telephone touch pad (e.g. "Have you ever smoked a marijuana cigarette? Press 1 for yes and 2 for no") are being developed for surveys. These telephone data entry devices can input the response directly into the respondent data file without the interviewer, or someone who happens to be eavesdropping on a second line, knowing how the respondent answered. In addition, completely automated surveys, in which a recorded voice asks the question and the respondent enters a response on the telephone touch pad, are being explored.

Conversational Aspects of Surveys

Although survey practitioners have always acknowledged the human side of survey research—that it consists of two people engaged in a conversation of sorts—the traditional approach has been to standardize the context in which the conversation occurs to minimize the influence that interviewers have on the response. Interviewers were (and are still, for the most part) trained to read questions exactly as they are written; to avoid getting into extended dialogue with the respondent about the meaning of terms; and, to the extent that explanations are necessary, to rely on those presented by the researchers in advance of the interview. Recently, a flaw in this practice has been noted. Although the purpose of the standardization is to improve the data, the rigidity of the conversation may lead to misunderstanding. This is because in most human encounters, meaning is conveyed through discourse. As a consequence, some survey researchers are attempting to understand the survey question-and-answer process from the standpoint of natural language. At the theoretical level, a survey question is an interrogative statement that communicates a space of uncertainty from which responses can be drawn. If the uncertainty space conveyed by the question does not match a space for which a respondent has propositions, then no meaningful answer can be obtained. This might be the case when a question asks a respondent about issues for which the respondent has no knowledge, or when a question contains obscure terminology. This may also be the case when a question asks the respondent to engage in a virtually impossible recall task, such as listing every item the respondent bought from a store during the past week.

Questions also have presuppositions, which are descriptive statements that indicate what the question is about. For example, the question, "What time do you usually leave for work in the morning?" presupposes that the respondent has a job and that it is a daytime job. Presuppositions restrict the uncertainty space. In the preceding question, the uncertainty space has been restricted to information about daytime jobs. If a respondent does not have a job or does not have a daytime job, then the uncertainty space is so restricted for that respondent that he or she cannot answer the question.

Presuppositions have another feature. They can serve as cues to answers and can result in leading the respondent to answer in a certain way. For example, research in cognitive psychology has shown that people who viewed a videotape of a car accident and were asked how fast the cars were traveling when they "crashed" gave higher speed estimates than people who were asked how fast the cars were traveling when they "hit." The presupposition provided by the descriptor crash was that the cars were going fast. This same principle may underlie the finding that a question asking whether something should be "forbidden" is less likely to receive agreement than a question asking whether that same thing should be "not allowed." Although the phrases have the same meaning, the presupposition provided by the word forbidden

seems to indicate much more finality or strictness than that provided by the phrase not allowed.

The attention to natural language in surveys is leading to methods that show promise in eliciting accurate responses to questions involving long-term recall of events. Sometimes it is useful to have information about the history of different events in a respondent's life. For example, a researcher may want to know about all of the places a person has lived during his or her life, how many times he or she has changed jobs and when the changes were made, the types of insurance coverage people have had, the number and type of cars that they have owned, their incomes in different phases of life, or whether their children received all the necessary medical examinations and inoculations prior to starting school. Traditional methods focus separately on each type of event and ask the respondent to start from a point in the past and report forward in time or to start at the present and report backward in time. New methods are being developed that take into account the idea that people may store events not as discrete entities in memories but as parts of narrative chains. Encouraging respondents to engage in these narratives when they are answering questions about their past, for example, allowing them to pick any time point within the range and to begin to tell a story about their circumstances at that time, has been shown to improve recall for target events.

Conclusion

This article is intended to provide an overview of social surveys with an emphasis on instructing the reader in basic concepts, discussing new thinking on questionnaire design, and illustrating how different sampling and questionnaire design issues affect the quality of the data. These and other important topics are covered in more detail in the books and articles listed in Further Reading. The interested reader should consult these and, to keep up with new developments, read journals such as *Public Opinion Quarterly* and *Survey Methodology*, as well as journals in the area of market research.

Survey research in the United States began in earnest in the mid-1930s. Since that time, survey researchers have taken advantage of new methodologies as they emerged to improve upon the information they gather. Early surveys were conducted using quota techniques until advances in sampling theory demonstrated how more accurate results could be obtained through the scientific sampling of survey respondents. For a long time, in-person household surveys were the norm, and they still play an important role in the collection of high-quality survey data. However, as telephones became more common in households throughout the United States, telephone survey methods were developed. Using telephone survey methodology, useful information could be obtained at far lower costs than in-person household surveys. Similarly, the mail survey technique has been highly developed, and it provides a cost-effective method for obtaining useful survey data. Household, telephone, and mail surveys each have their uses, strengths, and limitations. Survey researchers were also quick to take advantage of the development of personal computers, and personal data assistants offer more possibilities, as do portable scanning devices and touch-tone data entry from telephones.

The development of the WWW has also presented opportunities for survey researchers. Web-based surveys are becoming more common, and creative survey researchers are finding ways to use the Web to good advantage while overcoming the challenges Web surveys impose. This is the latest frontier of survey research, and it shows much promise. Based on past experience, it is likely that Web-based surveys will not completely replace the other survey modalities but will add an option to the survey researcher's tool chest that is appropriate in certain situations but not in others. Together, these methodologies give social scientists flexible, powerful, and efficient methods to study important social issues.

See Also the Following Articles

Clustering • Cognitive Research Methods • Interviews • Longitudinal Studies, Panel • Population vs. Sample • Stratified Sampling Types • Survey Questionnaire Construction

Further Reading

Belli, R. F., and Shay, W. L., and Stafford, F. P., and (2001). Event history calendars and question lists. *Public Opinion Q.* **65**(1), 45–74.

Belson, W. A., and (1981). *The Design and Understanding of Survey Questions.* Gower, Aldershot, UK.

Dillman, D. A., and (1978). *Mail and Telephone Surveys: The Total Design Method.* Wiley, New York.

Dillman, D. A., and (2000). *Mail and Internet Surveys: The Tailored Design Method.* 2nd Ed. Wiley, New York.

Droitcour, J., and (1999). *Survey Methodology: An Innovative Technique for Estimating Sensitive Survey Items.* U.S. Government Accounting Office, Washington, DC.

Groves, R. M., and (1989). *Survey Errors and Survey Costs.* Wiley, New York.

Groves, R. M., and Biemer, P. P., and Lyberg, L. E., and Massey, J. T., and Nicholls, W. L., and Waksberg, J. (1989). *Telephone Survey Methodology.* Wiley, New York.

Hippler, H. J., Schwarz, N., and Sudman, S. (1987). *Social Information Processing and Survey Methodology.* Springer-Verlag, New York.

Huttenlocher, J., Hedges, L. V., and Bradburn, N. M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *J. Exp. Psychol. Learning Memory Cognition* **16,** 196–213.

Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines.* National Academy Press, Washington, DC.

Kish, L. (1965). *Survey Sampling.* Wiley, New York.

Laumann, E. O., Gagnon, J. H., Michael, R. T., and Michaels, S. M. (1994). *The Social Organization of Sexuality: Sexual Practices in the United States.* University of Chicago Press, Chicago.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context.* Academic Press, New York.

Schwarz, N., and Sudman, S. (1995). *Answering Questions.* Jossey-Bass, San Francisco.

Sudman, S., and Bradburn, N. (1982). *Asking Questions: A Practical Guide to Questionnaire Design.* Jossey-Bass, San Francisco.

Sudman, S., Bradburn, N., and Schwartz, N. (1996). *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology.* Jossey-Bass, San Francisco.

Tourangeau, R. (1984). Cognitive science and survey methods. In *Cognitive Aspects of Survey Design: Building a Bridge between Disciplines* (T. Jabine, M. Straf, J. Tanur, and R. Tourangeau, eds.), pp. 73–100. National Academy Press, Washington, DC.

Tourangeau, R., Rips, L., and Rasinski, K. (2000). *The Psychology of Survey Response.* Cambridge University Press, New York.

Taxation

John Hudson
University of Bath, Bath, United Kingdom

Joweria M. Teera
University of Bath, Bath, United Kingdom

Glossary

compliance costs The costs incurred by individual firms or taxpayers in administering their tax affairs.
tax avoidance Tax legally avoided by taxpayers via the exploitation of loopholes in the law.
tax effort A measure of the effort a country makes in levying taxes, generally measured as a ratio (for example, of gross domestic product).
tax evasion Tax illegally evaded by taxpayers.
tax handles The means that facilitate the ability of governments to raise tax revenue.

Initially, perhaps, it would appear that taxation is one concept with which there should be few measurement problems. Tax rates and threshold points are clearly defined in virtually all, if not all, countries; using existing tax regulations, it is relatively easy to determine what a person with a given income and set of socioeconomic characteristics should pay in taxes. The relevant characteristics vary from country to country, but tax codes commonly relate to marital status and number of children or other dependents in the family and to what extent money is earned by work, is a transfer payment from government, and is accumulated from savings or investment in the stock market. However, even at this basic level, there are problems: direct taxation—the taxation of individual incomes—is only a part, and possibly a declining part, of total taxation. Alternative forms of taxation include indirect taxation (taxation on expenditures), taxation on firms, taxation on imports, and even taxation on exports. In addition, taxes can be levied, within the context of the United States, at the local, state, and national levels; within Europe, there is the possibility that taxes may at some point in the future be levied at the supranational level. In reality, the problem is much more complex than may appear at first sight. Among the many complexities, the focus here is primarily on three issues: (1) measuring the tax effort and other comparative measures of tax performance, (2) estimating the compliance costs of taxation, and (3) estimating the extent of tax evasion. In addition, a brief analysis is offered of some of the wider literature that focuses more on measuring, or estimating, the impact of taxation.

Tax Performance

In practice three concepts, tax ratio, tax effort, and taxable capacity, have been used to measure tax performance; sometimes these concepts are used interchangeably. The tax ratio is simply the ratio of tax revenue to some tax base, often gross national product (GNP) or gross domestic product (GDP). The tax effort appears to be widely used in policy circles. For example, two, of many, recent documents from the International Monetary Fund (IMF) have tax effort as a central feature of their dialogue with individual countries, including Pakistan and Paraguay. These documents mention the need to increase tax effort by strengthening tax collection and other measures, and on occasion provide specific targets for total tax revenue as a percentage of GDP.

It may be argued that tax ratios are rather simplistic as a measure of tax capacity, because capacity is also affected by, among other factors, the size of the population, the

availability of "tax handles," and the degree of monetarization. Taxable capacity is a concept that has been interpreted in a number of different ways; some define it as the amount of tax that could be justly or fairly imposed on an individual, or as the ability of people to pay tax and the ability of governments to collect. Others define it as the amount of revenue a country could have raised if it had applied the average effective rates (AERs) to its bases (i.e., if it had made an "average" effort).

Individual Tax Ratios

Tax ratios have not been limited to simplistic definitions; rather, their development has related to specific taxes. These more complex relationships still suffer from problems associated with taxable capacity, but allow a finer analysis of where tax effort is being made and also a comparison of trends over time and differences between countries. The starting point for analysis lies with the difficulties associated with measuring marginal effective tax rates; these problems have led a number of researchers to suggest relating actual tax revenues to certain macroeconomic aggregates. The resulting tax ratios are collectively known as "average effective tax ratios" (AETRs), or implicit tax ratios. This concept was operationalized in the 1990s in an approach subsequently updated by Carey and Tchilinguirian. In taking this approach, a number of assumptions need to be made. With respect to the AETR for capital (Ω_k), the following formula is used:

$$\Omega_k = \frac{\{\Omega_h \cdot (Y_{UB} + R) + T_Y + T_P + T_{FT}\}}{S_O}, \qquad (1)$$

where Y_{UB} is unincorporated business net income; R is interest dividends and investment receipts; T_Y is taxes on income, profit, and capital gains of businesses; T_P is recurrent taxes in immovable property; T_{FT} is taxes on financial and capital transactions, and S_O is operating surplus. The AETR on household income, Ω_h, is

$$\Omega_h = \frac{T_Y}{Y_{UB} + R + W}. \qquad (2)$$

In Eq. (1), $\Omega_h \cdot (Y_{UB}+R)$ represents estimated taxes paid by households on unearned income linked to business activity. In this estimation, it is assumed that the same rate of tax applies to earned and unearned income, an assumption clearly violated in many countries. The need to make such assumptions makes such measurements less than perfect. The other terms in the numerator, $T_Y+T_P+T_{FT}$, then represent the remaining business taxes. The changed assumptions made by Carey and Tchilinguirian vary from tax to tax and indeed from country to country. Thus with respect to Eq. (2) they take account additional taxes, such as wealth taxes and estate, inheritance and gift taxes which the original approach had ignored.

The results of doing these calculations for various taxes are shown in Table I. The figures are informative for several reasons. First, they illustrate the importance of assumptions. It can be seen, for example, that the revised calculations are substantially different from those using the original methodology, particularly with respect to capital taxation, whereby the estimates are substantially higher than previously. The importance of choice of methodology is also apparent with respect to the choice of base. When, for example, gross operating surplus is used, much smaller figures are derived compared to using net figures. This in itself is not surprising, but the impression created, of exceptionally high taxation, disappears when gross figures are used (for example, with respect to Japan when using net figures with the revised methodology). But the figures are also informative in terms of what they tell us about (1) taxation and trends in taxation and (2) the extent to which they confirm the predictions of economic theory. Thus, it can be seen that the United States and Japan have substantially lower AETRs on labor, compared to the European Union (EU) countries, but that within the EU, the United Kingdom is more closely aligned with non-EU countries. Using these figures, economists and commentators have argued that the majority of the EU countries are less competitive with respect to labor and labor costs compared to their competitors in international markets, and this may help to explain job losses in the EU at the end of the 1990s. It also apparent that there is a considerably greater degree of harmonization, as reflected by smaller standard errors, of tax rates within the EU than within the Organization for Economic Cooperation and Development (OECD) as a whole. This is what would be expected, because the close economic union of these countries is forcing harmonization via both legislation and the mobility of the factors of production and consumption. This is less obvious for labor than for consumption and capital, but if the United Kingdom (which, as already noted, is something of an outlier in the EU) is excluded, then it is also true for labor.

The Behavioral Approach

Perhaps the most satisfactory way of measuring tax effort is what is known as "the behavioral approach," where the tax ratio (T/Y) is regressed on a vector of variables ($\mathbf{X}$) that serve as proxies for a country's tax handles:

$$T/Y = f(\mathbf{X}) + e, \qquad (3)$$

where e is a stochastic error term and $f(\mathbf{X})$ then forms the predicted tax ratio against which the actual tax ratio is compared, so that a tax effort index is computed as the ratio of the actual to predicted tax ratio. Variables that have been included within $\mathbf{X}$ include income per capita,

Table I Estimating Average Effective Tax Ratios[a]

Methodology/ region	*Capital based on net operating surplus*			*Labour*			*Consumption*		
	1980–1985	*1986–1990*	*1991–1997*	*1980–1985*	*1986–1990*	*1991–1997*	*1980–1985*	*1986–1990*	*1991–1997*
				Original methodology					
United States	39.5	39.1	40.9 [27.3]	25.3	25.9	26.7	5.5	5.0	5.2
Japan	38.1	46.2	41.6 [24.1]	24.9	29.6	24.1	4.8	5.3	6.0
Germany	29.6	26.5	25.1 [15.5]	38.6	40.6	41.4	15.1	14.7	15.8
United Kingdom	67.8	61.2	48.2 [31.9]	27.5	25.2	23.7	16.5	16.7	16.7
France	28.7	26.3	26.8 [17.0]	42.6	45.9	47.2	20.5	20.2	19.1
Switzerland	27.8	36.4	35.0 [29.2]	50.9	54.4	52.3	7.6	8.2	8.0
European Union	32.2	33.6	32.6 [21.2]	38.8	41.2	42.8	17.1	19.2	19.3
OECD	32.4	34.9	34.7 [22.0]	33.1	35.4	36.8	14.4	16.1	16.5
OECD Std Dev	13.4	13.8	10.8 [6.0]	11.7	12.0	12.1	7.9	8.0	7.6
EU Std Dev	14.5	14.9	10.6 [5.9]	8.6	9.0	9.4	5.4	5.0	3.9
				Revised methodology					
United States	50.6	48.8	51.0 [31.1]	21.6	22.1	22.6	6.3	5.9	6.1
Japan	108.7	98.8	83.6 [32.6]	20.1	23.1	24.0	6.4	6.2	6.7
Germany	47.6	39.4	36.4 [19.9]	33.1	34.8	35.9	14.8	14.6	15.8
United Kingdom	53.3	41.5	41.4 [38.4]	24.3	22.3	21.0	16.0	16.4	16.9
France	53.3	41.5	41.4 [23.6]	35.4	38.5	40.2	18.8	19.0	18.0
Switzerland	49.2	71.8	75.6 [30.5]	27.2	28.1	30.2	8.5	8.9	8.4
European Union	48.4	46.9	45.3 [25.1]	33.0	35.3	36.8	16.6	18.6	18.7
OECD	51.7	52.2	52.2 [26.6]	30.0	32.2	33.4	16.1	17.2	17.1
OECD Std Dev	21.9	21.2	17.7 [6.2]	8.1	8.3	8.6	6.2	6.0	5.5
EU Std Dev	18.9	19.7	13.2 [6.2]	7.9	8.2	8.3	3.8	3.4	2.6

[a] Data from Carey and Tchillinguirian (2000). Numbers in brackets are based on gross operating surplus.

openness of the economy to trade (as measured, for example, by the ratio of imports or exports to GDP), a measure of industrialization (such as the share of agricultural income in GDP), and population density. The latter affects the efficiency of the tax collection system. Differences in the level of tax revenue have also been linked to the impact of government legitimacy, efficiency, and credibility on taxpayers' compliance. Such political and cultural factors clearly affect the taxable capacity of a country. But is it reasonable to link this with tax effort? To an extent, this is a semantic discussion, but, clearly, if a democracy has greater legitimacy than a dictatorship, and because of this tax compliance is greater, then in a real sense tax effort is greater in the democracy.

This approach has recently been extended to include, as well as the usual variables included in **X**, a measure of tax evasion. This is found by a variation on the cash balances method (see later). The impact of tax evasion on the ratio of tax to GDP is ambiguous in that theoretically, tax evasion can reduce both the numerator (tax revenue) and the denominator (GDP) in Eq. (3). Normally, the former would be expected to decline more than the latter, because parts of GDP are effectively free of, or pay little, tax (low earners). However, if the government is able to compensate for the tax shortfall by higher tax rates and more extensive taxes, then GDP will fall by more than tax revenue and tax evasion will have a positive impact on tax ratios. For individual taxes, the effect will depend on whether that tax is sensitive to evasion (see Table II). All of these approaches are static approaches that measure tax potential at a given point in time. It is also important in determining tax effort to look at changes over time. For example, a country with a low tax effort may nonetheless have been making considerable efforts in recent years to increase taxes, and hence the poor performance is a measure of the impact of past policies rather than current ones. This leads some scholars to argue for comparing the income elasticity of taxes. This can be estimated by regressing the log of tax revenue (T) on a constant and the log of income for a country over time:

$$\ln(T_t) = \alpha + \beta \ln(Y_t) + e. \quad (4)$$

The coefficient β gives the estimate of the income elasticity, i.e., the percentage change in tax revenue divided by the percentage change in income.

Table II Tax Effort Indices for Selected Countries[a]

Country	*Tax effort index*
Australia	0.924
Belgium	1.820
Canada	0.745
Denmark	1.391
France	1.609
Germany	1.185
Italy	1.490
Japan	0.540
Spain	1.090
Sweden	1.453
United Kingdom	1.401
United States	0.785

[a] Original calculations by J. M. Teera.

Measuring Compliance Costs of Taxation

Compliance costs may be defined as those costs that are associated with complying with the requirements of a tax system, over and above any distortion costs inherent in the tax. They do not include the administrative costs of taxation, which are borne by the tax authorities. The quantitative information on compliance costs is largely derived from responses to large-scale questionnaires, i.e., by directly asking taxpayers what are the costs they incur in paying such taxes. This important concept is linked to Adam Smith's fourth "canon of taxation." Government departments are increasingly concerned with minimizing compliance costs, but corrective policy requires an understanding of the causes of such costs, and this can be obtained only from regression analysis. One such study, with respect to the British income tax system, whereby firms deduct tax from employees at source and then transfer this money to the government, found significant economies of scale that put small firms at a considerable disadvantage to larger firms. Further unpublished work on this database for the Inland Revenue, the government department in the United Kingdom that is responsible for collecting income tax (the counterpart to which is the Internal Revenue Service in the United States), suggests that approximately 40% of compliance costs is due to inefficiencies. This is not surprising; especially for small firms, their expertise lies, e.g., in metal manufacture, not in being tax collectors. The chief problem in this area is collecting the data. In the study of the British income tax system, the data were obtained by a large-scale postal questionnaire using a stratified random sample of 5195 employers' payrolls. Stratification was based on the number of taxpayer records in each band. Respondents were chosen so as to ensure roughly equal sizes across bands. To ensure that the questionnaire was received by the appropriate person, initial telephone contact was made with all 5195 employers. The final response rate was 30.2%. Estimates of compliance costs were based on a "bottom up" approach, whereby respondents were asked detailed information on time taken and the

estimated hourly wage of the person engaged on this task. This information was then added to associated costs relating to computer hardware/software, tax advisors' fees, payroll bureau charges, and miscellaneous costs such as telephone, stationary, heating, etc. The data suggest that compliance costs amount to 1.3% of revenue raised. Similar figures for other taxes have been found in other studies.

Compliance costs are rarely less than 2%, and in the case of some taxes in The Netherlands and Australia, they are well above 10%. These costs are not insignificant and in some cases are relatively high, probably indicating an inefficient tax. In addition, such costs can have a significant impact on businesses and individuals. Estimates in one study put the compliance costs associated with corporation income taxes in the United States at over $2 billion. Such data have frequently led governments to seek ways to reduce compliance costs, as has specifically happened in the United Kingdom.

Measuring Tax Evasion

Tax evasion represents a potentially serious loss of revenue to governments, resulting in the possible underfunding of public service and an "unfair" burden falling on honest taxpayers. In the United States, it has been estimated that over 25% of all taxpayers underpaid their taxes by $1500 or more in 1988. In developed countries, tax evasion is frequently estimated to be at about the 20% level of tax revenue. The estimated loss in revenue in the United States in 1992 through underpaid federal income taxes was $95.3 billion. In developing countries, the problem may be worse; the loss in the Philippines, for example, has been estimated to be as much as 50% of income tax revenues.

Empirical work has also focused on the link between tax evasion and socioeconomic characteristics. There would appear to be considerable evidence that evasion declines with taxpayer age, and is more common among men and in households in which the head of the household is married. Measurement of the size of socioeconomic effects and the deterrent effects of audit probability, fines, or penalties necessitates the use of multiple regression analysis. The evidence that is available suggests that both penalties and audit probabilities have significant deterrent impacts on evasion, although the extent of the impact is not clear. In addition, it seems possible that the probability of detection is more important than the fine in deterrence.

Much of the empirical work on tax evasion is centered on the United States and is based on audit data or tax amnesty data. Both types of data suffer from an element of bias, the former because auditors are generally unable to detect all evasion, the latter because only those evaders who respond to the amnesty will be included in the data set. Survey data can be used to circumvent this problem, although self-reporting of actual evasion imports bias, even with confidential surveys. An alternative approach is to use survey responses to hypothetical questions. One group of researchers looked at a question related to hypothetical situations involving collusion with a builder; the builder would offer the individual a lower price to do a job if the individual would pay in cash, hence obviating the need for the builder to declare the income to the tax authorities. A further question relating to the evasion of income tax was also asked. In both cases, in excess of 50% of the population indicated that they would indeed engage in such behavior, a tendency that was greater for the young than for the old. This use of hypothetical questions to analyze real-world problems may have potential value in other areas.

Measuring the Extent of Tax Evasion

There are inherent and obvious difficulties in measuring the extent of tax evasion. Surveys are clearly inappropriate and hence recourse has to be made to indirect methods. Tax evasion is synonymous with the hidden or shadow economy that relates to unrecorded economic activity, generally for reasons of avoiding tax. The first attempt at estimating unrecorded national income was done by Nicholas Kaldor in 1956. Over the years, the methodology has steadily become more sophisticated. A methodology employed in the 1990s assumes that an economic activity M (frequently narrow measures of the money supply) is required in all k sectors/regions or industries of an economy; the level of activity M is determined by the income and other variables (Z_{jt}) related to the k sectors. The assumption is then made that

$$M_{jt} = f_j(Y_{jt}, Y_{\mathrm{H}jt}, \mathbf{Z}_{jt}). \tag{5}$$

In general, the jth sectoral/regional observations on M are unavailable and hence an estimate is made using multiple regression techniques:

$$M_t = \sum_{j=1}^{k} f_j(Y_{jt}, Y_{\mathrm{H}jt}, \mathbf{Z}_{jt}), \tag{6}$$

where Y_{jt} is legitimate (measured) income and $Y_{\mathrm{H}jt}$, hidden income. Of course, $Y_{\mathrm{H}jt}$ is unobservable and various proxies are used. These proxies, together with their estimated coefficients, allow construction of estimates of the size of the hidden economy. This approach does have its weaknesses (for example, in frequently ignoring the possibility that money demand, or whatever proxy is used, may be changing for reasons unrelated to the size of the hidden economy.

A variation on this theme is to estimate the size of the hidden economy on the assumption that the difference between the growth rates of measured GDP and electricity

Table III Estimates of Size of Shadow Economy as Percentage of Gross Domestic Product[a]

	Percentage	
Country	*1994–1995*	*1996–1997*
Australia	13.8	13.9
Canada	14.8	14.9
Germany	13.5	14.8
Italy	26.0	27.2
Sweden	18.6	19.5
Switzerland	6.7	7.8
United States	9.2	8.8

[a] Data from Schneider and Enste (2000); estimates derived using the currency demand approach.

consumption can be attributed to the growth in the shadow economy. All such approaches are based on simplifying assumptions; the electricity based approach, for example, is subject to the criticisms that (1) not all shadow economy activities require a considerable amount of electricity and other energy sources can be used and (2) that shadow economy activities do not take place solely in the household sector. An alternative approach pioneered in the 1980s used the multiple indicators, multiple causes (MIMIC) methodology. Essentially, this treats the size of the underground economy as an unobservable "latent variable." The latter is linked on one hand to a set of observed "causal variables," which are believed to be key determinants of the hidden economy. MIMIC methodology can use the following determinants of the hidden economy: direct tax share, indirect tax share, share of social security contributions, increase in direct tax share, share of public officials, tax immorality, rate of unemployment, and per capita disposable income. The "indicator variables," all of which are assumed to be partly constituted by the latent variable (the hidden economy), may be the male participation rate, hours worked, and the growth of real GDP. Of course, the effectiveness of this approach is determined by the appropriateness of the indicator and determinant variables.

Table III shows that the shadow economy in most countries is estimated to be of the order of 15%, but it is much higher for Italy and much lower for Switzerland. Again, these results are typical, as are the much higher figures that are obtained for developing or transition economies. Clearly, in this case, the data point to a considerable problem facing governments seeking to raise revenue to finance public expenditure.

Measuring the Impact and Incidence of Taxation

There is a considerable volume of literature that seeks to estimate the impact of taxation on individual behavior. There is a specific focus on the effects of taxes on labor supply with respect to individuals and the impact of tax on organizational form, investment decisions, merger and acquisition and dividend polices, accounting choices, and compensation decisions with respect to firms. Much of the literature spotlights the distortionary impacts of taxes that are viewed as undesirable. However, there is also a significant literature on the use of taxes to steer individual decision making in a beneficial fashion. Carbon taxes and taxes on tobacco and alcohol are common examples. An applied general equilibrium model was used to look at the impact of the carbon tax introduced in Norway in 1991. The conclusion was that, despite considerable tax increases, the impact was modest, amounting to just a 2% reduction. The majority of the empirical literature looking at the impact of taxes on behavior is based on simple linear reduced form regressions and it is questionable to what extent they can capture the relevant and complex linkages. However, an alternative that has also been used to analyze individual responses to taxes is to make use of an applied general equilibrium model. These models effectively mimic an economy in equilibrium. Individuals and firms are divided according to differing characteristics. For example, a model of the Dutch economy has focused on wage formation, labor supply and demand, and the process of job matching between vacancies and the unemployed. By including elements of wage bargaining and costly job matching, the model describes equilibrium unemployment in terms of the structure of the tax and social security system. The conclusion is that a more progressive tax system, by narrowing the gap between high- and low-paid workers, reduces labor supply. The bulk of this literature, as does implicitly that on AETRs discussed earlier, tends to assume that the agent on whom the tax falls effectively pays the tax. There is, however, a literature on tax incidence, which explores who effectively pays the tax. For example, when a tax is levied on a specific product, e.g., cigarettes, to what extent can the firm pass this price increase on to the consumer and hence to what extent does the firm pay the tax and to what extent does the consumer pay? The analysis also extends to analyzing the impact of hypothetical tax changes. For example, one study analyzed the incidence of a basic income/flat tax proposal using a simple theoretical general equilibrium model.

Conclusion

Different aspects of measurement problems may be assessed with respect to taxation. The solutions vary from the making of simplifying assumptions with available but incomplete data, to the painstaking collection of new data, to the use of sophisticated econometric techniques to estimate unobservable variables from observable behavior. Regression analysis and applied general

equilibrium analysis are also important in measuring individuals' responses to a change in the policy environment. With respect to overall tax effort, the problem is one of deciding on whether the simple tax ratio is a satisfactory measure of tax effort, and if not, then transforming it in such a way that it becomes more satisfactory. With respect to individual tax rates, the problem lies with the fact that the available data do not sufficiently match the theoretical counterparts. In such cases, simplifying assumptions need to be taken to allow us to derive suitable measures. With respect to compliance costs, the data can be collected only by painstaking sample surveys and then, of course, may quickly become dated. Finally, with respect to tax evasion, there is a need to deduce the level of tax evasion from behavior in legitimate areas of activity and by asking hypothetical questions.

Data can be used to validate the predictions of economic theory, to highlight economic problems, and to inform policymakers as they attempt to deal with these problems. Thus, there has been an increasing degree of harmonization with the EU, a harmonization with which the United Kingdom is slightly out of step. The measurement of compliance costs has led to an increasing awareness by governments of this problem and the introduction of measures to deal with it. Finally, the work on tax evasion has highlighted a considerable problem for governments and has provided the data allowing empirical analysis to suggest the effectiveness of possible solutions, such as higher penalties. However, in all these cases, assumptions are required, and hence there is the probability that the measures will in some way be biased as estimates of the real state of affairs. This then raises the question as to whether such data are still useful. The probable answer, depending on the error variance, is yes; some data are better than none as a background for making decisions, provided the potential inaccuracies are borne in mind by the decision maker, other commentators, and users of the data.

See Also the Following Articles

Economics, Strategies in Social Science • Social Economics

Further Reading

Andreoni, J., Erard, B., and Feinstein, J. (1998). Tax compliance. *J. Econ. Lit.* **36,** 818–860.

Atkinson, A. B. (1996). *Public Economics in Action. The Basic Income/Flat Tax Proposal. The Lindahl Lectures.* Clarendon Press, Oxford.

Bhattacharya, D. K. (1990). An econometric method of estimating the hidden economy. *Econ. J.* **100,** 703–717.

Bovenberg, A. L., Graafland, J. J., and de Mooij, R. A. (2000). Tax reform and the Dutch labor market, an applied general equilibrium approach. *J. Public Econ.* **78,** 193–214.

Bruvoll, A., and Larsen, B. M. (2004). Greenhouse gas emissions in Norway. Do carbon taxes work? *Energy Policy* **32,** 493–505.

Carey, D., and Tchilinguirian, H. (2003). *Average Effective Tax Rates on Capital, Labour and Consumption.* OECD Economics Department Working Papers, No. 258. Available on the Internet at www.oecd.org

Chelliah, R. J. (1971). Trends in taxation in developing countries. *IMF Staff Pap.* **18,** 254–325.

Creedy, J. (2001). *Taxation and Economic Behaviour. Introductory Surveys in Economics.* Volume **I,** Edward Elgar, Cheltenham.

Frey, B. S., and Weck-Hannemann, H. (1984). The hidden economy as an unobserved variable. *Eur. Econ. Rev.* **26,** 33–53.

Hudson, J., and Godwin, M. (2000). The compliance costs of collecting direct tax in the UK: An analysis of PAYE. *J. Public Econ.* **77,** 29–44.

Mendoza, E. G., Razin, A., and Tesar, L. L. (1994). *Marginal Tax Rates on the Use of Labour and Capital in OECD Countries*. NBER Working Paper, No. 4864. Available on the Internet at www.papers.nber.org

Orviska, M., and Hudson, J. (2003). Tax evasion, civic duty and the law abiding citizen. *Eur. J. Politic. Econ.* **19,** 83–102.

Sandford, C. (ed.) (1995). *Tax Compliance Costs: Measurement and Policy*. Fiscal Publ., Bath.

Schneider, F., and Enste, D. H. (2000). Shadow economies: size, causes and consequences. *J. Econ. Lit.* **38,** 77–114.

Slemrod, J. B., and Blumenthal, M. (1996). The income tax compliance costs of big business. *Public Finan. Q.* **24,** 411–438.

Telephone Surveys

Don A. Dillman
Washington State University, Pullman, Washington, USA

Glossary

aural communication Transmission of information through the sense of hearing.
coverage error Survey error that results from all members or units of a population not having a known, nonzero, chance of being sampled for data collection.
measurement error Inaccuracies in answers to survey questions that result from poor question wording, inadequate interviewing, and/or the respondent behavior.
nonresponse error Survey error that results from respondents to a survey being different than nonrespondents on characteristics relevant to the survey objectives.
random digit dialing A way of generating samples of households that depends on calling randomly selected telephone numbers.
sampling error Survey error that results from only a subset of the entire population or sample frame being selected for participation in the survey.

Telephone surveys are a means of collecting information from individuals, households, or other units of interest whereby potential respondents answer questions asked by an interviewer over the telephone. Perhaps its most common application occurs through the calling of randomly selected samples of households with telephones where the interest of the sponsor is in being able to determine through careful statistical inference from a relatively small number (hundreds or a few thousand) of interviews the occurrence of opinions or attributes in a large, carefully defined population from which the sample was drawn. For example, during election campaigns many organizations call national samples of households to interview likely voters for purposes of identifying voter preferences in order to predict who will win an upcoming election. Thousands of such telephone surveys, or polls as they are typically described, occur prior to major elections in the United States to ascertain voter opinions on candidates for national, state, and local races. Telephone surveys are also used regularly for marketing research. In addition, important national government-sponsored surveys on such topics as employment rates, health status, and educational behavior are done for use in formulating public policy.

Introduction

Origins

Prior to the 1970s, most important surveys were conducted by face-to-face interviews. However, as summarized by Nathan, the conduct of occasional surveys by telephone was reported in the literature prior to that time by marketing researchers, agricultural economists, and health researchers.

In the early 1970s, many organizations began testing use of the telephone for sample survey data collection. In rapid succession, three books were published that described the promise of telephone surveying for conducting a wide array of such surveys. The first of these, by Blankenship, described the use of telephone survey methods for doing marketing research. The second book, by Dillman, provided step-by-step instructions for designing and conducting telephone surveys. The third, by Groves and Kahn, reported a detailed comparison of the telephone and face-to-face interviews for conducting national surveys. Together, they showed that valid information could be collected over the telephone.

In the next few years an enormous amount of research was undertaken as researchers in many countries sought to understand the multiple issues—from sampling to

overcoming refusals—involved in conducting quality surveys by telephone. In 1988, the first comprehensive book on telephone survey methodology was published, the result of a national conference devoted solely to this mode of data collection. In only a decade, the telephone survey had moved from being an idea with promise to becoming the likely replacement for most face-to-face surveys.

Medium-sized survey organizations in the United States that had relied on face-to-face surveys rapidly shifted to use of the telephone, thus enhancing their ability to do larger surveys of more geographically dispersed populations. Small locality-oriented organizations that had relied mostly on mail survey methods also gained the ability to conduct regional and national surveys of much greater significance.

Reasons for the Growth of Telephone Surveying

Random Digit Dialing and Related Developments

The rapid expansion of telephone surveying in the 1970s and 1980s was the result of numerous developments that helped reduce survey errors and improve efficiency. For example, the evolution of telephone numbers to a standard 10-digit structure (area code plus 3-digit exchange plus final 4 digits) allowed the development of random sampling methods so that all telephone numbers could be assigned a known chance of being selected in samples. In addition, the fact that approximately 90–95% of households in the United States and other developed countries had telephones meant that most households had a chance of being selected for inclusion in survey samples.

The use of random digit dialing on a national basis was greatly facilitated when American Telephone and Telegraph files that showed all area code/prefix combinations became available to survey organizations. These developments, combined with the marketing innovation of WATTS (Wide Area Telephone Transmission Service) (i.e., long-distance calling service to interested customers at wholesale rates), provided the key ingredients that encouraged the creation of centralized telephone survey laboratories.

Minimal Measurement Differences

Research also revealed only minor differences between the results of telephone and face-to-face interview surveys. Thus, the fear that unseen interviewers asking questions of respondents over the telephone would adversely affect measurement began to dissipate. Also, the fact that response rates were relatively high, often approximately the same as those that could be obtained using face-to-face interviews, but without costly callback efforts reduced concern about nonresponse as a source of error.

Cost and Timeliness Concerns

Increased use of the telephone for surveying was also fueled by cost concerns. Face-to-face interview costs had risen rapidly as the number of callbacks required to obtain high response rates increased significantly. These higher costs resulted in part from the fact that in more households both adult males and females were employed. A related problem was the greater difficulty in hiring interviewers as more women, the typical interviewers, moved from part-time interview work to full-time employment.

Telephone interviewers could make repeated callbacks in much less time and without travel costs. The development of centralized telephone laboratories, with close supervision, allowed interviewers to become productive with less training investment and widened the pool for interviewer selection because the willingness to drive long distances was not a requirement. A corresponding decline in long-distance rates provided increasing cost advantages to surveying by telephone.

Surveys could also be done far more quickly by telephone than in person. The combination of time and cost savings made use of the telephone especially attractive to market researchers and pollsters who often needed quick assessments of people's behaviors and attitudes.

Methodological Dominance

By the early 1990s, telephone surveying had become the dominant survey methodology for marketing researchers and pollsters. Even surveys that could not tolerate omission of households without telephone numbers (e.g., the Current Population Survey, which establishes the United States unemployment rate each month) had come to rely on telephones for reinterviews of respondents in those households that had telephones.

How Telephone Sample Surveys Work

In order to conduct valid telephone surveys that allow one to generalize results from a few hundred or thousand completed interviews to an entire population of thousands (e.g., the student body of a university) or millions (e.g., households in a state or country), four sources of error—coverage, sampling, measurement, and nonresponse—described by Groves must be overcome. A well-done telephone survey must take into account all these sources of error when making efforts to generalize results to the total population from which the sample was drawn.

Coverage Error

This type of error is minimized by giving every survey unit in a population a known nonzero opportunity of being selected for the survey. For some survey populations (e.g., members of a professional organization or previous participants in a face-to-face survey), lists of everyone's telephone number may exist, so coverage is not a problem. However, for unlisted populations such as the general public, achieving acceptable coverage may be difficult.

Telephone directories are inadequate sample sources because many telephone subscribers, approximately 30–35% nationally, opt for their number not to be listed in directories. Random digit dialing, which in theory allows all telephone numbers to have a known chance of being selected, provided a means of overcoming this coverage problem. Because in some telephone exchanges many potential numbers (the last four digits) are not used, it was important that more efficient methods be developed for calling working numbers.

One procedure for allowing fewer nonworking numbers to be called was based on an unpublished memorandum by Mitofsky that was elaborated by Waksberg and became known as the Mitofsky–Waksberg method. Their procedure was based on the tendency for telephone companies to assign series of numbers for the last four digits rather than doing so randomly within each exchange. By generating the last two digits of telephone numbers randomly, and then attempting to call other combinations of numbers in this series of 100, only in the event that the first number contacted was a residential phone, an equal probability sample of all residential telephone households, could be obtained. Nathan has summarized numerous efforts to improve the efficiency of telephone sampling methods through the use of lists as well as efforts to improve coverage through combining telephone interviewing with the collection of a portion of the data by means of face-to-face interviews.

In household surveys in which the usual interest is making estimates for the adult population of a city or country, random respondent selection within each household is required so that the interviewees represent the entire U.S. adult population. Several procedures have been developed for accomplishing that objective, including a procedure proposed by Kish for face-to-face interviewing that requires listing all household members by age and gender and following a strict protocol for selecting the interviewee. A simplified method, developed by Trodahl and Carter in 1964, asked only the number of people of each gender living in a household. In 1983, Salmon and Nichols proposed selecting individuals on the basis of which adult has had the last (or next) birthday. Tests have shown that the results from the use of these methods are not always equivalent.

Sampling Error

A second source of survey error that must be reduced is sampling. It occurs because only some members of the entire population are selected for interviews. In general, reduction of sampling error is achieved by increasing the number of randomly selected population members who are surveyed. In simple random samples, approximately 100 completed interviews provide results with a precision of $\pm 10\%$, whereas samples of approximately 400 give a precision of $\pm 5\%$. Samples of approximately 1150 are capable of providing a precision of $\pm 3\%$. The latter is the approximate sample size often used for national election polls. The actual precision of results, from a sampling error perspective, depends on the desired degree of confidence in the sample estimates, how the sample was drawn, and related issues described by Lohr.

One of the achievements associated with telephone interviewing was to make possible simpler sample designs and, therefore, smaller samples to obtain the same levels of precision for the resulting sample estimates as those achievable in comparable national interviews done by face-to-face interviews. The latter method required multistage cluster sampling in which areas, segments, and smaller units within them were typically selected before several households were chosen to interview. This clustering was done to keep interviewing costs at acceptable levels. In national telephone surveys, there is generally little or no savings associated with interviewing respondents by telephone from households in the same neighborhood, as there is for in-person interviews.

Measurement Error

Measurement error may result from poorly worded questions that bias respondent answers, poor interviewing practices, and/or the withholding of accurate information by the person being interviewed. Such errors may occur when some interviewers do not read questions accurately or decide to read them out of order, with the result that different answers are given than would have otherwise been the case. Questions may be answered differently because of the order in which they are asked or even the order in which response categories are presented. In addition, sensitive or threatening questions may not be candidly answered.

Use of the telephone also introduced new perspectives in measurement. The complete dependence on aural communication introduced a trend towards question simplification. Because it was necessary for respondents to remember all aspects of a question (show cards were typically used as a visual aid in personal interviews), survey measurement evolved toward the use of fewer categories and fewer word labels (e.g., the use of polar-point

labeled scales, such as "where 1 means strongly agree and 5 means strongly disagree and you can also use any number in between").

In general, telephone interviews have been found to produce answers that are quite similar to those produced by face-to-face interviews. However, compared to self-administered surveys, telephone respondents are less likely to respond accurately to sensitive questions such as those about sexual or drinking behavior. This tendency to provide socially desirable answers (i.e., answers consistent with cultural norms) is only one of the ways in which telephone answers have been found to differ from responses to self-administered surveys. Others include a recency bias (i.e., choosing the last answers provided on a list) and a tendency to choose more extreme categories. These latter concerns may be a result of cognitive processing that stems from the complete dependence on aural communication and the way in which verbal information is retained and recalled by respondents.

Nonresponse Error

It is rare that all survey units sampled for inclusion in a survey are successfully interviewed. If those who do respond to the survey differ from those who do not respond in a way relevant to the survey (e.g., they have different opinions on a topic asked about in the survey), then nonresponse error is said to occur. Nonresponse error is therefore different than response rate (i.e., the percentage of sampled units that complete an interview). However, the mathematical potential that nonresponse error exists decreases as response rates increase.

Response rates to telephone surveys appear to have declined significantly during the 1980s and 1990s. Nathan described this phenomenon as partly the result of ambiguity associated with the increased technological use of household phones, such as fax machines, computers, and screening technologies (e.g., caller identification and answering machines) in people's homes. It also appears to result from frustration associated with receiving many unsolicited and unwanted marketing calls.

Many techniques have been used to improve response rates, including increasing the length of the calling period and the number of call attempts for each number, refusal conversion, token financial incentives sent with a prior letter informing the recipient of an impending call, and leaving messages on answering machines in anticipation of later call attempts.

Another response improvement strategy that has generated much interest in recent years is training interviewers to provide tailored responses to individuals' objections to being surveyed, a strategy detailed by Groves and Couper based on their analyses of personal interview data. In this strategy, when a respondent indicates that he or she does not have time, the interviewer would be expected to respond to this concern by indicating that was okay and he or she would call back at a more convenient time.

Although it is known that respondents to telephone surveys often differ significantly from nonrespondents, it is not clear whether special efforts to improve response rates will lower nonresponse error. A study by Keeter *et al*. did not show significant changes in respondent answers when they increased response rates from 36.0 to 60.6% through the combined use of advance letters and incentives for listed numbers, repeated callbacks, and refusal conversions over an extended time. This is a topic on which substantial research is now being done.

Technology and Its Consequences

Computer-Assisted Telephone Interviewing

Although early telephone interviews were conducted using regular paper questionnaires, this situation changed. Software was developed for mainframe computers that allowed questions to be displayed on computer screens. Early attempts to do this in the 1970s were plagued by limited question structure capabilities (e.g., not allowing answers to early questions to be incorporated into later questions), long lapses between the time an interviewer clicked an answer and the computer responded, and unpredictable computer downtime when no interviews could be conducted. However, this situation changed significantly during the 1980s as computer-assisted telephone interviewing (CATI) shifted from mainframe to dedicated minicomputers and later to networked personal computers. In addition, software was developed that automated important functions, such as creating random samples of numbers to be called, assigning new numbers as well as those to be called back to interview, and the rapid compilation of results. During this time of rapid development in information technologies, more of its advantages accrued to telephone interviewing than to any other form of survey data collection. A primary impact of these developments was that it became possible to conduct telephone surveys faster. Overnight surveying became popular on many topics so that in some cases less than a day elapsed between the time a survey was designed and the time the results were reported.

Questionnaire designers were able to take advantage of CATI developments to resolve potential measurement problems in surveys. For example, answer categories could be systematically rotated in order to eliminate potential effects of presentation order on respondent answers. It also became possible to automatically insert prior answers

into subsequent questions and branch respondents automatically to the next appropriate question.

Technological Developments That Inhibit the Conduct of Telephone Surveys

Technological developments have also influenced the conduct of telephone surveys in negative ways with respect to the potential for error. Today, most U.S. households have answering machines. Calls are monitored in some households and not answered until recipients of the calls know who is calling. In addition, caller-ID devices are used in many households with the result that calls from unknown numbers, or from those who choose to block display of their numbers, go unanswered. Call-blocking technology also makes it possible to prevent calls from ringing in from numbers that the owner chooses to avoid. In addition, technology allows telephone calls to be automatically forwarded to other numbers that may or may not be located in the same geographic or survey area. The effect of these technologies on nonresponse is to make it more difficult to reach some household telephones than in the past, but whether this results in dramatically lower response rates remains to be determined.

In addition, telephones in homes are connected to other household devices, such as computers, faxes, and security systems. Some telephones serve multiple purposes. This situation increases the likelihood of calling telephone numbers that are not answered by the household occupants and contributes to lower survey response rates.

However, the most profound technological change affecting telephone surveys is the increased use of cellular telephones and the tendency for individual household members to have their own phones. In some cases, households have eliminated traditional land-line phones and use cellular phones only, thus reducing the coverage of traditional phone numbers as a sample frame. There are formidable obstacles to including cellular telephone numbers in random digit sample frames. Some individuals pay by the minute for inbound calls and therefore pay for any calls made to them by others. Risks also exist with respect to calling individuals while in the midst of other activities such as driving a vehicle on a busy street.

Significant differences exist among countries with regard to ownership of cellular telephones and norms that govern their use. In some countries, cellular telephones are the dominant means of telephone communication, leading to substantial decreases in use of land-lines. In other countries, particularly those with poor land-line telephone infrastructures, cellular telephone technology has provided a means of offering telephone service where none previously existed. Cellular telephone numbers appear in directories in some countries but not in others. In addition, interrupting people in midactivity to conduct a survey does not seem to be a problem in some countries, but in others it is viewed as unacceptable. It remains to be seen how these concerns will be resolved.

Interactive Voice Response Surveys

A recent development in telephone surveying is the use of touchtone data entry, whereby respondents listen to prerecorded questions and instructions and use the touchtone numbers on telephones to enter their answers. In its more advanced form, interactive voice response, respondents can state their answers verbally. This technology is used for specialized surveys of employees and customers. It has also been used as a means of cutting costs for regular telephone surveys by having interviewers introduce the survey and then transfer respondents to the automated system.

The Future

It is striking that in a slightly more than 30 years, telephone surveys have moved from being viewed with skepticism as a potential survey methodology to becoming the dominant survey mode in the United States. Now, they are once again being viewed with skepticism.

The technological developments that made its extensive use as a survey methodology possible have been paralleled by other developments that threaten its use. However, it would be a mistake to think of these threats to coverage and response as only technological in nature. Advanced societies have undergone an enormous cultural change in how telephones are owned and used. When surveying began in earnest in the 1970s, telephones were limited to voice communication. Most households had only one telephone and answering machines did not exist, so a ringing telephone demanded to be answered. Telephones are now used for multiple functions in addition to voice conversations and have become a personally owned device with an automatic answering machine. In only three decades, the telephone has changed from controlling people to being controlled by them.

Nonetheless, it seems likely that the telephone will continue to be used for many surveys. Alternative survey methodologies (e.g., mail and the rapidly growing Internet) also suffer from significant coverage problems for household and many other types of surveys. It seems likely that future surveying will emphasize different modes—face-to-face, telephone, mail, or the Internet—for different situations. In addition, it seems likely that more surveys will use multiple modes in an effort to maintain response rates and overcome coverage problems associated with individual modes. Expecting an end to telephone surveying is as unthinkable as expecting

a continuation of telephone surveying only as it has been done in the past.

See Also the Following Articles

Non-Response Bias • Randomization

Further Reading

Blankenship, A. B. (1977). *Professional Telephone Surveys.* McGraw-Hill, New York.

de Leeuw, E., and Van der Zowen, H. (1988). Data quality in telephone and face-to-face surveys: A comparative analysis. In *Telephone Survey Methodology* (R. M. Groves, P. Biemer, L. Lyberg, J. T. Massey, W. L. Nicholls, II, and J. Waksberg, eds.), pp. 283–299. Wiley-Interscience, New York.

Dillman, D. A. (1978). *Mail and Telephone Surveys: The Total Design Method.* Wiley-Interscience, New York.

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method.* Wiley-Interscience, New York.

Dillman, D. A. (2002). Navigating the rapids of change: Some observations on survey methodology in the early 21st century. *Public Opinion Q.* **66**(3), 473–494.

Groves, R. M. (1987). Research on survey data quality. *Public Opinion Q.* **51,** 5156–5172.

Groves, R. M., Biemer, P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., II, and Waksberg, J. (eds.) (1988). *Telephone Survey Methodology*. Wiley-Interscience, New York.

Groves, R. M., and Couper, M. P. (1998). *Nonresponse in Household Interview Surveys.* Wiley-Interscience, New York.

Groves, R. M., and Kahn, R. L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews.* Academic Press, New York.

Groves, R. M., and Nicholls, W. L., II (1986). The status of computer-assisted telephone interviewing: Part II—Data quality issues. *J. Official Stat.* **2**(2), 117–134.

Keeter, S., Miller, C., Kohut, A., Groves, R. M., and Presser, S. (2000). Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Q.* **64,** 125–148.

Kish, L. (1949). A procedure for objective respondent selection within households. *J. Am. Stat. Assoc.* **44,** 380–387.

Lohr, S. (1999). *Sampling Design and Analysis.* Dunsbury, Pacific Grove, CA.

Nathan, G. (2001). Telesurvey methodologies for household surveys—A review and some thoughts for the future. *Survey Methodol.* **27,** 7–31.

Nicholls, W. L., II, and Groves, R. M. (1986). The status of computer-assisted telephone interviewing. Part I—Introductions and impact on costs and timeliness of survey data. *J. Official Stat.* **2**(2), 93–115.

Salmon, C. T., and Nichols, J. S. (1983). The next birthday method of respondent selection. *Public Opinion Q.* **47,** 270–276.

Tarnai, J., and Dillman, D. A. (1992). Questionnaire context as a source of response differences in mail vs. telephone surveys. In *Context Effects in Social and Psychological Research* (N. Schwarz and S. Sudman, eds.), pp. 115–129. Springer-Verlag, New York.

Trodahl, V., and Carter, R. E. (1964). Random selection of respondents within households in phone surveys. *J. Marketing Res.* **1,** 71–76.

Waksberg, J. (1978). Sampling methods for random digit dialing. *J. Am. Stat. Assoc.* **73,** 40–46.

Terman, Lewis

Henry L. Minton
University of Windsor, Windsor, Ontario, Canada

Glossary

achievement Knowledge acquired through learning.
gender identity The extent to which a person chooses to internalize the cultural norms of the masculine or feminine gender.
intelligence The ability to learn or understand from experience.
intelligence quotient (IQ) A quantified measure of tested intelligence, calculated on the Binet scales by dividing mental age by chronological age.
marital adjustment The degree to which married couples are satisfied with their relationship.
mental age On the Binet intelligence scales, the average score of the chronological ages of the standardization group.
mental test Any measure that assesses an individual's ability to learn.
nature–nurture debate Controversy over the extent to which intelligence is influenced by genetic or environmental factors.

Lewis M. Terman (1877–1956) played a central role in the development and establishment of psychological tests that assessed individual and group differences. He was committed to applying the technology of psychological measurement to fit the needs of the dynamically changing American society of the early 20th century. To meet the demands of a growing industrialized and urbanized nation, he sought to demonstrate that measures of intellectual and personality differences could be used to sort individuals into the social roles they were most qualified to fulfill. According to Terman, those individuals who excelled intellectually and motivationally had the potential to achieve the highest positions of responsibility and leadership. As a result of such a meritocratic structure, both individual and social efficiency would be maximized. In this article, Terman's accomplishments in pioneering psychological testing and achieving his social objectives are considered by reviewing and evaluating his career.

Early Life and Professional Training

Terman was born and raised on a farm in central Indiana, the 12th of 14 children. He attended a one-room school, completing the eighth grade when he was 12 years old. Determined to further his education and with the financial help of his parents, Terman left the family farm at age 15 to attend Central Normal College in Danville, Indiana. During a 6-year period, he earned three undergraduate degrees at Normal College. At age 17, with basic teacher preparation achieved, he obtained his first teaching position, and 2 years later he became a high school principal. While at Normal College, he met fellow student Anna Belle Minton (no relation to the author), whom he married in 1899.

With aspirations beyond school teaching, Terman enrolled at Indiana University in 1901, earning a master's degree in psychology in 2 years. With the encouragement of his Indiana mentor, Ernest H. Lindley, he went on to Clark University in 1903 for doctoral studies with G. Stanley Hall, one of the early leaders in American psychology. For his dissertation, Terman undertook an experimental investigation of mental tests in which he compared a "bright" and a "dull" group of 10- to 13-year-old boys. Since Hall did not approve of mental tests, Edmund C. Sanford became his dissertation adviser, and he received his Ph.D. in 1905. Hall, however, influenced Terman's thinking about the nature of intelligence. Consistent with Hall's evolutionary perspective

on individual and group differences, Terman believed that mental tests measured native ability.

During his tenure at Clark, Terman contracted tuberculosis. Although he made a successful recovery, he decided that when he completed his studies it would be desirable to work in a warm climate. He therefore accepted a position as a high school principal in San Bernardino, California. A year later, he was able to obtain a more intellectually stimulating assignment, teaching child study and pedagogy at the Los Angeles State Normal School. In 1910, his academic career was enhanced with an appointment at Stanford University's education department. He spent the remainder of his career at Stanford, becoming the head of the psychology department in 1922, a position he held until his retirement in 1942.

Pioneering the Measurement of Intelligence

The move to Stanford in 1910 coincided with Terman's physical ability to take on a more active academic workload. He thus resumed his research interests in mental testing and began to work with Alfred Binet's 1908 scale, the first widely accepted measure of intelligence. Henry H. Goddard had published translations of Binet's original 1905 scale and the subsequent 1908 revision. Terman's earliest revision of the Binet appeared in 1912, and with the assistance of a team of graduate students, the finished product—the Stanford–Binet—was published in 1916. An innovative feature of the Stanford–Binet was the inclusion of the "intelligence quotient" (IQ), a concept advanced by William Stern but not previously used in mental tests. In competition with a number of other American versions of the Binet, Terman's Stanford revision made use of the largest standardization sample and by the 1920s became the most widely used individually administered intelligence test.

As a result of the published Stanford–Binet, Terman became a highly visible figure in the mental testing movement. Attesting to his reputation, in 1917 he was invited to become a member of a committee that had been assembled at the Vineland, New Jersey, training school for the mentally retarded to devise tests for the U.S. Army. The United States had entered World War I, and Robert M. Yerkes, the president of the American Psychological Association, organized the psychologists' contribution to the war effort. The test committee, chaired by Yerkes, was composed of the leading psychologists in the mental testing field. Terman brought with him a new group-administered version of the Stanford–Binet that had been constructed by his doctoral student, Arthur S. Otis. This test served as the basis for the development of the army group tests (the Alpha and Beta examinations). Although serious questions have been raised about the significance of the psychologists' contributions to the war, there is no doubt that the war provided an enormous boost for the mental testing movement. Approximately 1.75 million men were tested, and on this basis recommendations were advanced with respect to job placements or immediate discharge from the army. The major weakness of the army testing program was the psychologists' failure to consider the impact of cultural differences on tested intelligence. Thus, the lower IQ scores found for foreign-born and poor native-born soldiers were attributed to low levels of native ability rather than such alternatives as limited acculturation and schooling. Terman, like the other members of the army testing committee, subscribed to the Galtonian theory that mental abilities were primarily a product of heredity.

Applying Psychological Testing to Education

After the war, Terman advanced the use of the army group testing methods for education. To this end, he and the other psychologists who constructed the army tests adapted them for school-age children. The resulting National Intelligence Tests for grades 3–8 were published in 1920. Terman promoted the use of intelligence tests as a means of reorganizing schools so that pupils could be categorized into homogeneous ability groups. During the 1920s, intelligence testing and the tracking system of ability grouping became popular practices in schools and Terman played a central role in fostering these policies. He was also a leader in the development of standardized achievement tests, which measured school learning. With a team of Stanford colleagues, he constructed the first achievement test battery—the Stanford Achievement Test. Terman believed that educational testing would be of great value to U.S. society. It would serve as the major means of achieving his vision of a meritocracy within the U.S. democratic ideal—a social order based on ranked levels of native ability. As a measure of native ability, intelligence tests could identify children who were cognitively gifted and therefore had the potential to become the leaders of society. Once these children were identified, it was the responsibility of the schools to devote the necessary time and effort to cultivate their intellectual potential.

Identifying and Enhancing Intellectual Giftedness

To accomplish his meritocratic objectives and with financial support from the Commonwealth Fund of New York,

Terman launched a longitudinal study of gifted children in 1921. This was the first follow-up study to use a large sample. Children with an IQ of at least 135 were categorized as gifted. Terman and his research team generated a sample of approximately 1500 gifted children, based on canvassing elementary and secondary schools in urban areas of California. In an effort to dispel the popular notion that gifted children were underdeveloped in nonintellectual areas, medical and physical assessments were included as well as measures of personality, character, and interests. The gifted sample was compared with a control group of California schoolchildren of comparable age.

In the first of a series of monographs on the gifted study, the major finding was that gifted children excelled in measures of academic achievement when matched for age with control children. The composite profiles of the gifted children also revealed that they were emotionally as well as intellectually mature. Based on these initial results, Terman strongly promoted a differentiated school curriculum that would place gifted children in special classrooms in which they could progress educationally according to their ability rather than their age. With additional research funding, Terman was able to follow up his sample for a period of 35 years. At midlife, the intellectual level of the gifted group continued to be within the upper 1% of the general population, and their vocational achievement was well above the average of college graduates. Moreover, as earlier reports had demonstrated, they showed few signs of such serious problems as insanity, delinquency, or alcoholism. The midlife report also included some marked gender differences. Whereas the men as a group had attained a high level of career success, few women had comparable levels of career achievement. As Terman observed in the 1959 monograph on the gifted sample at midlife, career opportunities for women were restricted by gender role conformity and job discrimination.

Terman's involvement with the gifted study entailed more than data collection and research reports. Especially after he retired in 1942, he devoted himself to the interests of gifted children by promoting gifted education and, through contacts with journalists, disseminated the results of the gifted study in newspapers and magazines. He also popularized his work by making a guest appearance on the radio show "The Quiz Kids." His appearance in 1947 coincided with the publication of the 25-year follow-up. These forays into the popular media also served as a vehicle for Terman to change the public's negative stereotypes of gifted children as maladjusted. In his work with the gifted, Terman experienced particular satisfaction in his personal contact with the participants under study. He maintained correspondence with many of them over the years and in some instances received them as guests in his home. For a number of the gifted who "grew up" and came to be identified as "Termites," he was a benevolent father figure and psychological counselor. By the early 1950s, with plans developing for the continuation of the gifted follow-up, Terman appointed Stanford colleague Robert Sears (who coincidentally was a member of the gifted sample) to succeed him as research director. The gifted sample was thus followed up through late adulthood.

Debating the Testing Critics

As one of the leading advocates of intelligence testing, critics of the testing movement often challenged Terman. These challenges began in the early 1920s when the results of the army testing became widely disseminated. The influential journalist Walter Lippmann wrote a series of highly critical articles about the army tests in the *New Republic*. Lippmann singled out Terman because of his development of the Stanford–Binet and asserted that there was no scientific basis for the claim made by Terman and the other army psychologists that the tests measured native ability. Terman responded in the *New Republic* by dismissing Lippmann as a nonexpert in testing who should thus stay out of issues that he was uninformed about. Lippmann, in fact, was quite technically sophisticated in many of his criticisms, but Terman chose to be evasive in his response to the points that Lippmann raised, such as an environmental interpretation of the correlation between tested intelligence and social class.

During the 1920s, Terman also engaged in a series of published debates about testing with psychologist William C. Bagley, another critic of the hereditarian view of intelligence. In an attempt to settle issues, Terman took on the task of chairing a committee that organized an edited book on the nature–nurture debate. In this monograph, published in 1928, leading advocates of each position marshaled evidence and arguments, but as in previous exchanges, nothing was resolved. In 1940, Terman was once again drawn into the nature–nurture debate, this time challenged by a team of environmentalist advocates at the University of Iowa led by George D. Stoddard. Stoddard campaigned for the limited use of intelligence tests because they were subject to environmental influences that compromised their usefulness in making long-term predictions. Terman was concerned that Stoddard's position against mass testing would threaten his career objective of establishing a meritocracy based on IQ differences. As in previous instances, the 1940 debate led to an impasse. No changes took place in the widespread use of intelligence tests in the schools. It would not be until the 1960s, as a consequence of the civil rights movement, that mass testing was seriously challenged. Terman did modify his position to some extent. In the 1930s, mindful of the racial propaganda of Nazi Germany, he resigned his long-standing membership in

the American Eugenics Society. After World War II, although he still held to his democratic ideal of a meritocracy, he no longer endorsed a hereditarian explanation of race differences, and he acknowledged that among the gifted, home environment was associated with degree of success.

Measuring Gender Identity and Marital Adjustment

Terman's interest in the measurement of individual and group differences extended beyond mental abilities and achievement. Deriving from his study of the gifted, he became interested in assessing nonintellectual traits. By measuring emotional and motivational characteristics, he hoped to demonstrate that the gifted had well-adjusted and well-rounded personalities. To tap this facet of human differences, he set out to measure gender identity, which was viewed as a composite of motivational and emotional traits that differentiated the sexes. He identified masculine and feminine interests from questionnaire preferences given by gifted boys and girls about their play activities, games, and amusements. The initial survey conducted in 1922 revealed that the gifted children were similar in gender orientation to the control children. In 1925, Terman received a National Research Council grant to investigate sex differences and, with his former student Catharine Cox Miles, constructed a masculinity–femininity (M–F) test, the first measure of its kind. The final version published in 1936, called the Attitude–Interest Analysis Test to disguise its purpose, was based on normative samples of male and female groups ranging in age from early adolescence to late adulthood, although the core of the sample was high school juniors and college sophomores. The test comprised approximately 450 multiple-choice items that assessed preferences for a variety of activities and interests, as well as responses to situations that might arouse feelings of anger or fear.

In an attempt to validate the M–F test, Terman was able to collect test protocols from a group of male homosexuals in San Francisco. As he predicted, the results showed that male homosexuals had high feminine scores. He therefore concluded that marked deviations from gender-appropriate behaviors and norms were psychologically unhealthy because such deviations would very likely lead to homosexuality. Even if this "maladjustment" did not develop, other problems could arise. Referring to those with cross-gender identities, Terman and Miles in their 1936 monograph, "Sex and Personality: Studies in Masculinity and Femininity," commented, "One would like to know whether fewer of them marry, and whether a larger proportion of these marriages are unhappy" (p. 468). Underscoring this point, they observed that "aggressive and independent females" could very well be at a disadvantage in the "marriage market" (p. 452). They also expressed the fear that too much competition between the sexes would not be socially desirable. In essence, the authors supported the conventional patriarchal relationship between the sexes. (The extent to which Catharine Cox Miles concurred with this position is not clear since Terman acknowledged prime responsibility for the conclusions in their book.) Terman's conclusions were based on the standardized norms he generated with his M–F test. What the test reflected were the gender norms of the 1930s, but Terman was insensitive to the cultural and historical limits of his measure. He chose to emphasize the need to raise and educate girls and boys so that they would conform to the existing gender norms that fostered a clear distinction between the sexes. As in the case of his vision of a social order ranked by native ability, Terman believed that sex differences also had to follow a prescribed ranking. Paralleling the need to cultivate ability differences to meet the needs of a changing society, in his view it was also important to ensure compatible sex roles in the face of potential conflict between the sexes. Many social scientists during the interwar era, mindful of the feminist challenge, preached the need for compatibility rather than conflict between the sexes.

Terman's interest in gender identity and sex differences expanded to research on marital adjustment. He conducted a large-scale survey study of several hundred married and divorced couples in the San Francisco area. The major finding, according to Terman, was that contrary to previous research results, sexual compatibility was less influential than personality and background factors in predicting marital happiness. He therefore stressed that the key to marital adjustment was the extent to which each spouse accepted the other's needs and feelings and did not fight to get his or her own way. To emphasize this point, he observed that happily married women could be characterized as being cooperative and accepting of their subordinate roles. Terman's conventional views on gender carried over from his masculinity–femininity study to his marital research.

Evaluating Terman's Contribution to Social Measurement

Terman's seminal contributions to the development of psychological testing and the study of the intellectually gifted ensure his position as one of the pioneers in devising social measurement. Perhaps more than any of the other advocates of the testing movement, he was successful in creating a wide variety of methods assessing individual and group differences. His interest in the gifted led

him far beyond the measurement of ability. As a consequence, he was in the vanguard of constructing indices of school achievement, gender identity, marital adjustment, and sexual behavior. Aside from these personal achievements, Terman has left us with an unfulfilled legacy. What he wanted to accomplish with his psychological tests and identification of the intellectually gifted was a more socially just and democratic society. A considerable part of Terman's project, however, has had an unintended dehumanizing effect. For racial and ethnic minorities and lower class individuals, his differentiated educational system based on IQ scores served as an obstacle for personal development and equal opportunity. His views on gender and homosexuality worked against the creation of a more pluralist society. What Terman failed to understand was the intricate way in which scientific knowledge reflects social power. By uncritically accepting the given power inequities of American society, he produced scientific knowledge and technology that functioned to perpetuate the status quo.

See Also the Following Articles

Binet, Alfred • Intelligence Testing • Psychological Testing, Overview

Further Reading

Chapman, P. D. (1988). *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890–1930.* New York University Press, New York.

Fancher, R. R. (1985). *The Intelligence Men: Makers of the IQ Controversy.* Norton, New York.

Gould, S. J. (1981). *The Mismeasure of Man.* Norton, New York.

Hernnstein, R. J., and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* Free Press, New York.

Jacoby, R., and Glauberman, N. (eds.) (1995). *The Bell Curve Debate: History, Documents, Opinions*. Random House, New York.

Kimmel, M. (1996). *Manhood in America: A Cultural History.* Free Press, New York.

Minton, H. L. (1988). *Lewis M. Terman: Pioneer in Psychological Testing.* New York University Press, New York.

Samelson, F. (1979). Putting psychology on the map: ideology and intelligence testing. In *Psychology in Social Context* (A. R. Buss, ed.), pp. 103–168. Irvington, New York.

Seagoe, M. V. (1975). *Terman and the Gifted.* Kaufmann, Los Altos, CA.

Sokal, M. M. (ed.) (1987). *Psychological Testing and American Society, 1890–1930.* Rutgers University Press, New Brunswick, NJ.

Zenderland, L. (1998). *Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing.* Cambridge University Press, New York.

Test Equating

Michael J. Kolen
University of Iowa, Iowa City, Iowa, USA

Glossary

alternate forms Versions of a test, each of which contains different items; constructed to be similar to one another in content and statistical properties.
equating design Process used to collect data for conducting equating.
equating method Statistical method used to define and estimate the equating relationship between alternate forms.
raw scores Scores on a test prior to transformation; often, the number of test questions correctly answered by an examinee.
scale scores Scores transformed to a common scale that allows for direct comparison of scores earned on different alternate forms.
test content specifications Detailed description of the numbers of test questions from each of a number of content areas; enables development of alternate forms having similar content.
test statistical specifications Detailed description of the statistical properties of test questions; enables development of alternate forms having similar statistical properties.

Test equating methods are statistical methods used to adjust test scores for differences in test difficulty among alternate forms of educational and psychological tests, with the goal being to use scores on the alternate test forms interchangeably.

Introduction

Alternate forms of educational and psychological tests are often developed that contain different sets of test questions. The alternate forms are administered on different occasions, which enhances the security of the tests and allows examinees to be tested more than once. Test content specifications detail the numbers of questions on a test from each of a number of content areas. Test statistical specifications detail the statistical properties (e.g., difficulty) of the test questions. Alternate forms of tests are built to the same content and statistical specifications, which is intended to lead to alternate forms that are very similar in content and statistical properties.

Although alternate forms are built to be similar, they typically differ somewhat in difficulty. Test equating methods are statistical methods used to adjust test scores for the differences in test difficulty among the forms. A requisite condition for applying test equating methodology is that the alternate forms be built to the same content and statistical specifications. Equating methods adjust for small differences in test difficulty among the forms. As emphasized by Kolen and Brennan (1995, p. 3), "equating adjusts for differences in difficulty, not for differences in content." The goal of equating is to enable scores on the alternate test forms to be used interchangeably.

Test equating is used when alternate forms of a test exist and examinees who are administered the different test forms are considered for the same decision. For example, the ACT Assessment (http://www.act.org) is a college entrance examination used in the United States. The ACT Assessment contains four tests in the areas of English usage, mathematics, reading, and science reasoning. All test questions are multiple choice. The test is administered annually in September, October, December, February, April, and June each year, with different test forms administered on each test date. Examinees who are applying for admission to a particular university might have taken the ACT Assessment on any of the test dates during the past year or even 2 years. Examinees who were administered the test on different test dates might be considered together for admission to a university for

the fall semester. In this situation, it is important that the scores on the test forms be interchangeable.

The implementation of test equating requires a process for collecting data, referred to as an equating design. Statistical equating methods are also a component of the equating process. A variety of equating designs and equating methods exists. Some of the more popular ones are considered here.

Test equating has been conducted since the early 20th century. The first comprehensive treatment of equating was presented by Flanagan in 1951. Subsequent treatments by Angoff in 1971, Holland and Rubin in 1982, Petersen *et al*. in 1989, and Kolen and Brennan in 1995 trace many of the developments in the field. In a 1999 publication, the American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing provide standards that are to be met when equating tests in practice.

The Scaling and Equating Process

Raw scores on tests are often computed as the number of test questions that a person answers correctly. Typically raw scores are transformed to scale scores. The use of scale scores facilitates score interpretation. Often, properties of score scales are set with reference to a particular population. For example, the ACT Assessment score scale was set to have a mean of 18 for a nationally representative group of examinees who indicated that they were planning to attend college. Often, the score scale is established using an initial alternate form of a test. Raw scores on a subsequent alternate form are equated to raw scores on the initial form. The raw-to-scale score transformation for the new form is then applied to the equated scores on the subsequent form. Later, raw scores on new forms are equated to previously equated forms and then transformed to scale scores. The scaling and equating process results in scores from all forms being reported on a common scale. The intent of this process is to be able to state, for example, that "a scale score of 26 indicates the same level of proficiency whether it is earned on Form X, Form Y, or Form Z."

Equating Designs

Equating requires that data be collected and analyzed. Various data collection designs are used to conduct equating. Some of the most common designs are discussed here.

Random Groups

In the random groups design, alternate test forms are randomly assigned to examinees within a test center. One way to implement the random groups design is to package the test booklets so that the forms alternate. For example, if two test forms, form X and form Y, are to be included in an equating study, the form X and form Y test booklets would be alternated in the packages. When the forms are distributed to examinees, the first examinee would receive a form X test booklet, the second examinee a form Y booklet, and so on. This assignment process leads to comparable, randomly equivalent groups being administered form X and form Y.

Assuming that the random groups are fairly large, differences between mean scores on form X and form Y can be attributed to differences in difficulty of the two forms. Suppose, for example, following a random groups data collection the mean raw score for form X is 80 and the mean raw score for form Y is 85. These results suggest that form X is 5 raw score points more difficult than form Y. Such a conclusion is justified because the group of examinees taking form X is randomly equivalent to the group of examinees taking form Y.

For example, the ACT Assessment is equated using the random groups design. The score scale was initially constructed in 1988 using an initial form. Subsequently, the initial form and a set of new forms were administered using the random groups design in specially selected test centers on a single test date using the random groups design. The new forms were equated to the initial form and then to the score scale using this design. The new forms were then used in later test dates during the first year. In the following years, a previously equated form and a set of new forms have been administered using the random groups design. In all cases, scores on new forms have been expressed on the common score scale.

Single Group Design

In the single group design, the same examinees are administered two alternate forms. The forms are separately timed. Typically, one of the forms has been equated in the past and one of the forms is to be equated. The order of the test forms is usually counterbalanced. One random half of the examinees are administered form X followed by form Y. The other random half are administered form Y followed by form X. Counterbalancing is used to control for context effects. For example, due to the effects of fatigue, examinees who take a form second might not do as well because they may be tired when taking the second form. Also, due to practice, examinees who take a form second might do better because they have had a chance to practice on the form taken first. Counterbalancing requires the assumption that taking form X

prior to form Y has the same effect as taking form Y prior to form X. If this assumption does not hold, then differential order effects are said to be present, and the data on the form taken second are discarded, resulting in a considerable loss of data.

Common-Item Nonequivalent Groups

In the common-item nonequivalent groups design, form X and form Y are administered to different (nonequivalent) groups of examinees. The two forms have items in common. There are two variants of this design. When using an internal set of common items, the common items contribute to the examinee's score on the form. With an internal set, typically the common items are interspersed with the other items on the test. When using an external set of common items, the common items do not contribute to the examinee's score on the form taken. With an external set, the common items typically appear in a separately timed section that is administered during the administration of the form taken.

When using the common-item nonequivalent groups design, the common items are used to indicate how different the group of examinees administered form X is from the group of examinees administered form Y. Strong statistical assumptions are used to translate the differences between the two groups of examinees on the common items to differences between the two groups on the complete forms.

Because scores on the common items are used to indicate differences between the examinee groups, it is important that the common items fully represent the content of the test forms. Otherwise, a misleading picture of group differences is provided. In addition, it is important that the common items behave in the same manner when they are administered with form X as with form Y. Therefore, the common items should be administered in similar positions in the test booklets in the two forms, and the text of the common items should be identical.

For an example of a common-item nonequivalent groups equating design, consider the SAT I (http://www.collegeboard.com). Like the ACT Assessment, the SAT I is used for college admissions in the United States. The SAT I tests are administered seven times per year. The SAT I tests are equated using the common-item nonequivalent groups design with an external set of common items. The SAT I tests contain objectively scored verbal and mathematics tests. Examinees are administered three mathematics sections, three verbal sections, and one "equating section." All examinees are administered the same operational (not including the equating section) form. The equating section contains an external set of common items, so scores on the common items do not contribute to examinees' scores. Examinees are randomly assigned to receive either a verbal common item section or a mathematics common item section. The examinees have no way of knowing which sections are operational sections and which sections are equating sections, so they are equally motivated on all sections of the test. Some of the common item sections were also administered during previous administration and are used to equate the new form to a form that was previously administered.

Comparison of Equating Designs

The benefits and limitations of the three designs can be compared on five dimensions: ease of test development, ease of administration, security of test administration, strength of statistical assumptions, and sample size requirements. Of the designs considered, the common-item nonequivalent groups design requires the most complex test development process. Common item sections must be developed that mirror the content of the total test so that the score on the common item sections can be used to give an accurate reflection of the difference between the group of examinees administered the old form and the group of examinees administered the new form. Test development is less complex for the random groups and single group designs because there is no need to construct common item sections.

However, the common-item nonequivalent groups design is the easiest of the three designs to administer. Only one test form needs to be administered on each test date. For the random groups design, multiple forms must be administered on a test date. For the single group design, each examinee must take two forms, which cannot be done in a regular test administration.

The common-item nonequivalent design tends to lead to greater test security than the other designs because only one form needs to be administered at a given test date. With the random groups and single group designs, multiple forms are administered at a particular test date to conduct equating. However, security issues can be of concern with the common-item nonequivalent groups design because the common items must be repeatedly administered.

The common-item nonequivalent groups design requires the strongest statistical assumptions. The random groups design requires only weak assumptions, mainly that the random assignment process was successful. The single group design requires stronger assumptions than the random groups design in that it assumes no differential order effects.

The random groups design requires the largest sample sizes of the three designs. Assuming no differential order effects, the single group design has the smallest sample size requirements of the three designs because, effectively, each examinee serves as his or her own control.

As is evident from the preceding discussion, each of the designs has strengths and weaknesses. Choice of design

depends on weighing the strengths and weaknesses with regard to the testing program under consideration. Each of these designs has been used to conduct equating in a variety of testing programs.

Statistical Methods

Equating requires that a relationship between alternate forms be estimated. Equating methods result in a transformation of scores on the alternate forms so that the scores possess specified properties. For traditional equating methods, transformations of scores are found such that for the alternate forms, after equating, the distributions, or central moments of the distributions, are the same in a population of examinees for the forms to be equated. Traditional equating methods focus on observed scores. Item response theory (IRT) methods make heavy use of test theory models.

Traditional Methods

Traditional observed score equating methods define score correspondence on alternate forms by setting certain characteristics of score distributions equal for a specified population of examinees. For example, in traditional equipercentile equating, a transformation is found such that, after equating, scores on alternate forms have the same distribution in a specified population of examinees. Assume that scores on form X are to be equated to the raw score scale of form Y. Define X as the random variable score on form X, Y as the random variable score on form Y, F as the cumulative distribution function of X in the population, and G as the cumulative distribution function of Y in the population. Let e_Y be a function that is used to transform scores on form X to the form Y raw score scale, and let G^* be the cumulative distribution function of e_Y in the same population. The function e_Y is defined to be the equipercentile equating function in the population if

$$G^* = G. \tag{1}$$

Scores on form X can be transformed to the form Y scale using equipercentile equating by taking

$$E_Y(x) = G^{-1}[F(x)], \tag{2}$$

where x is a particular value of X, and G^{-1} is the inverse of the cumulative distribution function G.

Finding equipercentile equivalents would be straightforward if the distributions of scores were continuous. However, test scores typically are discrete (e.g., the number of items correctly answered). To conduct equipercentile equating with discrete scores, the percentile rank of a score on form X is found for a population of examinees. The equipercentile equivalent of this score is defined as the score on form Y that has the same percentile rank in the population. Due to the discreteness of scores, the resulting equated score distributions are only approximately equal.

Because many parameters need to be estimated in equipercentile equating (percentile ranks at each form X and form Y score), equipercentile equating is subject to much sampling error. For this reason, smoothing methods are often used to reduce sampling error. In presmoothing methods, the score distributions are smoothed. In postsmoothing methods, the equipercentile function is smoothed. Kolen and Brennan discuss a variety of smoothing methods.

After raw scores on form X are equated to the form Y scale, typically the scores are transformed to scale scores using the raw to scale score transformation for form Y.

Other traditional methods are sometimes used that can be viewed as special cases of the equipercentile method. In linear equating, a transformation is found that results in scores on form X having the same mean and standard deviation as scores on form Y. Defining $\mu(x)$ as the mean score on form X, $\sigma(x)$ as the standard deviation of form X scores, $\mu(Y)$ as the mean score on form Y, $\sigma(Y)$ as the standard deviation of form Y scores, and l_Y as the linear equating function,

$$l_Y(x) = \sigma(Y)\left[\frac{x - \mu(X)}{\sigma(X)}\right] + \mu(Y). \tag{3}$$

Unless the shapes of the score distributions for form X and form Y are identical, linear and equipercentile methods produce different results. However, even when the shapes of the distributions differ, equipercentile and linear methods produce similar results near the mean. When interest is in scores near the mean, linear equating is often sufficient. However, when interest is in scores all along the score scale and sample size is large, then equipercentile equating is often preferable to linear equating. A common rule of thumb is that a minimum of 1000 examinees per form are needed for equipercentile equating, whereas fewer examinees are needed for linear equating.

For the random groups and single group designs, the sample data typically are viewed as representative of the population of interest, and the estimation of the traditional equating functions proceeds without the need to make strong statistical assumptions. However, estimation in the common-item nonequivalent groups design requires strong statistical assumptions. First, a population must be specified in order to define the equipercentile or linear equating relationship. Since form X is administered to examinees from a different population than is form Y, the population used to define the equating relationship is typically viewed as a combination of these two populations. The combined population is referred to as the synthetic population.

Three common ways to define the synthetic population are to equally weight the population from which examinees are sampled to take form X and form Y, weight the two populations by their respective sample sizes, or define the synthetic population as the population from which examinees are sampled to take form X. The definition of the synthetic population typically has little effect on the final equating results. Still, it is necessary to define a synthetic population in order to proceed with traditional equating using this design.

Kolen and Brennan describe a few different equating methods for the common-item nonequivalent groups design. The methods differ in terms of their statistical assumptions. Define V as score on the common items. In the Tucker linear method, the linear regression of X on V is assumed to be the same for examinees taking form X and examinees taking form Y. A similar assumption is made about the linear regression of Y on V. In the Levine linear observed score method, similar assumptions are made about true scores rather than observed scores. No method exists to directly test these assumptions using data that are collected for equating. Methods do exist for equipercentile equating under this design that make somewhat different regression assumptions.

IRT Methods

Unidimensional IRT models assume that examinee proficiency can be described by a single latent variable, θ, and that items can be described by a set of parameters or curves that relate proficiency to probability of correctly answering the item. For multiple-choice tests, the probability that examinees of proficiency θ correctly answer item g is symbolized $p_g(\theta)$. IRT models are based on strong statistical assumptions. The θ scale has an indeterminate location and spread. For this reason, one θ scale sometimes needs to be converted to another linearly related θ scale. If number-correct scores are to be used, then there are two steps in IRT equating. First, the θ scales for the two forms are considered to be equal or are set equal. Then, number-correct score equivalents on the two forms are found.

In many situations, the parameter estimates for the two forms are on the same θ scale without further transformation. In general, no transformation is needed in the following situations: (i) in the random groups design, (ii) in the single group design, and (iii) in the common-item nonequivalent groups design when form X and form Y parameters are estimated simultaneously. The typical situation in which a transformation of the θ scale is required is in the common-item nonequivalent groups design when the form X and form Y parameters are estimated separately.

After the parameter estimates are on the same scale, IRT true and IRT observed score methods can be used to relate number-correct scores on form X to number-correct scores on form Y. In IRT true score equating, the true score on one form associated with a given θ is considered to be equivalent to the true score on another form associated with that same θ. In IRT the true score on form X for an examinee of ability θ is defined as

$$\tau_X(\theta) = \sum_{g:X} p_g(\theta), \tag{4}$$

where the summation g: X is over items on form X. True score on form Y for an examinee of ability θ is defined as

$$\tau_Y(\theta) = \sum_{g:Y} p_g(\theta), \tag{5}$$

where the summation g: Y is over items on form Y. Typically, an integer score on form X is specified. The θ that leads to an equality in Eq. (4) is found by iterative means. This θ is then substituted into Eq. (5) to find the IRT true score equivalent of the integer score on form X. In practice, estimates of parameters are substituted for the parameters in Eqs. (4) and (5).

IRT observed score equating uses the item parameters estimated for each form along with the estimated distribution of ability for the population of examinees to estimate the number-correct score distribution for form X and form Y. Given the item parameters on a test form, Lord and Wingersky provided a recursive equation that can be used to find the distribution of number-correct scores, conditional on θ, which is symbolized as $f(x \mid \theta)$. Many IRT computer programs output an estimate of the distribution of θ, symbolized as $g(\theta)$. The distribution of number-correct scores on form X, $f(x)$, can be related to these two quantities by the following equation:

$$f(x) = \int_\theta f(x \mid \theta) g(\theta) d\theta. \tag{6}$$

Given the item parameters and distribution of θ, Eq. (6) can be used to estimate a smoothed distribution of number-correct scores in the population for form X. A similar process can be used to obtain a smoothed distribution for form Y. Standard equipercentile equating procedures are then used to equate these two smoothed distributions.

Any application of unidimensional IRT models requires that the test forms be unidimensional and that the relationship between ability and probability of correct response follows the model. In these applications, the fit of the models needs to be carefully analyzed.

Equating Error

Minimizing equating error is a major goal when developing tests that are to be equated, designing equating studies, and conducting equating. Random equating error is

present whenever samples from populations of examinees are used to estimate equating relationships. Random error depends on the design used for data collection, the score point of interest, the method used to estimate equivalents, and sample size. Standard errors of equating are used to index random error. Standard error equations have been developed to estimate standard errors for most common designs and methods, and resampling methods such as the bootstrap can also be used. In general, standard errors diminish as sample size increases. Standard errors of equating can be used to estimate required sample sizes for equating, for comparing the precision of various designs and methods, and for documenting the amount of random error in equating.

Systematic equating error results from violations of assumptions of the particular equating method used. For example, in the common-item nonequivalent groups design, systematic error will result if the Tucker method is applied and the regression-based assumptions that are made are not satisfied. Systematic error typically cannot be quantified in operational equating situations.

Equating error of both types needs to be controlled because it can propagate over equatings and result in scores on later test forms not being comparable to scores on earlier forms. Choosing a large enough sample size given the design is the best way to control random error. To control systematic error, the test must be constructed and the equating implemented so as to minimize systematic error. For example, the assumptions for any of the methods for the common-item nonequivalent groups designs tend to hold better when the groups being administered the old and the new forms do not differ too much from each other. The assumptions also tend to hold better when the forms to be equated are very similar and when the content and statistical characteristics of the common items closely represent the content and statistical characteristics of the total test forms. Another way to help control error is to use what is often referred to as double-linking. In double-linking, a new form is equated to two previously equated forms. The results for the two equatings are often averaged to produce a more stable equating than if only one previously equated form had been used. Double-linking also provides for a built-in check on the adequacy of the equating.

Selected Practical Issues

Due to practical constraints, equating cannot be used in some situations in which its use may be desirable. Use of any of the equating methods requires test security. In the single group and random groups designs, two or more test forms must be administered in a single test administration. If these forms become known to future examinees, then the equating and the entire testing program could be jeopardized. With the common-item nonequivalent groups design, the common items are administered on multiple test dates. If the common items become known to examinees, the equating is also jeopardized. In addition, equating requires that detailed content and statistical test specifications be used to develop the alternate forms. Such specifications are a prerequisite to conducting adequate equating.

Although the focus of this article has been on equating multiple-choice tests that are scored number-correct, equating can often be used with tests that are scored in other ways, such as essay tests scored by human raters. The major problem with equating such tests is that frequently very few essay questions can be administered in a reasonable time frame, which can lead to concerns about the comparability of the content from one test form to another. It also may be difficult, or impossible, when the common-item nonequivalent groups design is used to construct common item sections that represent the content of the complete tests.

Recently, computers have been used to administer tests. Often, adaptive tests are used in which questions are selected based on examinees' responses to previous questions. For example, when an examinee incorrectly answers a question, the subsequent question will tend to be an easier item. When an examinee correctly answers a question, the subsequent question will tend to be a more difficult question. In adaptive testing, items are typically selected from an item bank that has been scaled using IRT methods. Periodically, the item banks are updated or replaced. Wang and Kolen demonstrated that when the item banks are modified, it is important to ensure that the scores from the updated bank are comparable to those from the earlier item bank.

Processes Related to Equating

There are processes related to equating properly considered under the general category of "linking" in the terminology of both Linn and Mislevy. One of these processes is vertical scaling, which is often used with elementary school achievement test batteries. In these batteries, students are administered test questions that match their grade level. Scores from the tests administered at different grade levels are placed on the same score scale, enabling school personnel to chart an individual student's growth. Because the tests given at the different grade levels differ in difficulty and content, the process of placing the test levels on the same score scale is not equating. Other examples of linking include relating scores on one test to another (e.g., ACT scores and SAT I scores) and scaling tests within a battery so that they have the same score distribution. Although similar

statistical procedures are used in linking and equating, their purposes are different.

Conclusion

The goal of test form equating is to use scores from alternate test forms interchangeably. Test development procedures that have detailed content and statistical specifications allow for the development of alternate test forms that are similar to one another. These test specifications are a necessary prerequisite to the application of equating methods.

See Also the Following Articles

Education, Tests and Measures in • Item Response Theory

Further Reading

American College Testing Program (ACT) (1989). *Preliminary Technical Manual for the Enhanced ACT Assessment.* American College Testing Program, Iowa City, IA.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (1999). *Standards for Educational and Psychological Testing.* American Educational Research Association, Washington, DC.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In *Educational Measurement* (R. L. Thorndike, ed.), 2nd Ed., pp. 508–600. American Council on Education, Washington, DC.

Brennan, R. L. (ed.) (1989). *Methodology Used in Scaling the ACT Assessment and P-ACT+.* ACT Publications, Iowa City, IA.

Donlon, T. F. (ed.) (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests*. College Entrance Examination Board, New York.

Flanagan, J. C. (1951). Units, scores, and norms. In *Educational Measurement* (E. F. Lindquist, ed.), pp. 695–763. American Council on Education, Washington, DC.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Sage, Newbury Park, CA.

Holland, P. W., and Rubin, D. B. (1982). *Test Equating.* Academic Press, New York.

Kolen, M. J., and Brennan, R. L. (1995). *Test Equating: Methods and Practices.* Springer-Verlag, New York.

Linn, R. L. (1993). Linking results of distinct assessments. *Appl. Measurement Education* **6,** 83–102.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Erlbaum, Hillsdale, NJ.

Lord, F. M., and Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Appl. Psychol. Measurement* **8**(4), 453–461.

Mislevy, R. J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects.* ETS Policy Information Center, Princeton, NJ.

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1989). Scaling, norming, and equating. In *Educational Measurement* (R. L. Linn, ed.), 3rd Ed., pp. 221–262. Macmillan, New York.

Wang, T., and Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *J. Educational Measurement* **38**(1), 19–49.

Test–Retest Reliability

Chong Ho Yu
Cisco Systems/Aries Technology, Tempe, Arizona, USA

Glossary

carry over effect A type of data contamination caused by carrying over some influence to the item response from one condition to another, such as recalling answers from the first test administration.
coefficient of stability and equivalence The correlation coefficient yielded from a combination of test–retest and alternate form methods, in which different forms are given to the same subjects on different occasions.
Cohen's Kappa A measure of the degree of agreement between two sets of the frequency counts given that the data are categorical.
deliberative effect A non-spontaneous source of data contamination that could be caused by different motivations, such as deliberately taking the first test less seriously for practice or giving up on the test due to frustration.
intraclass correlation (ICC) A reliability estimate oriented toward the computation of inter-rater reliability, but which also can be employed to estimate the reliability of subjective scoring (same raters on different occasions) in the context of test–retest reliability.
Kendall's tau A nonparametric measure of the agreement between two rankings.
learning effect A type of carry over effect caused by the opportunity of practice in the first testing which results in gaining improvement in the second testing.
maturation effect The real changes in the traits, such as improvement in performance, which will likely occur when the time gap between two tests is very long.
testlet A cluster of items based upon the same scenario, such as items referring to a passage in a comprehension test.
Yule's Q A measurement of association based upon the odds ratio, which measures the ratio between the probability that an outcome would occur and the probability that the same outcome would not occur.

The "test–retest reliability estimate" measures the degree of stability of a test taken by the same subjects on different occasions. In contrast to internal consistency, test–retest reliability estimation compares a set of data with another set external to the first set, and thus test–retest reliability is also known as external reliability. Since data are collected at different times, it is also called temporal reliability. The attribute "stability" could be viewed as the means and "reproducibility" as the end. In other words, a stable test is considered a good test because using the same test would yield reproducible results. Thus, test–retest reliability is also conceived as test–retest reproducibility.

Meanings of Test–Retest Reliability

It is not uncommon that many students identify reliability as a mathematical formula or a computational procedure. For example, Cronbach Alpha is usually equated with internal consistency, whereas Pearson correlation coefficient is strongly associated with test–retest reliability. It is important to point out that reliability should be construed conceptually rather than computationally. For example, besides Cronbach Alpha, Kuder-Richardson 20 (KR-20) and split-half methods can also be employed for estimating internal consistency. By the same token, a test–retest reliability estimate could be computed in more than one way. To be specific, Pearson correlation coefficient is an appropriate indicator of the relationship between two sets of interval-scaled data, while Cohen's Kappa, Kendall's Tau, and Yule's Q are suitable to correlate the frequency of categorical data. But data that involves subjective scoring (same raters on different occasions) necessitate the use of intraclass correlation (ICC). In some other cases, researchers have gone even further to suggest that ICC is a better indicator of test–retest reliability than Pearson's, because Pearson's correlation does not imply stability, but ICC can in principle be interpreted

as a measure of stability. Further, some researchers have employed paired *t* tests, also known as correlated *t* tests, to detect subtle changes between two measures as a supplement to reporting correlation coefficients. When more than two measures are administered, researchers have expanded the *t* test approach to Analysis of Variance (ANOVA) repeated measures. Indeed, ICC can be derived from ANOVA models. Other innovative uses of psychometric procedures can be found in complicated research design. In short, the magnitude of stability could be estimated by more than one procedure. Interestingly enough, kappa, tau, Q, and ICC could also be used for estimating interrater reliability. Thus, these statistical procedures are not exclusively tied to a particular type of reliability. Their proper applications depend on the conceptual understanding of the data. Hence, it is recommended that researchers approach the issue of test–retest in a conceptual fashion rather than confining test–retest to particular computations.

Sample Issues

It is highly recommended that for estimating test–retest reliability, researchers should recruit more subjects than they need because when multiple tests are given to the same group of subjects, it is likely that in a later session some subjects may drop out from the study. Besides sample size, the quality of the sample is also important to test–retest reliability studies. No doubt the quality of the sample is tied to the sample representativeness. If a researcher designs a test for clinical groups, it is essential that the test–retest reliability information is obtained from those particular groups. For example, schizophrenics are said to be difficult to test. It is expected that the mental state of patients who suffer from schizophrenics is unstable, and thus sometimes the use of normal subjects is necessary. However, in a clinical setting a reliability estimate obtained from a normal sample for a test designed for use with abnormal samples may be problematic. Take the Drug Use History Form (DUHF) as another example. DUHF is designed to track usage of drugs among drug users, and its test–retest reliability information is crucial for clinicians to carry out treatment-effectiveness studies. However, due to the physical and emotional weaknesses of drug users, availability of self-report data from drug users may be scarce, and hence sometimes nondrug users participate in test–retest reliability studies of DUHF. As a result, contamination by nondrug users artificially inflates test–retest reliability coefficients of DUHF. The problem of sample representativeness could also be found in widely used diagnosis tests in education. Reports of test–retest reliability of Reading Disabilities (RD) tests have been questioned by certain researchers, because individuals who experience difficulty with reading may exhibit a limited range of reading performance; multiple measures of people who could not read at all would not yield a meaningful test–retest study result. To counteract this shortcoming, it is suggested that measures of RD should be based upon samples who have acquired basic reading skills by receiving interventions.

Another controversial aspect of sample representativeness is the use of convenience sampling rather than true random sampling. Traub criticized that the "canon" of sample representativeness has been overlooked by many social scientists because very often reliability studies are conducted on convenience samples, consisting of subjects who are readily accessible by the experimenter. Traub was doubtful of whether reliability estimates based upon "grab-bag" samples could be generalized beyond the sample itself. While Traub's criticism is true to some certain extent, other researchers have argued that "true random sampling" is an idealization. It is extremely difficult, if not impossible, to randomly draw samples from the target population to which the inference is made. For example, if the target population is drug users in the United States, ideally speaking subjects should be randomly selected from drug users in all 50 states, but in practice this goal may be very difficult to accomplish. Some have suggested that broader generalizations with convenience samples is still possible when different reliability studies using the same instrument are carried out in different places, and then meta-analytical techniques are employed to synthesize these results.

Sources of Errors

Uncontrollable measurement errors are inherent in every test and the test administrators can do virtually nothing to avoid this kind of error. For instance, a subject may quarrel with a spouse before taking the test, and thus this emotional state deeply affects the test performance. Other uncontrollable sources of measurement errors may be physical illness, fatigue, bad weather, or malfunctioning air-conditioning in the test centers. Some uncontrollable errors are specific to a test–retest design. For example, regardless of how much careful control is exercised, the conditions of the second testing would never be exactly the same as that of the first testing. This phenomenon is called person-by-occasion interaction. The estimates of the standard error of measurement would no doubt be inflated by this interaction. However, it is impossible to produce an estimate of test–retest reliability that is totally free of this interaction. In the following, focus is directed to certain controllable errors. These errors either could be avoided at the stage of experimental design or could be taken into account at the stage of computation.

Errors from the Subjects

Carry Over Effect

A carry over effect, as its name implies, is an effect on the item response that "carries over" from one condition to another. In other words, events in the first testing may influence performance in the second testing. This contamination could happen in several ways. In most cases, the test–retest reliability is inflated because the subjects still remember what they answered the first time and thus tend to give the same answer in the second test. In some situations, the carry over effect is the learning effect, also known as the practice effect. In this case, the skill level of the subjects improves because the first test provides them an opportunity to practice.

There is a subtle difference between the carry over effect and the learning effect. The learning effect is a subset of the carry over effect, but the carry over effect may not necessarily be the learning effect. To be specific, on the second occasion of a survey the subject may recall what they have answered on the first occasion. Some influence is definitely carried over from one situation to another, but in this case no learning or skill improvement is involved at all. In an ability test the subjects could put down the same wrong answers in the second test as what they did in the first testing. It could happen when the time interval between the two tests is very close and thus the subjects does not have a chance to look up the right answers. Again, there is no learning effect in this type of carry over.

Some researchers argued that the effect of the regression to the mean may balance out the learning effect. Regression to the mean is a statistical concept invented by Francis Galton in the 19th century. At that time, this notion was used as an argument against the natural selection theory proposed by Charles Darwin. In Darwinism, certain traits of species would get better and better in terms of survival fitness. But Galton argued that very tall parents do not necessarily give birth to very tall children. Instead, it is likely that the height of their offspring would approach the mean height of the population. This phenomenon is called regression to the mean. Galton believed that in the long run the regression effect will cancel out the short-term improvement. In social measurement regression to the mean is regarded by some researchers as a counter-balance against the learning effect. However, it is important to note that the alleged counter-balance effect due to the regression to the mean is most likely to occur in the long run. Due to the fact that in most studies of test–retest reliability subjects are tested two or three times within a short period of time, some researchers have argued that it is difficult to imagine how the regression effect could counteract the learning effect in such a short term. In brief, the threat of the learning effect should still be taken seriously.

Randomization of items is a common technique as a countermeasure against the carry over effect. When the order of items and order of options within an item are shuffled, it reduces the probability that the test takers can recall the answers in the first testing. Another technique is to introduce alternate forms into a test–retest study. In this case, virtually nothing could be recalled from the first test. This will be discussed below in the section Coefficient of Stability and Equivalence.

Deliberative Effect

In the previous discussion, the learning effect is spontaneous. But on some occasions, the learning effect is a result of deliberation. For example, in a certification examination administered in the information technology industry, such as the Microsoft Certified System Engineer Exam, examinees are allowed to retake the exam over and over. In the first testing, some examinees just take it as a learning experience by getting a preview of the test content and format. This preview serves as a guideline for them to study and to take the next exam seriously. Needless to say, the test–retest reliability is affected by this intention.

Other kinds of intentions and deliberate acts that deviate from normal test-taking behaviors could also seriously affect test–retest reliability. For instance, uncooperative test takers may object to the second testing and deliberately mismark the second test. In an ability test, some examinees who performed poorly in the first testing may hope to improve their performance in the second one. But when they find that the second test is equally as challenging as the first one and they are unlikely to get a higher score, they just deliberatively give up the test by skipping questions or answering them arbitrarily. In both cases, the test–retest reliability is deflated. In some situations, some subjects may seek advice from other test takers about how to improve their scores. This behavior leads to a violation of the independence of test scores, in which the measurement for one examinee must not be influenced by or have influenced on the measurement for any other examinee.

Taking all the preceding effects into consideration, it is advised that data patterns should be carefully examined by exploratory data analysis in order to spot observations that display strange patterns. For example, when a test taker is discouraged by the difficulty of the second test and stops devoting effort to answering questions, the scores of the items near the end would be much poorer. These observations in question could be put aside to avoid contamination of the reliability estimates. By the same token, when a test taker did not take the first test seriously but used it for practice only, the gap between his/her two test scores would be substantively large.

Errors from the Administration

Nonidentical Administrative Procedures

The circumstances during the test administration of the first testing should be identical to the circumstances of the second testing. Although it sounds common sense, nonidentical administrative procedures do happen from time to time when different administrators preside on different occasions. For example, in the first testing if a subject leaves the test center to go to the restroom and returns 10 minutes later, the test administrator may give him/her an extra 10 minutes as compensation. But in the second testing another administrator might count the time for the restroom as a part of the testing time, and thus inconsistency arises. In order to achieve identical test administrations, it is strongly recommended that the same people should administrate the tests on all occasions or/and a clear protocol is given to the test administrators.

Poor Test Instruction

The instruction of the test is a commonly overlooked source of error. When the instruction is poor, the subjects may not know what they are supposed to do immediately. This problem is often found in computer-based testing. For example, if a subject is asked to perform simulation-based tasks in a computer test while the instructions are unclear, subjects may do better on one occasion than the other.

Subjective Scoring

If the test format is subjective, such as consisting of essay-type questions, subjective scoring could be a source of measurement error. To be specific, errors from subjective scoring can come from two different sources. One type of error results from different raters (interrater) and the other is caused by the same rater conducting grading on different occasions (intrarater). This issue could be very complicated since it involves multiple sources of errors, or an interaction effect between the time factor (test–retest) and the rater factor. The generalizability theory, which will be discussed later, is proposed as an effective strategy to address the problem of multiple sources of errors. In addition, the intraclass correlation (ICC) coefficient can be computed to estimate the intrarater and interrater reliability. This estimate is based on mean squares obtained by applying ANOVA models. To be specific, in an ANOVA model the rater effect is indicated by the mean square (MS) of the between-subject factor while the multiple measures on different occasions are shown in the mean square of the between-measure factor. The reliability estimate is calculated by

$$r = \mathrm{MS}_{\text{between-measure}} - \mathrm{MS}_{\text{residual}} / \mathrm{MS}_{\text{between-measure}} + \left(\mathrm{df}_{\text{between-subject}} \times \mathrm{MS}_{\text{residual}}\right)$$

In the context of ANOVA, there are separate coefficients for three models:

1. One-way random effects model: Raters are perceived as a random selection from possible raters, who grade all subjects, which are a random sample.
2. Two-way random effects model: Raters, also conceived as random, rate a subset of subjects chosen at random from a sample pool.
3. Two-way mixed model: All raters rate all subjects, which are a random sample. This is a mixed model because the raters are a fixed effect and the subjects are a random effect.

Too Wide or Too Narrow Time Gap

As expected, time gap is a major factor that influences test–retest reliability. If the interval between the two test administrations is short, say three minutes, the correlation between the two test scores, needless to say, is boosted because of the carry over effect. In contrast, if the interval is very long, say three years, the correlation is expected to be low. In the latter case, the ability can substantively change due to the maturation effect, which is a real change in the trait under study. In this case, a low test–retest correlation may indicate low reliability, real changes in the individuals measured, or a combination of both.

Pedhazur and Schmelkin argued that because of these serious deficiencies of the test–retest approach, this should not be used or should be used with caution. Nevertheless, some researchers counterargued that these deficiencies are not insurmountable. Experienced researchers who are familiar with the subject matter usually know how to choose a proper time gap to minimize both the carry over effect and the maturation effect. Further, these so-called deficiencies are not inherent in every test–retest. Crocker and Algina argued that in a test designed for assessing the level of an infant's psychomotor development, it is unlikely that the baby can remember the previous responses. Last but not least, researchers could design test–retest studies with different temporal gaps to provide users with both short-term and long-term reliability estimates.

Errors from the Test

Nature of the Subject Matter

If the researcher attempts to measure personality, interest, attitude, or a transient state or mood, such as anger, fear, and anxiety, test–retest reliability would be bound to be low. On the other hand, test–retest reliability estimates tend to be higher when stable traits are measured. According to Crocker and Algina, among the highest test-test coefficients reported for commercially published tests are aptitude tests, such as the Wechsler Adult Intelligence Scale (WAIS). Thus, researchers are encouraged to carefully examine the appropriateness of

the subject matter before adopting test–retest reliability estimation.

Difficulty Levels of Items

In an ability test, the test–retest reliability tends to be high if the questions are too easy, simply because the subjects will get them right on both occasions. In contrast, if the items are extremely difficult, it will lead to the same effect since the subjects will miss most items both times. Nevertheless, neither case is a good sign because when a test is too easy or too difficult, it cannot discriminate people of high proficiency from those of low proficiency.

Test Length and Item Randomization

Very few people see test length as a source of error for test–retest reliability, but low test–retest reliability could occur when the test is long and the item/option sequence is randomized. As mentioned before, item/option randomization is a way to counteract the carry over effect. However, when an ability test is composed of too many items, test takers may be too fatigued to answer items near the end, and thus performance toward the end of the test may be relatively poor. If the test is an aptitude test, the subject may be too bored to answer questions near the end seriously. As a result, the quality of the answers is affected. In both cases, since the items near the end at the second administration are not the same as those at the first, it is expected that test–retest reliability is affected. As a remedy, it is recommended that the test length be kept short so the quality of answers to randomized items would not be affected by fatigue or boredom.

Test–Retest Reliability in a Wider Perspective

It is a widespread impression that test–retest, internal consistency, and alternate forms are separate methods in both computational and conceptual senses. Actually, test–retest and alternate forms could be blended together for estimating the coefficient of stability and equivalence. In addition, some researchers have proposed that the generalizability theory could be employed to take different sources of error into account, and also stability as a unified theme for all three types of reliability estimates.

Coefficient of Stability and Equivalence

Reliability coefficients can be estimated by combining the test–retest and the alternate form approaches. Instead of giving the same test to the same subjects on two different occasions, in this approach the test administrator gives two different forms to the same subjects in two different situations. The advantage of this method is that the carry over effect is avoided because items in the two tests are not the same, although the contents are equivalent. The tradeoff is that another potential source of error is introduced—the content effect. As a result, there will be two sources of measurement errors in this mixed method. One source is content sampling in the form construction, and the other is change in subject performance over time. The correlation coefficient between the two sets of scores is termed the coefficient of stability and equivalence.

Generalizability Theory for Addressing Multiple Sources of Error

Addressing multiple sources of error is an interesting idea, but classical test theory directs researchers to focus on one source of error with different computing methods. For example, if one computes a test–retest reliability coefficient, the variation over time in the observed score is counted as error, but the variation due to item sampling is not. If one computes Cronbach coefficient Alpha, the variation due to the sampling of different items is counted as error, but the time-based variation is not. This creates a problem if the reliability estimates yielded from different methods are substantively different. To counteract this problem, Marcoulides suggested reconceptualizing classical reliability in a broader notion of generalizability. Instead of asking how stable, how equivalent, or how consistent the test is, and to what degree the observed scores reflect the true scores, the generalizability theory asks how the observed scores enable the researcher to generalize about the examinees' behaviors given that multiple sources of errors are taken into account.

Stability/Reproducibility as a Unified Theme of all Reliability Estimates

Although the meanings of stability, consistency, and equivalency are different from each other, they all share a common theme: All of them address the extent to which scores obtained by a test taker will be the similar if the same person is retested by the same test on different occasions. At first glance, this common thread describes stability or reproducibility rather than consistency and equivalency. Nevertheless, when a researcher computes Cronbach Alpha to obtain an estimate of internal consistency, he/she is not satisfied with the following inference: "The response pattern of this test is internally consistent. Externally speaking, it does not give me any information about what the response patterns would look like when the same set of items are used for another sample on another occasion. Hence, this test can be applied to this local sample only." Actually, this is not the goal of the researcher. He/she would certainly collect more data from other sources and to verify whether the Cronbach

coefficient Alpha is stable across various samples. It is hoped that the same instrument would yield reproducible results elsewhere. The same goal can be found in the use of alternate forms. Thus, while test–retest is a direct way of measuring test–retest reliability, other forms of reliability estimation are also regarded as an indirect means of seeking stability and reproducibility information.

Special Applications

Although the "generalizability theory" and "stability" are proposed as two unified themes of internal consistency, test–retest, and alternate forms, test–retest reliability estimate nevertheless has certain specific applications. To be explicit, in some situations only test–retest could be used for reliability estimation. In the following, three examples will be discussed.

Frequency of Categorical Data

In some tests stability is the only psychometric attribute that can be estimated and thus only test–retest can be applied. The Rorschach test, also known as the Rorschach inkblot test, is a good example. The test is a psychological projective test of personality in which a subject's interpretations of abstract designs are analyzed as a measure of emotional and intellectual functioning and integration. Cronbach pointed out that many Rorschach scores do not have psychometric characteristics that are commonly found in most psychological tests. Nonetheless, researchers who employ the Rorschach test can encode the qualitative-type responses into different categories and frequency counts of the responses can be tracked. Hence, the stability of these tests can be addressed through studies of test–retest reliability. The following three approaches are widely adopted.

Cohen's Kappa

Cohen's Kappa coefficient, which is commonly used to estimate interrater reliability, can be employed in the context of test–retest. In test–retest, the Kappa coefficient indicates the extent of agreement between frequencies of two sets of data collected on two different occasions.

Kendall's Tau

However, Kappa coefficients cannot be estimated when the subject responses are not distributed among all valid response categories. For instance, if subject responses are distributed between "yes" and "no" in the first testing but among "yes," "no," and "don't know" in the second testing, then Kappa is considered invalid. In this case, Kendall's Tau, which is a nonparametric measure of association based on the number of concordances and discordances in paired observations, should be used instead. Concordance could be found when paired observations co-vary, but discordance occurs when paired observations do not co-vary.

Yule's Q

Even if the subject responses distribute among all valid categories, researchers are encouraged to go beyond Kappa by computing other categorical-based correlation coefficients for verification or "triangulation." Yule's Q is a good candidate for its conceptual and computational simplicity. It is not surprising that in some studies Cohen's Kappa and Yule's Q yield substantively different values and the discrepancy drives the researcher to conduct further investigation.

Yule's Q is a measurement of correlation based upon the odds ratio. In Table I, a, b, c, and d represent the frequency counts of two categorical responses, "yes" and "no" recorded on two occasions, "Time 1" and "Time 2." The odd ratio, by definition, is $OR = ad/bc$. Yule's Q is defined as $Q = (OR - 1)/(OR + 1)$.

Testlet

The test–retest approach is recommended when testlets are included in a test. A testlet is a cluster of items based upon the same scenario. A typical example is a comprehension test using the same passage on which a group of items is based. The absence of local independence is a serious problem in assessing internal consistency. But in estimating test–retest reliability, the issue of violation of local independence is less serious, because test–retest reliability is not locally assessed; rather, the response pattern of each item on one occasion is paired with that of another occasion. Even if response patterns of testlet items are internally correlated, this does not increase the measurement error of external correlation.

Single-Item Measures

It is a common practice that psychologists and sociologists use a group of items to measure a single construct. However, certain clinical tests make a direct measurement of a narrowly defined variable. Those tests might contain just a single item. Expanded Disability Status Scale (EDDS) and Functional Systems (FS) are two good examples. Using Cronbach Alpha, which relies on interitem association, to measure reliability of such tests is out of question. Obviously, the most appropriate method

Table I 2 × 2 Table of Two Measures

Time 1 \ *Time 2*	*Yes*	*No*
Yes	a	b
No	c	d

of estimating reliability of single-item measures such as EDSS and FS is the test–retest approach.

Controversies

Cutoff of Test–Retest Reliability

There is no universal agreement about how a test–retest reliability estimate is considered adequate. It is not surprising that different textbooks give different recommendations. A popular suggestion is that a high correlation should be 0.80 or above. However, quite a few researchers warn that this suggestion is too arbitrary to be taken seriously. Critics draw an example in internal consistency as a counterexample: In 1978 Nunnally suggested that a reliability of 0.7 is acceptable, but about a decade later he changed the cutoff from 0.7 to 0.8. Actually, the variable being measured will change the expected strength of the correlation. Rust and Golombok suggested that for IQ tests a reliability of 0.9 is acceptable, but for personality tests, which measures less stable traits, a reliability of 0.7 is good enough. Some have argued that clinicians should hope for a correlation of 0.9 or above in tests that will be used to make decisions about individuals. Other high-stakes examinations should also follow this strict principle.

Moreover, since there is more than one procedure to measure test–retest reliability estimates, the acceptability of a reliability level should also depend on the procedure the researcher adopted. For example, in Kappa coefficient, even if 70% of two datasets concur with each other, it does not mean that the measurements are reliable enough. Since the outcome is dichotomous, there is a 50% chance that the two measurements will agree. Thus, Kappa coefficient demands a higher degree of matching to reach reliability. Fleiss suggested that a Kappa coefficient of 0.75 or above is considered excellent, coefficients between 0.4 and 0.75 represent fair agreement, and a reliability of 0.4 or less is unacceptable. For ICC, Strout and Fleiss regarded 0.7 as the cutoff between acceptable and unacceptable reliability values. Landis and Koch provided a detailed classification of the quality of ICC, as shown in Table II.

As mentioned before, uncontrollable errors always sneak into the measurement process no matter how carefully we plan and implement a study. Thus, rather than reporting a fixed reliability coefficient only, it is a good practice to report also the confidence interval (CI) of the reliability estimation.

Table II Classification of ICC

Slight	$0 < \text{ICC} = 0.2$
Fair	$0.2 < \text{ICC} = 0.4$
Moderate	$0.4 < \text{ICC} = 0.6$
Substantial	$0.6 < \text{ICC} = 0.8$
Almost perfect	$0.8 < \text{ICC} = 1$

Stability as a Psychometric or Datametric Attribute

Although stability is considered a vital property in test–retest and other forms of reliability estimate, in recent years, stability as an indispensable property of test reliability has been challenged by Thompson and Vacha-Haase. They suggested that "psychometrics is datametrics." To be specific, in their view reliability, especially in the form of stability, is not a psychometric attribute of a test; the phrase "the reliability estimate of the test" is misleading. Rather, reliability is associated with the data and thus it inherently fluctuates from sample to sample. Thompson and Vacha-Haase introduced a research methodology called meta-analyses of reliability across studies, also known as reliability generalization studies. Based on their massive empirical studies employing meta-analysis, they found that reliability information fluctuates from sample to sample and thus the so-called stability of a test is a myth. No doubt this notion poses a serious challenge to the concept of test–retest reliability estimation, which is expressed in terms of stability, and to a lesser degree, to other forms of reliability estimate.

Nevertheless, most researchers still accept that stability is a psychometric, not datametric, property for a number of reasons. With the threats of many sources of measurement errors and "noise," it is not surprising to see fluctuations of reliability estimates across data sets. Many researchers realize that testing methods in the social sciences are inevitably imprecise. Within the fields of physical and engineering sciences, direct measurements are possible, such as the strength of electrical-magnetic interference and the throughput of a fiber optics cable. In contrast, in the social sciences, measurements are often indirect; researchers need to relate observables to the latent construct. Therefore, instead of denying reliability as a psychometric attribute, the goal of a test constructor should be to reduce as many measurement errors as possible, and also to identify the relationship between the observable and the hidden construct. On one hand, it may not be appropriate for a test developer to apply a "universal" test to all populations. On the other hand, he/she should retain the goal of constructing a fairly "stable" test, in which the psychometric properties are invariant within a broad range of occasions.

See Also the Following Articles

Alpha Reliability • Reliability • Reliability Assessment

Further Reading

Allen, M. J., and Yen, W. M. (1979). *Introduction to Measurement Theory.* Brooks/Cole, Monterey, CA.

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Holt, Rinehart, and Winston, New York.

Groth-Marnat, G. (1990). *Handbook of Psychological Measurement,* 2nd Ed. Wiley, New York.

Kline, P. (2000). *Handbook of Psychological Testing,* 2nd Ed. Routledge, New York.

Litwin, M. S. (1995). *How to Measure Survey Reliability and Validity.* Sage, Thousand Oaks, CA.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA.

McGraw, K. O., and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1,** 30–46.

Rust, R., and Golombok, S. (1989). *Modern Psychometrics: The Science of Psychological Assessment.* Routledge, London.

Theory, Role of

David Byrne
University of Durham, Durham, United Kingdom

Glossary

classification The process both of constructing a set of categories into which cases may be assigned and the actual assignation of cases to those categories.
complexity theory An interdisciplinary understanding of reality as composed of complex open systems with emergent properties and transformational potential. A crucial corollary of complexity theory is that knowledge is inherently local rather than universal.
constructionism An ontological position which asserts that the social world is the product of human action and is socially constructed.
conventionalism An epistemological position which asserts that the objects of knowledge are specified by agreement among those who work with terms describing them and do not necessarily have any "real" counterparts separate from intellectual use.
methodology The application of ontological and epistemological understanding to the actual practice of research and critique of the methods employed in research.
realism A meta-theory that accepts the existence of a world separate from and preexisting our knowledge of it, understands processes of causation as inherently complex and contingent, and recognizes that the production of knowledge is an inherently social process.

Theory, the role in social measurement: consideration is given to the implications of both ontological and epistemological positions for the understanding of the objects of social measurement and the nature of measurement as a process.

The Necessity for Theory in Social Measurement

There is a general agreement that theoretical understanding is necessary for the proper conduct of social measurement. Blalock emphasized the relationship between conceptual formation and underlying theoretical assumptions in any social research arguing that these must always be associated in the actual practice of research. By theoretical concerns Blalock means substantive social theory which must always be linked to our methodological considerations. Blalock's position typifies that of mainstream quantitative sociology in principle, although the practice has seldom if ever succeeded in establishing the necessary linkages.

Another approach, which has been particularly influential in psychology, emphasizes general theories of measurement applicable in all contexts. Here the theory of measurement—often highly formalized in axiomatic terms—drives the whole measurement process. Kyburg identifies two strands—a mathematical focus on mapping empirical structures onto the structure of real numbers and an experimental focus concerned with the establishment of valid scales that measure real variate aspects. Cicourel in the enormously influential *Method and Measurement in Sociology* drew on both strands in constructing a theoretical justification for the rejection of a programme of quantification in sociology since sociological measurement was incapable of meeting the necessary criteria of either abstract measurement theory or scaling.

Pre-1990 discussions of social measurement often debated the ontological validity of positivism's assertion that the social and physical worlds could be understood as essentially similar, but generally accepted an understanding of the nature of measurement as essentially concerned with quantifying social variables. More recently, developments in the philosophy of science, the innovation in methods, and the history of social quantification have moved the argument in a radically different direction. Although these labels are seldom employed by the protagonists, we can identify a realist tendency and a constructionist tendency. Realists accept the existence

of a social reality external to and separate from observers but recognize that causal processes are complex and contingent, and that any account of reality is socially constructed, albeit that it usually has some relationship to the reality it is seeking to describe. Social constructionists, typified by Desrosières, do not deny the reality of measurements but consider that this is a product of the general social use of them—a conventionalist approach. Social measurements are real because people act on them as if they were.

New Metatheories and Social Measurement

Realism offers a social ontology that can combine social constructionism and an acceptance of the existence of the objects of measurement. The essential relevant realist proposition is that social outcomes are complex, which means that they are generated by a multiplicity of processes and their interaction effects. In consequence, we cannot understand them through analytical processes nor access them through controlled experimentation.

This is compatible with the social constructionist account of social structure, but it does not correspond to Desrosières' extreme formalism in which the reality lies in the conventionally established social measurement that cannot be established as corresponding to anything existing prior to its construction. The realist position regards the objects that we measure as real, although recognizing that the social process of measurement shapes the form in which we know them in quantitative terms.

Realist approaches must be contrasted with the traditional understanding of measurement as constructed around real variables. Abbott has characterized 20th century social science as obsessed with causal analysis purporting to describe the relationships among variables that have a real existence separate from the social entities, macro, meso, or micro, from which analysis has abstracted them. Traditional measurement theory seems doubly Platonist in its assertion first of the necessary isomorphism of mathematical structures with social reality and second in its insistence on the reality of the things—abstracted variables—which vary separate from the cases for which that variation is measured.

The insistence of traditional discussions of measurement on the significance of validity is revealing. Measurement processes are considered to be of value to the extent to which they achieve a valid measurement, i.e., to the extent to which the measurement generated corresponds to the true value of the real variable across the range of cases for which measurements are made—construct validity. Yet, in general, validity can only be established not in content terms, which would necessarily involve an ability to specify the character of the domain of interest prior to and beyond the measurement process, but through predictive validity where it is the performance of the measure which signifies. The influence of factor analytical techniques on such theorization of measurement is considerable since these techniques seem to provide a mathematical solution to the problem of accessing the real—the factors—through indirect measurement of attributes of cases. However, factor analysis cannot establish the reality of the factors. Moreover, factors are intrinsically abstracted from data about cases; they exist separate from any real case.

The Classical Approach—Measuring Variables

Operationalization

All actual measurements of variables are achieved by a process of operationalization in which some rules for the measurement process are specified and these, when enacted, generate the measurement. These rules of process constitute the operational definition of the variable measured. Carley points out in a discussion of the special case of social indicators something which has general force in describing the relationship between a set of measurements based on an operational definition and the underlying concept to which the measurement set is supposed to correspond. We must accept, and properly should specify, the nature of the causal linkage between the observable phenomena which give rise to the measurement and the unobservable underlying reality. For Carley, this requires the specification of a theory of relation between empirical observations and the underlying system that gives rise to them.

Auxiliary Theories

Blalock reiterated his longstanding argument that we cannot rely on a single theory of measurement to sustain the relationship between the enactment of operational definitions and our conceptual formations. Rather, we are forced to bring into play a whole set of auxiliary theories, many of which cannot not be tested. In contrast with approaches in psychology which distinguish between direct and indirect measurement, Blalock suggests that we can never achieve direct measurement but that all social measurement is necessarily indirect and dependent on a causal model that incorporates to the best of its abilities the auxiliary measurement theories that inform the construction of that model.

Measurement by Fiat

Cicourel asserted, correctly, that generally the approach adopted by researchers in order to resolve these problems

was simply to measure by fiat. In effect, whatever indicator is conveniently available is declared to be an indicator of an underlying theoretical construct. There is an interesting literature on definitions of social class which illustrate this issue exactly. We can say with some confidence that whenever things are measured without a specific justification of correspondence between underlying concept and operational definition then the measurement is by fiat alone.

Classification—Problem or Solution?

The Problem of Multidimensionality

The apparently most elementary form of measurement is classification: the assignation of a value on a nominal variable to a case in order to indicate membership of a particular set—for example, to use the number one to indicate membership of the set of males and the number two to indicate membership of the set of females. Traditional variable centered analysis has tended to be scornful of mere categorization. This is in part because until the development of logistic regression and related techniques, it was difficult to employ categorical variables in causal modeling. However, Blalock made a more subtle point when he noted the tendency of social scientists to rely upon classification as a way of handling multidimensional variation. In general, social scientists did not properly engage with issues of multidimensionality in their data. Blalock's specification of the issue again involved intrinsic reification of variables, although the variables were now understood in a more complex way as ordered along multiple rather than single dimensions. Here, classifications can be understood as devices for taking note of multidimensionality which did not handle the real nature of such multidimensional real variables.

Theories of Classification

In general, the theorization of classification has been quite distinctive from the theorization of measurement. Essentially, the establishment of categories has been regarded as the major problem in taxonomy with the assignation of cases to membership of those categories—the actual process of nominal measurement—being secondary. Recently, the actual processes of typing have been subject to theoretical consideration. Bowker and Star, drawing in part on the conventionalist tradition that informs Desrosières' work, have returned to the distinction between Aristotelian and prototypical approaches to classification. In the Aristotelian scheme class membership depends on the possession of a given set of one or more distinctive attributes—nominal variable values in conventional measurement terms. The prototypical approach assigns cases to categories through comparison with a framed "ideal" of the category in a rather holistic sense. This basic distinction has considerable implications for our understanding of measurement because the Aristotelian approach is essentially analytical and compatible with a belief in the reality of variables, while prototypical categorization does not depend on analysis and hence does not require the measurement of variables as such. Of course, real classifications as social practice do, as Bowker and Star point out, usually involve some mixture of the two approaches but the distinction has considerable significance for our understanding of what measurement is actually dealing with.

Generating Typologies through Numerical Taxonomy

In measurement, the development of technology can have a profound influence on the actual processes by which things are measured, and one of the crucial tasks of theory in relation to measurement is to examine this relationship and its outcomes. Any theory of instrumentation must recognize that instruments are part of the process through which measurements are constructed and that developments in them have profound implications for what we are actually doing. Of particular significance in relation to classification has been the development of a set of techniques in which measures of variation across a large number of attributes for a large number of cases can be used not only to assign those cases to a preexisting set of categories but to actually generate categorical schemes ab initio. The original techniques for doing this were based on clustering algorithms using matrix algebra approaches. More recently, neural nets have been used for the same purpose. Although the underlying programming basis of the two approaches is very different, output is much the same and that is what matters here.

In numerical taxonomy, information about a large set of variate attributes of the members of a large set of cases is employed to make comparisons among those cases and to assign the cases to sets which are based on the principle of maximizing within set similarity and minimizing similarities among sets themselves. It is important to realize that this is not a simple Aristotelian polythetic scheme. In such a scheme, which corresponds exactly to locating a case within a cell in a multidimensional contingency table, cases must share all the attributes to be assigned to the set. Numerical taxonomies, which originally required that the attributes used to classify be measured as continuous data, do not impose this stipulation. This applies even when categorical variable attributes are used as the classifying principle. This, intuitively, has much in common with Ragin's understanding of fuzzy sets. The actual process of a hierarchical cluster analysis would

seem to have a considerable prototypical component in that clusters are established on the basis of most similar (in practice least dissimilar) cases and then other cases are progressively assigned to clusters in a process of fusion with the most similar being joined together to form a new cluster. Iterative relocation is possible at any given stage. The mathematically complicated process generates results with considerable intuitive appeal—an outcome which can readily be understood in the light of recent developments in cognitive theory.

Taking Cases as the Focus

Establishing Causality from Case-Based Data

Traditional causal analysis has employed regression-derived techniques that rely on an underpinning general linear model in order to specify causal relationships among variables. In sociology, a radically different approach has been available since Znaniencki's formulation of the approach of analytic induction, which was derived from a contrast between the population based approaches underpinning probabilistic statistical causal reasoning and the focus on single cases of bench scientists. In a series of works culminating in his discussion of fuzzy set approaches, Ragin has developed a method of establishing causal configurations which in contrast with statistical reasoning's emphasis on the (single) causal model, recognizes that particular outcomes may be produced by different combinations (configurations) of causes. The fuzziness indicates that set membership is not an absolute but may be a matter of degree. Although Ragin does not discuss realist metatheory, his approach has much in common with realism's understanding of the nature of causation. Coming from a different direction, Karl Popper's attention in his later work to "single case probabilities" has similar implications for our understanding of measurement.

The issue is that case-based approaches regard the case as the center of attention, not some reified variable that is abstracted out with the case. The contributors to Ragin and Becker review the implications of this important distinction. Traditional measurement theory was concerned with validation of variables. Case-based approaches are concerned more with answering in the particular, the general question posed by Ragin and Becker: What is a case? In quantitative social science the case is typically an entity—a country, a city region, a household, a firm. It might seem that the cases of this kind have a clear and coherent identity, but consideration of the example of city–region shows that this is not necessarily so. Therefore, the establishment of boundary conditions—delimiting cases from other cases—becomes a crucial task of measurement.

Handling Large Numbers of Cases in Causal Inference

Ragin's work was developed for macrosocial comparisons in relation to radical social changes. However, the same case-based logic of measurement can be applied to the elucidation of complex and multiple causal processes when dealing with large numbers of cases at the micro- or mesosocial levels. This requires us to think dynamically, to think about processes of change through time as crucial for social understanding and to recognize that a primary purpose of measurement is documenting the character of such changes through time. Here, a commonplace of general social theory becomes crucial to our understanding of what measurement is for. "Scientific measurement" as part of the Newtonian programme of science has been concerned with documenting changes of degree, hence the status accorded to continuous scale measurement. Social theory is much more concerned with changes of kind, with quality as type rather than quantity as number.

There is a variety of time-ordered approaches in quantitative methodology but they are typically variable centered. Time-ordered numerical taxonomies offer interesting possibilities for measurement in the sense of specifying trajectories of cases through time understood in a socially appropriate fashion. The time dimension for an ensemble of cases need not be calendar time, although that should always be recorded, but rather time of process, for example, stages in the treatment regimes (the plural is necessary) for patients going through a career as a person with a mental illness. We have no difficulty in demarcating the boundaries of the case understood as individual patient, but a crucial task of measurement is the specification of the boundaries of specific stages of process. Such approaches are in the early days of development, but they do seem to have considerable potential for elucidation of complex causation.

The Implications of Complexity Theory

Recent developments in complexity theory have profound implications for our general approach to measurement. Once our interest is focused on complex systems in which the interaction of parts of the system with each other, with the system as a whole, and with other systems becomes central to our process of understanding, then the purpose of measurement can no longer be understood in terms of a variable centered approach to the understanding of causality. Complex systems are characterized by emergence and the character of their trajectories, and in particular, changes of kind (phase shifts) cannot be understood by a programme of analysis. Byrne has proposed that our measurements must now be understood

not as measurements of real variables, except in the special circumstances where some external variation is imposed on the system, but rather as descriptions of variate traces of the system. This usage has two potential advantages. It reminds us that what matters is the system and that the measurement is a temporary, albeit useful, description of a characteristic of the system. It also emphasizes the significance of dynamism in measurement. Measurements are most useful when they enable us to record processes of change. Necessarily measurement of variate traces of complex systems must be multidimensional although there are interesting questions as to direct and indirect measurement to be considered here.

Measurements are external to the system and the variate traces measured do not necessarily correspond to any component of the system. They can be used to type the system. Examination of interaction among variate traces can also offer some clues as to the nature of complex processes driving emergence within the system and in engagement with other systems.

Complexity theory reinforces the significance of categorization as fundamental to social measurement. Complex systems are characteristically robust. Most of the time they remain much the same in character although by no means static. However, they can and do undergo phase shifts—transformations of kind—which we might consider metaphorically as metamorphoses. They become something very different without ceasing to exist. Most social sciences are profoundly engaged with such transformations and we can only identify them if we can specify kinds and measure change of kind, including the possibility that change of kind might involve not just the relocation of a complex system in an existing taxonomy but the creation of a whole new set of available categories of kind. Theorization of both the processes of taxonomy and the nature of transformation is essential to the measurement of complex change.

Cognition and Measurement—A Radical Departure

Lakoff, together with a range of coauthors, has argued for a radical recasting of our understanding of the relationship among human cognition and perception, language, and knowledge. The key phrase in this argument is embodied mind and the proposition is that knowledge as a product of human action derives from the relationship between human cognition and language and the world itself. Our systems of ideas are grounded in bodily experience—a wholly rational and modern version of Heidegger's antirational and antimodern assertion of dasein. Lakoff and Johnson subtitled the first chapter of their book "How cognitive science reopens central philosophical questions." This is a bold claim but they make a very convincing argument. Here, we can simply say that the notion of embodied mind is as fundamental for our theorizing of measurement as it is for everything else in the episteme. It is worth noting that these approaches seem wholly compatible with an emergent realist understanding of measurement. We come equipped—have been rendered equipped through evolutionary process—to understand and to know in a way that reflects and is shaped by our embodied selves' relationship with a real external world. Our inherent classificatory abilities are just one instance, albeit a very important one, of this capacity.

Conclusion—Theory and Measurement at the Beginning of the 21st Century

Twentieth century theorization of measurement was profoundly influenced by the idea of the variable as real. There have always been countervailing currents, particularly from that part of empirical sociology that was explicitly informed by social theory, but the notion of the variable and the problem of validating measurement of it dominated discussion and debate. In the last quarter of the 20th century, a range of new ideas has emerged—from empirical research in terms of the development of case-based methods, from philosophy of social science as critical realism, from systems theory with the development of understanding of complex systems, and from cognitive science with the proposition of the embodied mind. These have the most profound implications for the way we think about what it is we are measuring, what our measurements are, and what our measurements are for. The elements identified above come together as realist in their understanding of the answers to all these questions.

In contrast, the conventionalist approach, which in a somewhat different guise informed the more sophisticated versions of 20th century positivism, has been reinvigorated by arguments grounded in the sociological and historical examination of data construction. We must note however that in the hands of the one of the most sophisticated proponents of this position, Desrosières, even the conventions become real as we act in accordance with the implications of our social measurements. It seems likely that the debate between the conventionalist and realist understandings of measurement will be key to our understandings in the first part of this new century.

See Also the Following Article

Complexity Science and the Social World

Further Reading

Abbott, A. (1998). The causal devolution. *Sociol. Methods Res.* **27**(2), 148–181.

Blalock, H. M. (1982). *Conceptualization and Measurement in the Social Sciences.* Sage, Beverly Hills, CA.

Bowker, G. C., and Star, S. C. (1999). *Sorting Things Out.* MIT Press, Cambridge, MA.

Byrne, D. S. (1998). *Complexity Theory and the Social Sciences.* Routledge, London.

Byrne, D. S. (2001). *Interpreting Quantitative Data.* Sage, London.

Carley, M. (1981). *Social Measurement and Social Indicators.* Allen and Unwin, London.

Cicourel, A. V. (1964). *Method and Measurement in Sociology.* Free Press, New York.

Cilliers, P. (1998). *Complexity and Postmodernism.* Routledge, London.

Desrosières, A. (1998). *The Politics of Large Numbers.* Harvard University Press, Cambridge, MA.

Kyburg, H. E. (1984). *Theory and Measurement.* Cambridge University Press, Cambridge, UK.

Lakoff, G., and Johnson, M. (1999). *Philosophy in the Flesh.* Basic Books, New York.

Lakoff, G., and Nuñez, R. E. (2000). *Where Mathematics Comes From.* Basic Books, New York.

Ragin, C. C., and Becker, H. S. (1992). *What is a Case?* Cambridge University Press, Cambridge.

Ragin, C. C. (2000). *Fuzzy-Set Social Science.* University of Chicago Press, Chicago.

Theta Reliability

Mike W. L. Cheung
The University of Hong Kong, Hong Kong

Paul S. F. Yip
The University of Hong Kong, Hong Kong

Glossary

alpha coefficient A measure of internal consistency of a composite score.
composite score A score created by summing over several weighted or unweighted scores.
essentially tau-equivalent tests Measurements of the same construct (true score) with the same units of measurement, with differences by a constant, and possibly with different precision (error variance).
internal consistency The degree of interrelatedness among items.
parallel tests Measurements of the same construct (true score) in identical units and with the same precision (error variance).
principal component analysis A multivariate technique to reduce a large number of items to a smaller number of independent composite scores.
reliability coefficient An estimate of reliability; can be interpreted as the ratio of true score variance to observed score variance.
theta coefficient An alpha coefficient that is maximized by using optimal weights in creating the composite score.

The theta (θ) coefficient is an index of internal consistency for a composite score. It was proposed in the 1970s by David J. Armor as an alternative measure of internal consistency in tackling several problems encountered by the alpha (α) coefficient, which had been introduced in the 1950s by Lee J. Cronbach. Theta reliability can be interpreted as an alpha coefficient maximized by using optimal weights in creating a composite score. It is directly related to principal component analysis. In this article, a real example is offered to demonstrate the procedures for estimating the theta coefficient, and how these procedures contrast against those used for estimating the alpha coefficient is discussed.

Introduction

Reliability is a general term that is used frequently in daily life. For instance, a train schedule may be said to be very reliable, meaning that trains arrive and depart according to the times indicated in the schedule. However, real-world interpretations of reliability are quite different from scientific usages of the reliability concept in social measurement. In social sciences, reliability refers to the dependability, consistency, or repeatability of test scores for a particular population; validity refers to what the tests are supposed to measure. Using the scientific definition of reliability, for example, trains are reliable if they arrive and depart at nearly the same time, regardless of the times indicated on the schedule. If the trains consistently arrive and depart at the same time, as well as match the times indicated in the schedule, then the train schedule can be considered valid. Returning to the definition of reliability, then, there are two types of reliability: (1) consistency of items in a scale and (2) stability (test–retest) of the scores across time. For the purpose of this discussion, the term "reliability" will refer to internal consistency, as opposed to stability.

Reliability

Two concepts of the reliability coefficient merit discussion: the reliability coefficient for a single score and the reliability coefficient for a composite score; the alpha coefficient is an estimate of the latter.

Reliability of Single Scores

The most frequently used model to define reliability is the classical test, or true score, theory. In the classical test

theory, the observed score x is assumed to be an additive combination of the true score t and the measurement error e. The basic equation of the classical test theory is

$$x = t + e. \tag{1}$$

Two reasonable assumptions are generally made. The measurement error, in the long term, is zero and the measurement error and the true score are not correlated. That is,

$$E(e) = 0 \tag{2}$$

and

$$\mathrm{cov}(t, e) = 0, \tag{3}$$

where $E(\cdot)$ is the expected value operator. It can be shown that the variance of the observed score σ_x^2 can be decomposed into the variance of the true score σ_t^2 and the variance of the measurement error σ_e^2,

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2. \tag{4}$$

The reliability coefficient of the measurement x, $\rho_{xx'}$, where x and x' are two parallel tests, is defined as the ratio of true score variance to the observed score variance. That is,

$$\rho_{xx'} = \frac{\sigma_t^2}{\sigma_x^2}. \tag{5}$$

There are several interpretations of the reliability coefficient. One direct interpretation, as indicated in Eq. (5), is the ratio of the true score variance to the observed score variance. When the reliability coefficient of a test score is high in a particular sample, the suggestion is that a large portion of the variation of observed scores is due to the true score rather than to the measurement error. A second interpretation of the reliability coefficient is the correlation between two parallel tests or items. Thus, the higher the reliability coefficient, the higher the correlation between two parallel tests.

Reliability of Composite Scores

The single-score reliability model is based on a single item x. It is generally accepted that a single item alone is not sufficient to measure constructs in social sciences. Constructs in social sciences are complicated by nature. To measure or define a construct properly, several items are usually required. In addition, the reliability of a single item measurement is usually too low for applied research; a high degree of reliability is a necessary (though not singularly sufficient) condition for a high degree of validity for a measurement. Therefore, social science researchers prefer measurements that include multiple items that can capture constructs comprehensively and reliably.

Suppose that there are several (say p) items $\mathbf{x} = (x_1, x_2, \ldots, x_p)$, where $\mathbf{x}$ is a $p \times 1$ observed vector taken to be distinct indicators of a theoretical variable of interest; the classical test theory for the multivariate case is

$$\mathbf{x} = \mathbf{t} + \mathbf{e}, \tag{6}$$

where $\mathbf{t}$ and $\mathbf{e}$ are the $p \times 1$ true scores and the $p \times 1$ measurement errors. Following Eqs. (2) and (3), it is also assumed that the expected value of the measurement error $\mathbf{e}$ is a $p \times 1$ null vector $\mathbf{0}$ and the variance–covariance matrix between $\mathbf{t}$ and $\mathbf{e}$ is a null matrix $\mathbf{0}$,

$$E(\mathbf{e}) = \mathbf{0}, \tag{7}$$

and

$$\mathrm{cov}(\mathbf{t}, \mathbf{e}') = \mathbf{0}. \tag{8}$$

Then the variance–covariance matrix of $\mathbf{x}$ can be decomposed into two parts,

$$\Sigma_x = \Sigma_t + \Sigma_e, \tag{9}$$

where Σ denotes the $p \times p$ variance–covariance matrix. Two important things are worth mentioning here. First, measurement errors are usually and reasonably assumed to be uncorrelated with each other. Thus, Σ_e is a diagonal matrix in which the diagonals represent the variances of the measurement error for $\mathbf{x}$, whereas the off-diagonals (covariances of the measurement errors) are all zeros. Second, the off-diagonals of Σ_x and Σ_t are exactly the same, because Σ_e is a diagonal matrix. This means that only the observed variances in Σ_x are inflated by the measurement error, whereas the covariances in Σ_x are not inflated by measurement error.

Now consider the composite score y as a linear combination of $\mathbf{x}$ with any nonnull $p \times 1$ vector of weights $\mathbf{w}$,

$$y = \mathbf{w}'\mathbf{x}. \tag{10}$$

Then the variance of the composite y can be expressed as a sum of true score variance and measurement error variance by

$$\begin{aligned} \mathrm{var}(y) &= \mathbf{w}'\Sigma_x\mathbf{w} \\ &= \mathbf{w}'\Sigma_t\mathbf{w} + \mathbf{w}'\Sigma_e\mathbf{w}. \end{aligned} \tag{11}$$

Similar to the definition of the reliability for a single score, the reliability coefficient $\rho_{yy'}$ for the composite score y can be defined as

$$\begin{aligned} \rho_{yy'} &= \frac{\sigma_t^2}{\sigma_y^2} \\ &= \frac{\mathbf{w}'\Sigma_t\mathbf{w}}{\mathbf{w}'\Sigma_y\mathbf{w}} \\ &= \frac{\mathbf{w}'\left(\Sigma_y - \Sigma_e\right)\mathbf{w}}{\mathbf{w}'\Sigma_y\mathbf{w}} \\ &= 1 - \frac{\mathbf{w}'\Sigma_e\mathbf{w}}{\mathbf{w}'\Sigma_y\mathbf{w}}. \end{aligned} \tag{12}$$

Because Σ_t and Σ_e are both latent (unobserved) scores, different estimates may have different reliability coefficients.

Alpha Coefficient

One of the most popular estimates of the reliability coefficient is Cronbach's alpha coefficient proposed. The alpha coefficient is used to estimate the internal consistency of a composite score. Estimating the reliability requires estimating a diagonal matrix represents the variances of measurement error. Assuming the items are essentially tau-equivalent (i.e., the difference between any two true scores is a constant only), internal-consistency approaches operationalize Σ_e as

$$\Sigma_e = \mathbf{D} - \bar{c}\mathbf{I}, \tag{13}$$

where $\mathbf{D}$ is the $p \times p$ diagonal matrix taken the diagonal elements of Σ_y, $\bar{c}$ is the average item covariance of the off-diagonal elements of Σ_y, and $\mathbf{I}$ is the $p \times p$ identity matrix.

The general idea is that the matrix $\bar{c}\mathbf{I}$ will be close to the expected variances of $\mathbf{x}$ when there is no measurement error. Thus, the differences between the observed variances $\mathbf{D}$ and the expected variances without measurement error $\bar{c}\mathbf{I}$ are used as the estimates for the elements in Σ_e. Following this rationale, Eq. (12) can be operationalized as follows (which is also equivalent to the general form of the alpha coefficient):

$$\alpha = \frac{\mathbf{w}'[\Sigma_y - (\mathbf{D} - \bar{c}\mathbf{I})]\mathbf{w}}{\mathbf{w}'\Sigma_y\mathbf{w}}. \tag{14}$$

It can be shown, with some calculations, that the alpha coefficient can be simplified to

$$\alpha = \frac{p}{p-1}\left(1 - \frac{\mathbf{w}'\mathbf{D}\mathbf{w}}{\mathbf{w}'\Sigma_y\mathbf{w}}\right). \tag{15}$$

If a composite score is formed by an unweighted method, that is, $\mathbf{w}$ is taken to be a $p \times 1$ unit vector of 1s, the alpha coefficient can be reduced to the familiar form,

$$\alpha = \frac{p}{p-1}\left\{1 - \sum_{i=1}^{p}\frac{[\mathrm{var}(x_i)]}{\mathrm{var}(y)}\right\}. \tag{16}$$

It is well known that the alpha coefficient equals the reliability coefficient only when all the items are essentially tau-equivalent. In instances when essentially tau-equivalence is not established, the alpha coefficient sets a lower bound for the reliability coefficient.

Theta Reliability

The alpha coefficient has several unrealistic assumptions imposed on the items. Several researchers have tried to release some assumptions required by the alpha coefficient (for instance, the maximum alpha proposed by Peter M. Bentler and the theta coefficient proposed by Armor). Armor proposed the theta coefficient as an alternative index for the internal consistency of a composite score, whereas the alpha coefficient assumes items are essentially tau-equivalent, meaning that true scores of all items are different only by a constant. Therefore, equal weightings are used to calculate the composite score and to estimate the internal consistency.

Practically speaking, the assumption of essentially tau-equivalence is difficult to satisfy. First, items may measure more than one single construct. Second, they may measure one single construct differently by a different proportion of true score variances, even if they are measuring only one single construct. In other words, items contribute different proportions of true score variances to the composite score when items are not essentially tau-equivalent. Thus, using unit weights for the alpha coefficient is not the optimal choice in this situation.

Formula

As indicated in Eq. (15), the estimated internal consistency depends on the chosen weighting vector $\mathbf{w}$, given the same set of data. A unit vector of 1s is chosen as the weighting vector in the alpha coefficient, whereas Armor proposed to estimate the theta coefficient by finding an optimal weighting vector $\mathbf{w}$ where the alpha coefficient is the maximum. From Eq. (15), the alpha coefficient will be at a maximum when $\mathbf{w}'\mathbf{D}\mathbf{w}/\mathbf{w}'\Sigma_y\mathbf{w}$ is at a minimum, or $\lambda = \mathbf{w}'\Sigma_y\mathbf{w}/\mathbf{w}'\mathbf{D}\mathbf{w}$ is a maximum. Thus, if the maximum of λ is found, a maximized alpha coefficient can also be obtained directly. It is usually more convenient to define a $p \times 1$ vector $\mathbf{u} = \mathbf{D}^{1/2}\mathbf{w}$, or equivalently, $\mathbf{w} = \mathbf{D}^{-1/2}\mathbf{u}$, such that λ can be rewritten as

$$\hat{\lambda} = \frac{\mathbf{u}'\mathbf{D}^{-1/2}\Sigma_y\mathbf{D}^{-1/2}\mathbf{u}}{\mathbf{u}'\mathbf{u}}. \tag{17}$$

It is then recognized that $\mathbf{D}^{-1/2}\Sigma_y\mathbf{D}^{-1/2}$ is the correlation matrix (say $\mathbf{R}$) of $\mathbf{x}$. Thus, λ can be simplified as

$$\lambda = \frac{\mathbf{u}'\mathbf{R}\mathbf{u}}{\mathbf{u}'\mathbf{u}}. \tag{18}$$

To maximize λ, it can be differentiated with respect to $\mathbf{u}$ to give

$$\frac{\partial\lambda}{2\,\partial\mathbf{u}} = \frac{\mathbf{R}\mathbf{u} - \lambda\mathbf{u}}{\mathbf{u}'\mathbf{u}}. \tag{19}$$

Setting the derivative to zero yields

$$(\mathbf{R} - \lambda\mathbf{I})\mathbf{u} = \mathbf{0}. \tag{20}$$

Then maximizing λ is equivalent to finding the eigenvalues of the correlation matrix $\mathbf{R}$. Because $\mathbf{u}$ is a nonnull vector, this suggests that

$$|\mathbf{R} - \lambda\mathbf{I}| = 0, \tag{21}$$

where $|\cdot|$ denotes the determinant of a matrix. There is a well-known solution for this problem in the context of principal component analysis. The solution for λ in Eq. (21) equals the largest root (say λ_1) in the principal component analysis. The theta coefficient can then be expressed as

$$\theta = \frac{p}{p-1}\left(1 - \frac{1}{\lambda_1}\right), \tag{22}$$

with

$$\mathbf{w} = \mathbf{D}^{-1/2}\mathbf{u}, \tag{23}$$

where $\mathbf{u}$ is the eigenvector of $\mathbf{R}$ corresponding to λ_1.

Relationship to Principal Component Analysis

From the preceding formulas of the eigenvalue equation, the theta coefficient is directly related to principal component analysis. Indeed, it can be shown that the proportion of variance explained by the first principal component equals $1/[p - \theta(p-1)]$. Thus, the results of the principal component analysis are generally used to estimate λ_1 and its corresponding eigenvector $\mathbf{u}$ to calculate the theta coefficient and its corresponding $\mathbf{w}$ for creating the composite score. It is important to note that although principal component analysis and theta reliability are closely related, they serve different purposes and interpretations. Theta reliability measures the internal consistency (interrelatedness) for a set of items whereas principal component analysis measures the dimensionality of a set of items. Items with high internal consistency do not necessarily imply that they are unidimensional.

Advantages

Armor suggested two conditions in which the theta coefficient is more useful than the alpha coefficient in estimating internal consistency. The first condition is when items measure two or more independent constructs either equally or unequally. The second condition is when the items measure a single construct, but do so unequally. For the first condition, the theta coefficient does not require all items to be unidimensional. Because theta reliability is related to the largest eigenvalue and its corresponding eigenvector, items not related to the primary dimension (the one with the largest eigenvalue) are downweighted with a smaller eigenvector. Thus, the effects of items not related to the first dimension will be minimized in the composite score by using smaller weightings. Second, when items are not loaded equally on the true scores, theta reliability uses optimal weights to create the composite score and to estimate its reliability. Therefore, the theta coefficient is always larger than (or at least equal to) the alpha coefficient. Because the estimate of the theta coefficient is based on the correlation matrix of these items, it is also invariant to the scales of the items. In other words, changing the scales of the items has no effect on the theta coefficient.

Procedures and Illustrations

To illustrate the procedures for calculating theta reliability and contrasting it with the alpha coefficient, a real data set on the Center for Epidemiologic Studies Depression Scale (CES-D) conducted in 2001 by the Family Planning Association of Hong Kong was used. The data were collected from 2864 Secondary 3 (Grade 10) to Secondary 7 Hong Kong Chinese students. The scale consisted of 20 5-point Likert scale items measuring a unidimensional factor. The descriptive statistics are shown in Table I. The alpha coefficient is 0.90, which suggests that the items, on average, are reasonably interrelated. As suggested by many researchers, the values of the alpha coefficient are affected by two factors: (1) the average interitem correlation and (2) the number of items. Consider, for example, the alpha coefficient for a set of weakly interrelated items can still be very high when there are numerous items in the scale. Moreover, the alpha coefficient may not be good enough to indicate which items are problematic (not related to other items).

By conducting a principal component analysis, the largest eigenvalue λ_1 is 8.22. Substituting λ_1 into Eq. (22), the estimated theta coefficient is 0.92. The principal component loadings, also known as factor loadings, and the weightings for creating the composite score with the theta coefficient are shown in Table II. The optimal weights for the theta reliability are created by the principal component loadings divided by their corresponding standard deviations. All of the principal component loadings are reasonably high except for the recoded items. As shown in Table II, the recoded items (x_4, x_8, x_{12}, and x_{16}) appear to be problematic. To adjust for this, these items were downweighted in the composite score.

Apart from using the theta coefficient, an attempt can be made to delete the problematic items and recalculate the alpha coefficient. After deleting the four problematic items, the alpha coefficient is 0.93. The largest root, λ_1, after deleting the problematic items is 8.12 and the theta coefficient is 0.94. In this case, the alpha coefficient and the theta coefficient are quite similar.

Conclusions

Theta reliability provides an alternative measure of internal consistency for a set of items and has been shown to be a good index for measuring internal consistency. For instance, it is less influenced by the violation

Table I Descriptive Statistics of the Center for Epidemiologic Studies Depression Scale[a]

	Item																			
Item	x_1	x_2	x_3	x_4^*	x_5	x_6	x_7	x_8^*	x_9	x_{10}	x_{11}	x_{12}^*	x_{13}	x_{14}	x_{15}	x_{16}^*	x_{17}	x_{18}	x_{19}	x_{20}
x_1	1.00																			
x_2	0.45	1.00																		
x_3	0.53	0.48	1.00																	
x_4^*	0.19	0.18	0.15	1.00																
x_5	0.40	0.31	0.47	0.24	1.00															
x_6	0.53	0.43	0.66	0.16	0.54	1.00														
x_7	0.48	0.36	0.52	0.18	0.56	0.61	1.00													
x_8^*	0.12	0.08	0.06	0.40	0.13	0.06	0.10	1.00												
x_9	0.40	0.36	0.50	0.03	0.42	0.54	0.56	−0.07	1.00											
x_{10}	0.48	0.39	0.55	0.11	0.42	0.58	0.55	0.03	0.59	1.00										
x_{11}	0.37	0.41	0.44	0.17	0.34	0.45	0.41	0.04	0.43	0.51	1.00									
x_{12}^*	−0.04	−0.01	−0.15	0.32	0.03	−0.17	−0.06	0.41	−0.20	−0.13	−0.13	1.00								
x_{13}	0.30	0.28	0.37	0.15	0.33	0.42	0.34	0.07	0.37	0.36	0.33	−0.08	1.00							
x_{14}	0.43	0.35	0.57	0.12	0.38	0.60	0.48	0.02	0.53	0.56	0.42	−0.18	0.53	1.00						
x_{15}	0.38	0.31	0.42	0.18	0.36	0.47	0.43	0.07	0.44	0.47	0.40	−0.11	0.42	0.52	1.00					
x_{16}^*	0.01	0.01	−0.11	0.33	0.02	−0.13	−0.05	0.48	−0.18	−0.11	−0.06	0.69	−0.05	−0.14	−0.10	1.00				
x_{17}	0.41	0.40	0.44	0.06	0.26	0.43	0.33	0.00	0.41	0.48	0.37	−0.11	0.27	0.41	0.35	−0.05	1.00			
x_{18}	0.47	0.39	0.59	0.10	0.43	0.68	0.52	0.02	0.54	0.59	0.45	−0.24	0.43	0.65	0.53	−0.18	0.52	1.00		
x_{19}	0.41	0.33	0.51	0.10	0.43	0.58	0.50	0.06	0.55	0.53	0.42	−0.12	0.41	0.62	0.60	−0.10	0.40	0.63	1.00	
x_{20}	0.45	0.37	0.54	0.14	0.57	0.62	0.60	0.03	0.55	0.58	0.44	−0.12	0.41	0.57	0.49	−0.12	0.39	0.63	0.60	1.00
SD	0.88	0.87	0.94	1.00	1.01	0.96	0.91	1.05	0.96	0.92	0.97	1.04	0.97	0.99	0.89	1.04	0.96	0.98	0.92	0.98
Mean	2.41	2.21	2.37	2.91	2.76	2.51	2.49	3.22	2.17	2.22	2.18	3.40	2.36	2.33	2.28	3.25	1.97	2.35	2.32	2.45

[a] $N = 2864$; asterisks denote negatively worded items that were recoded here.

Table II Optimal Weightings for the Theta Coefficient

Items	*Principal component loadings*	*SD*	*Optimal weights*
x_1	0.66	0.88	0.75
x_2	0.58	0.87	0.66
x_3	0.76	0.94	0.81
x_4	0.20	1.00	0.20
x_5	0.64	1.01	0.63
x_6	0.82	0.96	0.86
x_7	0.74	0.91	0.81
x_8	0.07	1.05	0.06
x_9	0.73	0.96	0.76
x_{10}	0.77	0.92	0.84
x_{11}	0.63	0.97	0.65
x_{12}	−0.18	1.04	−0.18
x_{13}	0.57	0.97	0.59
x_{14}	0.77	0.99	0.78
x_{15}	0.68	0.89	0.77
x_{16}	−0.15	1.04	−0.14
x_{17}	0.60	0.96	0.62
x_{18}	0.81	0.98	0.83
x_{19}	0.76	0.92	0.83
x_{20}	0.79	0.98	0.81

of unidimensionality and unequal loadings of the true scores on the items. It is also scale invariant, meaning that it is not affected by scale changes. It is not difficult for applied researchers to estimate theta reliability and to calculate the optimal weights by using results from principal component analysis. Practically speaking, the alpha coefficient is generally close to the theta coefficient if the items are reasonably good. The discrepancy between them is large only when there is more than one independent dimension and/or when the items are loaded differently on the true scores.

Acknowledgments

The authors thank the two referees for their comments. The work is supported by the Hong Kong Jockey Club Charities Trusts.

See Also the Following Articles

Alpha Reliability • Reliability

Further Reading

Armor, D. J. (1974). Theta reliability and factor scaling. In *Sociological Methodology*, Vol. 5 (H. L. Costner, ed.), pp. 17–50. Jossey-Bass, San Francisco, CA.

Bentler, P. M. (1968). Alpha-maximized factor analysis (alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika* **33,** 335–345.

Carmines, E. G., and Zeller, R. A. (1979). *Reliability and Validity Assessment.* Sage Publ., Beverly Hills, CA.

Cronbach, L. J. (1951). Coefficient alpha and the internal consistency of tests. *Psychometrika* **16,** 297–334.

Greene, V. L., and Carmines, E. G. (1980). Assessing the reliability of linear composites. In *Sociological Methodology*, Vol. 11 (H. L. Costner, ed.), pp. 160–175. Jossey-Bass, San Francisco, CA.

Thorndike, Edward L.

Richard E. Mayer
University of California, Santa Barbara, California, USA

Glossary

achievement test An instrument intended to measure what has been learned, such as arithmetic computation skills.
aptitude test An instrument intended to measure learning potential or capability to learn; college entrance examinations, for example, measure aptitudes.
Army Alpha A standardized job placement test used by the U.S. Army in World War I; was the largest mass testing effort in human history up to that time.
law of effect A principle of learning proposed by E. L. Thorndike stating that responses that are followed by satisfaction become more strongly associated with the situation and responses that are followed by dissatisfaction become less strongly associated with the situation.
learning curve The quantitative functional relationship between a measure of experience (such as number of practice sessions) and a measure of amount learned (such as time to perform a task).

Edward Lee Thorndike (1874–1949), the world's first educational psychologist, was influential in shaping the field to include educational measurement. Thorndike viewed rigorous quantitative measurement as the key to turning educational psychology (and other social sciences) into scientific enterprises. "Whatever exists at all exists in some amount." This quote from E. L. Thorndike in 1918 epitomizes his unwavering faith in quantitative measurement, and is the premise underlying many of his contributions to educational and social measurement. Thorndike contributed to improving educational methods by measuring learning outcomes, improving college admissions and personnel selection by measuring human characteristics, improving school dictionaries by measuring word frequencies, and improving communities by measuring quality of life.

The Life of E. L. Thorndike

Edward Lee Thorndike was born on August 31, 1874 in Williamsburg, Massachusetts, and remained in New England until his final year of graduate study in 1897. His childhood was shaped by time (i.e., growing up in the late 1800s), place (i.e., living in the uncomplicated and self-sufficient communities of New England), and family (i.e., being the son of a Methodist minister who moved from congregation to congregation). Thorndike's most prolific biographer, Geraldine Joncich, notes that growing up in a "clergyman's household, combined with a New England setting, was the best predictor of a future career in science" for those of Thorndike's generation.

Thorndike's education did not at first concentrate on psychology. He received a B.A. degree from Wesleyan University in 1895, where he took some psychology courses but, according to Joncich, "never publicly committed himself to psychology." Late in his undergraduate career, Thorndike read William James' *Principles of Psychology*, which Thorndike said had been more stimulating than any book he had previously read. Thorndike moved to Harvard for graduate study, where he hoped to take a course from James, but he still listed English and French as his course of study. Within his first year, he dropped English and French in favor of psychology, and became William James' doctoral student. After the Harvard administration refused to allow him to study children, Thorndike selected chickens—and eventually cats and dogs—as the focus of his soon-to-be-classic research on learning. When William James was unable to secure lab space on campus for Thorndike's research, the project was moved to the attic of the James family house. Thorndike has written that the "nuisance" (to Mrs. James) of his presence was "somewhat mitigated by the entertainment to the two youngest children." A final challenge to Thorndike's work was that his advisor, William James,

had given up on conducting experimental research in psychology and was moving back to philosophy.

After receiving a Master's degree from Harvard in 1897, Thorndike moved to Columbia University; continuing the research he began at Harvard, he received a Ph.D. from Columbia in 1898 under the sponsorship of James McKeen Cattell. His thesis, later published as *Animal Intelligence*, revolutionized the field of learning but did not interest his teachers. In the thesis, he articulated the law of effect, which was to become one of psychology's most important principles. Following a year of teaching education students at the College for Women of Western Reserve University in Cleveland, Thorndike accepted a faculty position at Teachers College, Columbia University in 1899. Thorndike remained at Teachers College for the next 50 years. Although he retired in 1940, he remained active in emeritus status until his death in 1949. During his career, Thorndike produced more than 250,000 pages of original writing in more than 500 publications. He served as President of the American Psychological Association in 1912 and as President of the American Association for the Advancement of Science in 1934. His incessant research studies revolutionized the fields of learning, transfer, and individual differences. Most importantly, he established the field of educational psychology as a science, promoting the scientific method for educational research and demonstrating its application in educational practices.

Measurement to Improve Behavioral Research: The Learning Curve

"Thorndike's lifelong preoccupation with measurement was first expressed," wrote Joncich, by the learning curves of his chickens, cats, and dogs reported in his landmark book, *Animal Intelligence*. In a typical study, a hungry cat (e.g., a 4- to 6-month old called "No. 12") was placed in a puzzle box and could escape to food outside by clawing at a loop of string that would open a small door. Over the course of many sessions, Thorndike placed the cat in the puzzle box and measured the time needed to get out. During the first session, the cat required 160 seconds; the cat took 30 and 90 seconds during the second and third sessions, respectively, and was down to 7 seconds by the 24th (and final) session. Figure 1 shows a learning curve for cat number 12; the curve represents the quantitative relationship between a measure of learning on the *y* axis (i.e., time needed to escape) and a measure of training on the *x* axis (i.e., the number of sessions). Thorndike summarized his time curves as curves for which the "lengths of one millimeter along the abscissa represent successive experiences in the box, and heights of one millimeter above it each represent ten seconds of time." Thus, Thorndike joined Hermann Ebbinghaus (who wrote *Memory* in 1885), being among the first to demonstrate a quantitative relation between amount of training and amount of learning. It is interesting to note that early training had a much stronger effect than did later training—that is, the learning curve is steep at first and then flattens out. Thorndike's meticulous measurements resulted in this characteristic shape of the learning curve that was to be replicated countless times over the ensuing century.

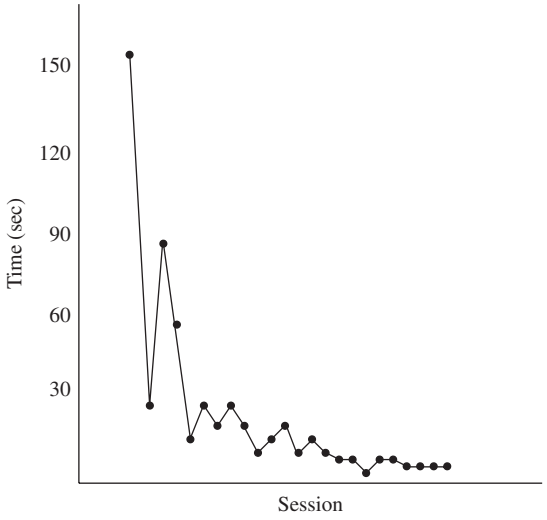


Figure 1 Thorndike's learning curve, showing the quantitative relationship between amount of experience (i.e., quantified as session number) and learning (i.e., quantified as time needed to escape from the puzzle box).

The precision of Thorndike's measurements helped reorient the field of animal learning from the realm of pseudoscientific speculations based on vague qualitative observations to an experimental science based on quantitative measurement. The precision of Thorndike's learning curves allowed him to propose the law of effect, which for more than a century has been a central pillar in psychology:

> *Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation so that when it recurs, they will be more likely to recur; those which are accompanied or followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.*
>
> [Thorndike, *Animal Intelligence*, 1911]

Most importantly, Thorndike's learning curves helped establish the fact that it was possible to quantify learning.

By measuring changes in learners due to their experience, Thorndike contributed to a breakthrough that had tremendous implications for education.

Measurement to Improve Educational Methods: Achievement Tests of Learning Outcomes

The most fundamental challenge facing Thorndike in his efforts to shape the new science of educational psychology was how to promote learning in students. He quickly realized that the improvement of education depended on appropriate quantitative measures of learning outcomes. In short, in order to determine whether a particular instructional method worked better than another did, it would be necessary to measure what was learned. Thus, educational measurement was at the center of Thorndike's vision of educational science.

In his first educational psychology textbook, *Educational Psychology*, published in 1903, Thorndike laid out the role of quantitative measurement as a vehicle for improving education: "The work of education is to make changes in human minds and bodies. To control these changes we need knowledge of the causes which bring them to pass. Such knowledge necessitates some means of measuring mental and bodily conditions; adequate knowledge necessitates accurate and complete measurements. . . . " Further, he recognized that the improvement of education depended on the use of high-quality measurements, writing in the textbook that "Commonly our measurements of mental conditions and so of the changes due to any educational endeavor are crude, individual, and incomplete . . . An adequate measurement of mental traits will be one that is precise enough for us to draw the conclusions we desire; objective or subject to individual repetition by another observer, and complete enough to take in all the features of the condition that are important for our purpose." Thorndike thus called for the creation of tests for every school subject, measuring arithmetic skills, music appreciation, English language writing abilities, and even skills of dexterity, such as the ability to use woodworking tools. To accomplish this goal, Thorndike often had to adapt new forms of measurement to academic tasks. For example, how might the handwriting of elementary school students be measured? Thorndike's solution was to have several trained raters judge the quality of handwriting using a 20-point rubric consisting of examples of each level. In creating such scales, Thorndike recommended providing "specimens at each level of goodness," ensuring that "differences between each level and the next are known with some exactitude," and making sure that "the scale extends down to a true zero." Foreshadowing more recent developments in the measurement of reliability, Thorndike also called for "measurement by a consensus of judges" in which "it is sound practice to correlate the rating by half of the judges with the rating by the other half and to increase the number of judges, if necessary, to obtain a half-with-half correlation of at least 0.90." Overall, Thorndike's advice on how to measure learning outcomes still makes sense more than a century later.

Thorndike's prescription for educational improvement was to conduct what today might be called clinical trials. In Thorndike's 1903 textbook, he wrote that "We should be able to state exactly the difference between any two human beings, between the condition of any one before and after any course of study or other educational influence; we could compare the results of different systems of education, describe the changes. . . . In the instance just quoted, A could say: 'The 600 children in my school under the old method made an average gain of 4 per cent in a year in arithmetic knowledge. . . . Under the new method the [gain is] 6 percent.'" For Thorndike, improving the effectiveness of educational methods was intimately linked to the creation of useful measures of learning outcomes.

Throughout his career, Thorndike was concerned with precise mental measurement in schools; this was exemplified in 1904 by his now classic book, *An Introduction to the Theory of Mental and Social Measurement*. One important result of Thorndike's call for quantified measures of student learning outcomes was a series of standardized tests and scales for school-related performance. These tests and scales included quantitative measures of English, composition, drawing, handwriting, reading ability, and arithmetic ability. Although the specific tests are no longer in wide use, measuring student learning has become a standard practice in education and training. Overall, Thorndike's relentless efforts to measure the learning outcomes of students resonates well with current calls for accountability based on educational standards.

Measurement to Improve Academic Admissions and Personal Selection: Aptitude Tests of Human Characteristics

Although Thorndike created or instigated many achievement tests (i.e., tests intended to measure what has been learned), he also was involved in the creation of many aptitude tests (i.e., tests intended to measure potential to learn or to perform some task). For example, suppose it is necessary to determine which college applicants should be admitted to Columbia University or which Army recruits were best suited for which jobs in the Army.

During his career, Thorndike was called on repeatedly to devise tests for college admissions and job placement.

In 1925, Thorndike devised an examination to be used for college entrance screening. To ensure proper sampling of relevant cognitive skills, he included four different types of cognitive tasks most likely to measure academic intelligence: sentence completion, arithmetic, vocabulary, and directions (the exam was thus named the CAVD exam). To ensure reliability, he devised many statistically parallel forms for each of the subtests. To insure validity, he coupled intelligence items and school content to prevent high scores by bright but poorly prepared students. The CAVD exam was not used as extensively as Thorndike had expected, partly because of the 3 hours required to take it and partly because Thorndike chose not to call it a test of general intelligence. Yet, the CAVD exam is an excellent example of Thorndike's inventory approach to mental testing—the test focused on four well-defined cognitive skills and contained a representative inventory of items measuring each of them. In addition to his contributions to general college admissions, Thorndike developed entrance tests for professional schools for engineering students and law students. He considered standardized tests, as opposed to a variety of entrance examinations, to be prognostic of future success, rather than being measures of previous educational opportunities.

In 1917, Thorndike was recruited into the Committee on Classification of Personnel for the U.S. Army in World War I; the committee was charged with the task of determining appropriate job classifications for the overwhelming number of soldiers called into action. The result was the creation and implementation of the Army Alpha, which, at the time, was the largest mass testing effort in human history. Thorndike was also instrumental in creating the Army Beta for people not literate in English. The Thorndike biographer G. J. Clifford, in the 1984 book *Edward L. Thorndike: The Sane Positivist*, noted that "by the spring of 1917 a small but active testing movement is evident with Thorndike near its center." Thorndike had already devised personnel selection tests for industry leaders, including businessmen for the American Tobacco Company. Joncich noted in his 1968 piece in the *American Psychologist* that "modern personnel divisions in industry may be dated from the time that Metropolitan vice-president Dr. Lee K. Frankel approached Thorndike to request a new kind of examination." In 1921, Thorndike and two of his colleagues, Cattell and Woodworth, established the Psychological Corporation to foster the development of tests and other kinds of measurements useful to business and industry. The goal was to apply the methods and principles of psychological science.

Are there any basic principles for designing aptitude tests? In one of his last books, *Human Nature and the Social Order*, published in 1940, Thorndike summarized his approach to the measurement of mental abilities: "There are two simple golden rules: Measure all of the ability. Measure nothing but it." Concerning the first rule, Thorndike explained that mental measurement involves taking a representative sample of the target skills: "To measure all of it does not, however, require measuring every item of it, but only that the sample be large enough and well-proportioned enough to give the same result that would be had if every item had been measured. For example, if the ability is knowledge of the meanings of English words (excluding proper names) . . . a test with even only a thousand will measure accurately enough for most purposes." Concerning the second rule, Thorndike recognized that, even though it was desirable, "to measure nothing but it" did not require obtaining a perfectly pure sample, free from all contamination by other abilities. If pure samples are unobtainable or obtainable only at enormous cost of time and effort, it is possible to manage the situation by determining the amount of contamination and allowing for it. In summary, Thorndike wrote that "measuring a human ability is usually more like taking an inventory than using a tape, or balance, or thermometer."

Is it possible to describe the character of person using numbers? For Thorndike, the answer was a resounding "yes." In his 1911 book *Individuality*, Thorndike argued that "All intelligible differences are ultimately quantitative. The difference between any two individuals, if describable at all, is described by comparing the amounts which A possesses of various traits with the amounts which B possesses of the same traits. . . . If we could list all the traits, each representing some one characteristic of human nature, and measure the amount of each of them possessed by a man, we could represent his nature—read his character—in a great equation." Thus, Thorndike was a strong proponent of the factor theory of human ability, namely, the idea that people differ along a number of dimensions. He rejected the idea that people should be classified into types: "The customary view has been that types, or particular combinations of amounts of human traits, could be found so that any individual would be much like some type and much less like any of the others. But no one has succeeded in finding such types." Thus, "there is much reason to believe that human individualities do not represent ten or a hundred or a thousand types, but either one single type or as many types as there are individuals". Overall, Thorndike's view of ability as a collection of small skills is consistent with modern views.

Measurement to Improve Educational Materials: Dictionaries for Each Age Level

How can quantitative measurement be used to improve school materials? Consider the case of school dictionaries.

Thorndike was the first to design school dictionaries based on empirical quantitative measurement. Because dictionaries at each age level could contain only a limited number of words, those words should be words that are most commonly found in normal usage. To determine the most commonly used words, Thorndike collected a corpus of 10 million running words sampled from 279 common publications such as newspapers, school books, adult literature, and even the Bible. He counted how many times each word occurred in these publications, and for each of the 20,000 most common words he determined the average number of occurrences per million words of printed text. On the basis of his word frequency counts, he was able to exclude words and technical terms not used by school children.

Thorndike then determined the specificity of his definitions based on the frequency count of the to-be-defined word. Simple words with high frequencies, such as "spoon" or "little," should be defined as simply as possible, because young readers would be most likely to look up the simple words. Difficult words with low frequencies, such as "factitious" or "feminine," could be defined more formally, because older readers would be more likely to look these up. Thus, Thorndike constructed his definitions so that the words used in definitions were as common as or more common (i.e., had the same or higher word frequency counts) than the word being defined. The following definition of "cow" violates Thorndike's principle: "The mature female or any bovine animal, or any other animal the male of which is called bull." Thorndike would replace that definition with one using terms at least as common as "cow," i.e., "(1) The large animal that furnishes us with milk, butter, and cheese. (2) The female of various animals, such as a buffalo cow, an elephant cow, and a whale cow." To make definitions clear, Thorndike included an illustrative sentence. For example, the definition of "facilitate" was "make easy; lessen the labor of; forward; assist. A vacuum cleaner facilitates housework." When a word had multiple definitions, Thorndike's dictionaries grouped similar meanings together and ordered them based on frequency of usage. For example, the most common usage of "club" is as "a stick," so this definition would be first; a less common usage of is as "a group of members," so this definition would come later. For example, the definition of "club" would be "(1) A heavy stick of wood, thicker at one end, used as a weapon. (2) Beat with a club or something similar. (3) A stick or bat for some games played with a ball, such as golf clubs. (4) A group of people joined for some special purpose, such as a social club, tennis club, yacht club, and nature-study club. (5) The building or rooms used by a club. (6) Join together for some purpose. The children clubbed together to buy their mother a plant for her birthday."

Thorndike's data-based strategy helped prune and organize definitions. For example, the *Oxford English Dictionary* available to Thorndike listed five definitions of "amenable." Thorndike would cut this list to the two definitions that actually had occurred in his frequency counts—"open to advice" (which was most common) and "accountable" (which was less common). Thorndike's dictionaries began with the *Junior* version in 1935, followed by *Senior* in 1941, *Revised Junior* in 1942, and *Beginning* in 1945. These volumes instantly became the most widely used student dictionaries in the United States, setting the standard for all subsequent student dictionaries. By applying his skills in quantitative measurement, Thorndike became one of the most important lexicographers of the era. In addition to his creation of curricular materials in language arts, he created arithmetic books (nicknamed *Thorndike Arithmetics*) that become so successful that they provided more income than his salary from Columbia.

Measurement to Improve Society: Ranking the Quality of Life of U.S. Cities

In his later years, E. L. Thorndike turned his attention to social issues such as the question of how to improve communities. This issue led Thorndike to the measurement challenge of determining what makes a good community. Consistent with his lifelong record of inventive applications of quantitative measurement, Thorndike sought ways to quantify what he called the "goodness" of all U.S. cities with populations above 30,000 in 1930. In 1939, he published *Your City*, in which he measured 310 U.S. cities on 297 quantitative dimensions, ranging from the infant death rate, to the per capita expenditures for schools, to the per capita number of telephones. Later, he published a follow-up study that examined the goodness of smaller cities. Foreshadowing similar rankings that are commonplace today, ranging from retirement communities to graduate schools, Thorndike demonstrated how it is possible to use a single number to express the quality of life of each city.

In order to create an index of the "General Goodness of Life" (which he labeled the "G score"), Thorndike selected a weighted collection of 37 measures. The measures focused on health (e.g., the infant death rate and the general death rate), education (e.g., per capita public expenditures for schools, percentage of persons 16 to 17 attending schools, and average salary of a high school teacher), recreation (e.g., per capita public expenditures for recreation), economic and social items (e.g., percentage of extreme poverty and percentage of working boys or girls ages 10 to 14), creature comforts (e.g., per capita domestic installations of telephones and per capita domestic installations of electricity), and other items (e.g., ratio of value of schools to value of jails and per capita

circulation of *Better Homes and Gardens*, *Good Housekeeping*, and the *National Geographic Magazine*). Based on this goodness index, each U.S. city with a population above 30,000 could be assigned a goodness score (or G score) ranging from 0 to above 1000. In 1930, Pasadena came in first, followed by Montclair, Cleveland Heights, Berkeley, Brookline, Evanston, Oak Park, Glendale, Santa Barbara, and White Plains.

Thorndike's task was to determine what makes a city good, and his solution was to use quantitative data and lots of it. In addition to four versions of his 37-item goodness scale, he created an 11-item P-scale of "certain desirable personal qualities" (including measures such as per capita number of high school graduates), a 9-item I-scale of "per capita private income" (including measures such as per capita number of income tax returns of $2500 or more), and a 10-item "City Yardstick" that simulated the G-score but contained items that "anyone can obtain for almost any city in a few hours." Overall, he collected more than a million facts and he seemed to revel in exploring them. For example, he noted that "per capita membership of the Boy Scouts correlates 0.56 with G" whereas "church membership correlates negatively with G." Thorndike's study of the "what makes a city good" reflects his faith in the value of measurement as a tool for improving society. He tells the reader that "Industry and business have found it profitable to use measurement as a major factor in control and improvement . . . Citizens may well do the same." Thorndike recognizes that not everyone shares his faith in numbers: "Certain humanists who abominate all efforts to measure human values, will object to the list of items and to the scores computed from them." For example, critics could argue that the G-score does not include the important things: "Radio sets, free schools, swimming pools and baby clinics cannot atone for bigotry and bad taste. What use are free libraries when people read trash?"

Thorndike's response to such criticism of his attempts to measure goodness of life were twofold: that the G-index is imperfect but "is good as far as it goes," and that "if those personal qualities which the humanist rightly admires and finds neglected by our list could be measured, and the three hundred cities were rated according to them the results would correlate with [the G-score] positively."

Thorndike's quantitative measurement of the quality of life in American cities led to some controversial observations. Commentary ran from "it is good for a city to have few very poor families," to "too much unskilled labor is bad for a city," to "disparity in income does no harm whatsoever." For Thorndike, these were factual statements based on his measurements and statistical analyses. In perhaps his most controversial analysis, he noted a strong negative correlation (−0.60) for the "percentage of Negro families with G," leading him to observe that "the fewer the Negro families, the better the score of the city on G." Yet, Thorndike saw his role as an honest presenter of facts. He stated that "The recital of these facts is in no sense an attack upon the Negro race. It would be no kindness to hide the truth." In interpreting the implications of his data, Thorndike calls for providing better treatment of minorities and full opportunities "to all human beings" as a way of eliminating this correlation, but insists that "the truth about . . . one's group is far better than misleading silence or flattery." Further, he condemned racially segregated ghettos as "wasteful and dangerous, as well as cruel" and stated that "a city should try to improve the personal qualities of all its residents."

Thorndike offers an optimistic vision of how society can use social measurement, such as his G-index, to improve life for everyone. "Any city can improve itself. Not by trying to be bigger . . . Not by building factories, shops, or offices . . . except where they are needed." In reviewing his measurements, Thorndike concluded that a city's quality of life is highly dependent on the education level and personal wealth of its inhabitants: "At least four-fifths of the differences of cities in goodness is caused by the personal qualities of the citizens and the amount of their incomes. These then are the main things to improve . . . attracting good people to their city and earning more money." Thus, improving a city requires improving the educational and economic status of the human beings who reside in it: "A community becomes great and good by giving opportunities to those who crave and deserve them." Except for his peculiar advocacy of eugenics (the "science and art" of breeding humans), much of Thorndike's evidence-based advice for improving cities still seems highly relevant today.

Conclusion

Thorndike was a quantifier. Who but Thorndike would have counted the number of times each word appeared in a large corpus of common printed materials, ranging from newspapers to school books to the Bible, resulting in a list of 10,000 or 20,000 or 30,000 common words and their rates of occurrence per million words? Each time he was confronted with a fundamental educational or social problem, his first step was to count something. He demonstrated that it was possible to assign a number to an amazing array of human activities, including the quality of a student's handwriting, the level of a college applicant's academic skill, the frequency of use of words in common publications, and the goodness of life in American cities. Thorndike's career is replete with many other examples of his quantification of social measures.

Thorndike was a pioneer who shaped the field of educational measurement. In 1901, he established in the United States the first true course in educational measurement. In 1904, he wrote *An Introduction to the Theory of Mental and Social Measurements*, which

G. J. Clifford called "the first complete theoretical exposition and statistical handbook in the new area of social-science measurement." Also, from the start, Thorndike provided a guiding vision for the field that the task of education to alter the human experience cannot be evaluated without means of measurement.

More than a century ago, Thorndike foresaw his legacy as a quantifier. In *Educational Psychology*, he wrote that "In education everything is said but nothing is proved. There is a plentiful lack of knowledge while opinions more and more abound. . . . The science of education when it develops will like other sciences rest upon direct observations of and experiments on the influence of educational institutions and methods made and reported with quantitative precision . . . Long after every statement about mental growth made in this book has been superseded by a truer one . . . the ideals of accuracy and honesty in statistical procedure by which I hope it has been guided will still be honored." Clearly, Thorndike's use of educational and social measurement created a formidable collection of products, including achievement tests, placement tests, dictionaries, and city rankings. Yet his most important contribution rests in demonstrating the inseparable relation between measurement, science, and the solution of social problems. To solve social problems (such as how to teach language arts or how to place job candidates or how to improve cities), Thorndike turned to science, and in applying science to human life, the first step involved measurement. Thus, Thorndike's most important legacy, demonstrated through his daunting list of accomplishments, is the value he placed on quantitative measurement as an indispensable feature of educational and social research. Indeed, his articulation of the role of measurement has an unmistakably contemporary ring. In an article in *Teachers College Record* in 1921, he wrote that "In proportion as it becomes definite and exact, this knowledge of educational products and educational purposes must become quantitative, taking the form of measurements."

See Also the Following Articles

Alpha Reliability • Education, Tests and Measures in

Further Reading

Barnhart, C. L. (1949). Contributions of Dr. Thorndike to lexicography. *Teachers Coll. Rec.* **51,** 35–42.

Clifford, G. J. (1984). *Edward L. Thorndike: The Sane Positivist.* Wesleyan University Press, Middletown, CT.

Joncich, G. (1968). E. L. Thorndike: The psychologist as professional man of science. *Am. Psychol.* **23,** 434–446.

Lorge, I. (1949). Edward L. Thorndike's publications from 1940 to 1949. *Teachers Coll. Rec.* **51,** 42–45.

Mayer, R. (2003). E. L. Thorndike's enduring contributions to educational psychology. In *Educational Psychology: A Century of Contributions* (B. J. Zimmerman and D. H. Schunk, eds.), pp. 113–154. Erlbaum, Mahwah, NJ.

Russell, J. E. (1940). Publications from 1898 to 1940 by E. L. Thorndike. *Teachers Coll. Rec.* **41,** 699–725.

Russell, W. F. (1949). Edward L. Thorndike, 1974–1949. *Teachers Coll. Rec.* **51,** 26–28.

Thorndike, E. L. (1903). *Educational Psychology.* Science Press, New York.

Thorndike, E. L. (1904). *An Introduction to the Theory of Mental and Social Measurements.* Science Press, New York.

Thorndike, E. L. (1911/1965). *Animal Intelligence.* Hafner, New York.

Thorndike, E. L. (1911). *Individuality.* Riverside Press, Cambridge, MA.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurement of educational products. *Natl. Soc. Study Edu. Yearbook* **17**(2), 16–24.

Thorndike, E. L. (1921). *The Teacher's Word Book.* Teachers College, Columbia University, New York.

Thorndike, E. L. (1921). Measurement in education. *Teachers Coll. Rec.* **22,** 371–379.

Thorndike, E. L. (1931). *The Teacher's Word Book of 20,000 Words.* Teachers College, Columbia University, New York.

Thorndike, E. L. (1935). *Thorndike-Century Junior Dictionary.* Scott, Foresman, Chicago, IL.

Thorndike, E. L. (1939). *Your City.* Harcourt, Brace and Company, New York.

Thorndike, E. L. (1940). *Human Nature and the Social Order.* Macmillan, New York.

Thorndike, E. L. (1940). *144 Smaller Cities.* Harcourt, Brace and Company, New York.

Thorndike, E. L. (1944). *The Teacher's Word Book of 30,000 Words.* Teachers College, Columbia University, New York.

Thucydides

Peter A. Furia
Wake Forest University, Winston–Salem, North Carolina, USA

Ari Kohen
Duke University, Durham, North Carolina, USA

Glossary

Alcibiades Athenian general initially chosen to lead the Sicilian expedition. He was recalled from the expedition and condemned to death by Athens, but he escaped to Sparta and helped them strategize against Athens toward the end of the war.
Athens The leading democratic city–state in Ancient Greece, of which Thucydides was a citizen.
Cleon Vengeful Athenian general and rival of Thucydides whom Thucydides consistently portrays in a negative light.
Melian dialogue Diplomatic exchange presented by Thucydides in which Athenian ambassadors justify their subsequent slaughter of the male population of the island of Melos.
Pericles Creator of the Athenian empire and the leader of democratic Athens at the beginning of the Peloponnesian War.
Sparta (Lacadæmon) The leading aristocratic city–state in Ancient Greece; the eventual victor in the Peloponnesian War.
The Peloponnesian War A 27 year military conflict between Athens and Sparta that took place between 431 and 404 B.C.E.

Thucydides (5th Century B.C.E.) is arguably the first person to engage in systematic social measurement. His lone surviving masterwork, *The Peloponnesian War*, stands as a founding text for the disciplines of history and political science. On its surface, *The Peloponnesian War* comprehensively and objectively chronicles the 27 year military struggle between Athens and Sparta from which the latter emerged victorious in 404 B.C.E. Yet Thucydides' aim and approach in recounting the Peloponnesian War is extensively debated, often reflecting enduring disagreements about the philosophy and methodology of social science.

Thucydides the Person

We know little about Thucydides' life beyond the few things he tells us in *The Peloponnesian War* itself. An Athenian citizen, Thucydides was probably born a few years prior to 460 B.C.E., likely dying shortly after the end of the Peloponnesian War. Perhaps significantly, we do know that he was elected to the office of general (*strategos*) in the eighth year of the war (424 B.C.E.) but then exiled from Athens shortly thereafter (for failing to prevent Sparta's capture of the strategically important city of Amphipolis). Thucydides also tells us that he began work on his history at the war's outset (I 1) and lived through its entire 27 years (II 65, V 26). He did not, however, live to complete the narrative, which breaks off abruptly in the autumn of 411, the 21st year of the conflict.

Overview of the Work

Far from a single cohesive nation, Ancient Greece in the 5th century B.C.E. was composed of dozens of independent city–states (*poleis*). Most of these states were, however, militarily subject to either Athens or Sparta, the two great powers in Greece at the time. At its most basic, therefore, the Peloponnesian War was a struggle for regional hegemony between these two powers: Although democratic Athens has exerted a greater influence on

the thinking of Thucydides' contemporary interpreters than has aristocratic Sparta, the latter of these two powers was clearly the stronger at the war's beginning. Indeed, as Thucydides tells us, most Greeks initially believed that the Spartans and their allies would emerge victorious from the conflict within three years (VII 28). Sparta's obvious military superiority on land, however, was offset by the supremacy of Athenian sea power, which prolonged the war and left its outcome uncertain until the fall of Athens and its empire.

Book I of Thucydides' chronicle features a wide-ranging "Archaeology" of Greek society and politics prior to the Peloponnesian War. As Thucydides recounts, the Athenians and Spartans had once been allies, forming the Hellenic League in 481 B.C.E. in order to help all of Hellas rebuff a Persian invasion. Despite major victories against their common enemy, relations between the two powers were strained, eventually resulting in the withdrawal of Sparta to lead its prewar alliance, the Peloponnesian League. The Athenians in turn created an alternative alliance, the Delian Confederacy, which accepted Megara, a defecting member of the Peloponnesian League. War broke out between the two alliances in 460 and continued until 446, when both sides accepted a peace treaty. The remainder of the narrative commences with the broken peace of 431.

From the beginning of Book II through the beginning of Book V, Thucydides recounts the first, indecisive phase of the conflict often called the Archidamian war. While the Spartans regularly attacked the Attic countryside, the Athenians responded by ravaging the Peloponnesian coastline. In this 10-year period, the Athenians also suffered a plague, intervened in Sicily, and routed the Spartans at Pylos-Sphacteria; they subsequently lost important holdings in battles near Thrace. This phase of the war culminated with what was to be a 50-year peace treaty, the Peace of Nicias, in 421.

The treaty, however, collapsed after only eight years (many of which were actually spent in open dispute). In 415, the Athenians intervened a second time in Sicily, assisting their allies, the Egestæans, and attempting to expand their empire through the conquest of Syracuse, the preeminent Spartan ally in Sicily. The subject of Books VI–VII, the so-called Sicilian war, ended with a crushing defeat for the Athenians. The effect of this defeat on Athens was all the more pronounced because it coincided with the resumption of hostilities by the Spartans and their allies closer to home. Facing Spartan troops permanently based in the Attic countryside, an increasing number of subjects in revolt, and hostile naval forces subsidized by their old enemy, Persia, the Athenians were overmatched. Despite these unfavorable circumstances, this final phase of the conflict, the Ionian or Declean war, lasted another 10 years. It ended, however, with the destruction of the Athenian navy in 405 and (after being starved into submission) the surrender of Athens in 404.

Thucydidean Method

What exactly Thucydides hoped to achieve via his history of the Peloponnesian War has been endlessly debated. Though some might make a claim for Herodotus, the more rigorous Thucydides is generally considered the inventor of descriptive history as such. Noting the absence of "romance" in his narrative, Thucydides avers that it might nonetheless "be judged useful by those inquirers who desire an exact knowledge of the past (I 22)." In turn, many have suggested that Thucydides' aim is no more and no less than to provide an accurate empirical description of the events that he recounts. Indeed, historians often remind us that Thucydides' narrative contains an astounding volume of painstaking detail, detail that Thucydides' more social-scientifically inclined readers may ignore at their peril. Yet while we cannot underestimate the importance of Thucydides' invention of comprehensive historical description, social scientists are understandably drawn to the few "causal" or "explanatory" claims that he appears to be making throughout the work.

Perhaps the most significant of these claims is Thucydides' famous comment about why the Peloponnesian War began: "The growth of the power of Athens, and the alarm which this inspired in Lacedæmon, made war inevitable" (I 23). At first blush, this passage seems to imply that Thucydides possesses a broadly "materialist" view of empirical causation. Specifically, he seems to suggest that the onset of the Peloponnesian War is fully explained by the existence of a single, concrete and observable variable: namely, *increasing Athenian power*. To be sure, many have noted that an accurate measurement of Athenian power proves elusive. In discussing the Peloponnesian War, as in other cases, debate among international relations scholars and military strategists abounds with disagreements over the degrees to which, e.g., population, wealth, geography, and munitions determine a state's "military capabilities," and, in turn, over whether it is "absolute" or "relative" military capabilities that are most significant.

More important, however, this single sentence of *The Peloponnesian War* is characteristic of Thucydides' occasional "editorial" statements in that it has likewise spawned more radical disagreement among his interpreters. Specifically, numerous scholars have noted that the most proximate cause of the war mentioned by Thucydides is not the material change in Athenian power, but rather, the more "ideational" variable of fear or "alarm" that this change inspired in Sparta. In

other words, debate over this passage closely reflects debate between idealist and materialist approaches to social science in general (as evidenced most famously via Marx's methodological critique of Hegel). Indeed, generation after generation of philosophers of social science have claimed a methodological pedigree from Thucydides. Contemporary constructivist and postmodernist scholars, for example, argue that the onset of the Peloponnesian War is explained by the breakdown of social and diplomatic discourses in 5th Century Hellas. In point of fact, however, debate over Thucydides' account of the onset of the Peloponnesian War provides only one example of how arguments about Thucydides reflect broader debates about social measurement. If anything, Thucydides' incomplete account of the outcome of the Peloponnesian war—that is, his account of why Athens loses—has proven still more suggestive to those interested in social-scientific theory and method.

As noted, a purely materialist account of Athens' defeat in the war is by no means difficult to construct. If we attend simply to the balance of power between the Spartan and Athenian sides, the fact that Athens avoided defeat as long as it did is arguably the greater mystery. Understandably, however, the balance of military power between the two sides is by no means the only variable that scholars have noted as contributing to the war's eventual outcome.

Many of Thucydides' interpreters, for example, stress the critical role in the conflict played by individual Athenian leaders. The charismatic Pericles is seen as crucial to Athens' initial pursuit of empire, the vengeful Cleon blamed for the subsequent alienation of her subject states, the perplexing Alcibiades alternately celebrated and condemned for his role in the initiation and execution of the Sicilian expedition. Despite the no doubt significant role played by these individual citizens, however, a still more common mode of explaining Athens' defeat centers on the political context in which they came to power: namely, as leaders of the Athenian democracy. Even the most prominent defenders of the benefits of democracy in the conduct of foreign policy, including Michael Doyle, note that the experience of Athens during the war casts the *demos* in a not entirely favorable light: "It is here, in Thucydides' *History*, that democracy first acquired its reputation for such disastrous factionalism" (79). In Athens, of course, the *demos* exercised much more direct control over foreign policy than it does in representative democracies today. Moreover, even insofar as it delegated some of this authority to the elected *strategoi*, Thucydides notes that the most jingoistic of these Athenian generals, such as Cleon, were often successful in appealing to what he seems to have regarded as the lowest common denominator of popular support.

We return to the question of Thucydides' moral stance in regard to the Athenian polity and its foreign policy in a moment. At present, it is enough to note that *The Peloponnesian War* raises many of the classic causal questions about whether a state's system of government influences its foreign policy: Advocates of the "democratic peace" proposition, for example (sometimes called the one "iron law" of contemporary international relations scholarship) argue that even if Athens was bellicose in its relations with Sparta and other autocracies, it avoided conflict with democratic states. In contrast, those who see democracies as possessing a particularly pronounced tendency for the foolish overextension of empire have likewise found an early cautionary tale in Athens. Last but not least, scholars intrigued by the fact that modern democracies rarely lose the wars they enter—the so-called "powerful pacifists" hypothesis—look back to the counterexample of Athens in hopes of determining whether anything inherent in democracy predicts its military success.

Yet is it appropriate for scholars to cavort across the millennia in search of universal social-scientific laws? Is not everything that Thucydides says about democracy and foreign policy so wrapped up in a bygone social context that to try and apply it to today's world constitutes a fateful *hubris*? To argue as much is certainly credible, but, to claim that Thucydides himself would have made this argument strikes us as less so. For just as he can be claimed as the inventor of "ideographic" methods and "thick description," so too can Thucydides be noted as the first researcher to possess a universalizing or "nomothetic" urge. He makes this evident, of course, in telling us that "the future ... in the course of human things must resemble if it does not reflect [the past]," and in turn that, "In fine, I have written my work, not as an essay which is to win the applause of the moment, but as a possession for all time" (I 22). It is this justly famous remark, perhaps, that constitutes Thucydides' most important and controversial contribution to the history of social measurement.

Measurement and Morality

Even if Thucydides is attracted to the project of a nomothetic social science, however, this as yet tells us nothing about how this may or may not comport with his other aims in writing *The Peloponnesian War*. He is, after all, read far less by methodologists than he is by political theorists, and apart from the question of his empirical aims stands the question of what, if any, moral lesson he wishes us to take from the narrative. Asking this question is interesting for various reasons. First, there is the possibility that Thucydides displays normative bias in recounting the events of the war, leading us to think twice about his status as an objective historian and social scientist. Alternatively, it could be that Thucydides felt that even an unbiased presentation of the facts of the war would lead the reader to certain moral or ethical conclusions. But just what are Thucydides' moral and

ethical commitments, if any? Many readers have seen Thucydides as the father of *realpolitik*, that is, of the view that morality and politics are incompatible. Such an ethos was, to be sure, frequently articulated in Ancient Greece, as, for example, by some of Socrates' more famous interlocutors. In Thucydides, *realpolitik* is most clearly defended by the Athenian ambassadors to Melos. For various reasons, however, most scholars now aver that Thucydides is little more sympathetic to these ambassadors than is Plato sympathetic to, e.g., Thrasymachus or Callicles.

While the preceding section discussed the claim that Sparta began the war out of fear, Thucydides also notes that the Spartans were compelled by the Corinthians to oppose Athenian injustices at Potidæa (I 71). The Athenians, for their part, argued that they were obligated to acquire and expand their empire out of fear of external threats, despite the knowledge that to do so would result in widespread opprobrium. In addition to this line of thought, they also suggested that "honour and interest afterwards came in" (I 75). These claims, in addition to the Athenian assertion that they are uniquely entitled to rule others, serve as the ground upon which Thucydides bases his discussion of Athens.

The most famous defense of the ethical underpinnings of Athenian war aims comes from Pericles, in an oration for the war's first fallen soldiers. He encourages the Athenians to love their city as the soldiers have and to be ready to make the same sacrifice should they be called on to do so (II 43). Athens, he says, is so beautiful and noble as to inspire love among its citizens, not merely because it is powerful but also because it acts for reasons beyond self-interest (either individual or collective). Arguably, however, this still falls well short of acting justly. As David Bolotin points out, Pericles "boasts that Athens has everywhere established everlasting memorials of evils as well as goods . . . And when he speaks of everlasting memorials of evils and goods, he has in mind the evils that Athens has suffered as well as the harm it has done to others" (20).

While Pericles may exaggerate the virtues of the Athenian people, Thucydides himself is often critical of the city in the post-Periclean era. At times, to be sure, Thucydides clearly approves of the workings of Athenian democracy. After a lengthy siege succeeded in crushing a rebellion at Mytilene, the Athenians initially determined that all the male citizens should be punished with death, and the women and children sold into slavery (III 36). A day after dispatching a ship to deliver the order, however, they experienced a change of heart and Thucydides gives a lengthy recounting of their second debate, with speeches by his perennial rival Cleon, who favored the original decree, and Diodotus, who opposed it. In the end, the Athenians adopted the position of Diodotus and hastily sent a second ship to prevent the original order from being carried out (III 49). As Michael Walzer observes, "It is the appeal to interest that triumphs—as has often been pointed out—thought it should be remembered that the occasion for the appeal was the repentance of the citizens. Moral anxiety, not political calculation, leads them to worry about the effectiveness of their decree" (9).

In stark contrast, Thucydides provides no account of any democratic process leading to the decision to attack the island of Melos and, ultimately, to kill its male citizens and enslave its women and children. Unlike Mytilene, a former ally of Athens that rebelled and joined the Spartans, Melos had chosen to remain neutral until the Athenians violently encroached on their territory (V 84). All that is recorded, this time, is the exchange between the Athenian generals, Cleomedes and Tisias, and the Melian representatives prior to the official outbreak of hostilities. Here, the Athenians remove the notion of justice from the discussion at the very outset:

> *For ourselves, we shall not trouble you with specious pretences—either of how we have a right to our empire because we overthrew the Mede, or are now attacking you because of wrong that you have done us—and make a long speech which would not be believed; and in return we hope that you, instead of thinking to influence us by saying that you did not join the Lacedæmonians, although their colonists, or that you have done us no wrong, will aim at what is feasible, holding in view the real sentiments of us both; for you know as well as we do that right, as the world goes, is in question only between equals in power, while the strong do what they can and the weak suffer what they must (V 89).*

In the end, the Melians chose not to subject themselves to Athens and were besieged. After months of fighting, Melos was betrayed by a number of its citizens and yielded to the Athenians, who put to death all the men, and sold the women and children into slavery (V 116). According to Walzer, "We are to understand that Athens is no longer itself. Cleomedes and Tisias do not represent that noble people who fought the Persians in the name of freedom . . . They represent instead the imperial decadence of the city state" (7). Whether or not the potential for imperial overreach was always present in Athenian democracy, however, it is eventually this imperial impulse that carries the day, ultimately leading to the ill-fated Sicilian expedition.

What did Thucydides himself think of Athenian imperialism and this final campaign to extend it? Thucydides clearly does not possess the outright aversion to imperialism that Walzer and most of the rest of us do today. He presents Pericles, the father of Athenian empire, in a highly favorable light, and various commentators note grounds on which the Athenian empire would have been viewed as a progressive enterprise at the time. (It is, for example, generally agreed that

the limited democracy brought by Athenian rule was embraced by the lower classes in her subject cities.) Even if this is the case, however, it is unclear that Thucydides himself was sufficiently fond of the political system of post-Periclean Athens to advocate its export to other Greek states.

In any case, a quite different account of Thucydides' stance regarding Athenian imperialism suggests that he actually condemns it for being too hesitant. This reading centers upon the mysterious destruction of the city's statues of Hermae that occurred just prior to the Sicilian expedition. The expedition was decided upon at a time of great factional conflict within Athens, and citizens sympathetic to Alcibiades, the brilliant general chosen to lead the expedition, were widely blamed for the statues' destruction. Eventually, this led to Alcibiades being stripped of his post. Thomas Pangle and Peter Ahrensdorf, among others, argue that the expedition might have succeeded were it not for this strange turn of events: "According to Thucydides, the Athenians could have conquered Sicily, and consequently could have won the war, if only they had retained the services of Alcibiades" (26).

Such a reading returns us to a view of Thucydides as an advocate of *realpolitik*. That is, rather than responding to the mutilation of the Hermae with the cool rationality that Thucydides, it is argued, himself recommends, the Athenians took drastic action in the face of a religious crime that they also interpreted as a sign of divine displeasure with their imperial ambition. "It would seem," Pangle and Ahrensdorf argue, "that the Athenians interpret the mutilation of the Hermae not in the light of their own argument on justice and self-interest but in the light of their suspicion or fear that they are guilty of injustice" (27–28). Like much else in Thucydides, however, the suggestion that he sympathizes with Alcibiades is controversial: Alcibiades is likewise responsible for undoing the Peace of Nicias, and Nicias is perhaps the only figure for whom Thucydides expresses even greater personal fondness. Commenting on his "unwarranted butchering" during the war, Thucydides remarks that "of all the Hellenes in my time, [Nicias] least deserved this fate, seeing that the whole course of his life had been regulated with strict attention to virtue" (VII 86).

Many scholars thus point to Thucydides' remark about Nicias, among other passages, as evidence that he views war as fundamentally tragic. While it would be a stretch to view Thucydides as a thoroughgoing pacifist, a case can certainly be made that he came to see the Peloponnesian War as a mistake—not just for Athens, but for all of Greece. Such a view is perhaps most strongly evidenced in Thucydides' discussion of prewar Hellenic society and politics at the beginning of Book I. For here, Thucydides speaks less of Athens and Sparta than he does of the Pan-Hellenic "country" (I 2) and "race" (I 1). Indeed, on more than one occasion, Thucydides seems to regret the fact that, except for during the war against Persia, the states of Greater Hellas proved "incapable of combination for great and national ends" (I 15), much less of uniting in "a spontaneous combination of equals" (I 16). While hardly a contemporary global citizen, then, it would seem that Thucydides is, like Socrates, drawn to the Pan-Hellenic ideal of cooperation and perhaps even confederation among Greek states (an ideal by no means uncommon in his time). Viewed from this perspective, *The Peloponnesian War* becomes a tragic story indeed, one in which an unwarranted and essentially "civil" conflict slowly engulfs a divided Hellas.

Conclusion

Obviously, we will never know exactly what sort of justice, if any, Thucydides felt that his native Athens owed to the rest of Greece and the world beyond. Nor, for that matter, will we know in what sense, if any, he thought it inevitable that states powerful enough to engage in hegemonic war will do so. We should, perhaps, take seriously the possibility that Thucydides changed his views of these issues while writing a very long narrative about a very long conflict. In any event, not only the diverse ethical thinking that Thucydides has inspired but also the numerous methods that he introduced to social measurement will continue to prove important gifts to posterity.

See Also the Following Article

Political Violence

Further Reading

Bolotin, D. (1987). Thucydides. In *History of Political Philosophy* (L. Strauss and J. Cropsey, eds.), 3rd Ed. University of Chicago Press, Chicago.

Connor, W. R. (1994). *Thucydides.* Princeton University Press, Princeton, NJ.

Crane, G. (1998). *Thucydides and the Ancient Simplicity: The Limits of Political Realism.* University of California Press, Berkeley.

Doyle, M. W. (1997). *Ways of War and Peace.* Norton, New York.

Forde, S. (1989). *The Ambition to Rule: Alcibiades and the Politics of Imperialism in Thucydides.* Cornell University Press, Ithaca, NY.

Garst, D. (1989). Thucydides and neorealism. *Int. Studies Quart.* **33,** 1.

Johnson Bagby, L. M. (1994). The use and abuse of Thucydides in international relations. *Int. Organ.* **48,** 1.

Lebow, R. N. (2001). Thucydides the constructivist. *Am. Polit. Sci. Rev.* **95,** 3.

Luginbill, R. (1999). *Thucydides on War and National Character.* Westview, Boulder.

Pangle, T. L., and Ahrensdorf, P. J. (1999). Classical realism: Thucydides. In *Justice Among Nations: On the Moral Basis of Power and Peace.* University of Kansas Press, Lawrence.

Thucydides (1982). *The Peloponnesian War* (R. Crawley, transl.; T. E. Wick, ed.). Modern Library, New York.

Walzer, M. (1977). *Just and Unjust Wars: A Moral Argument with Historical Illustrations.* Basic Books, New York.

Thurstone's Scales of Primary Abilities

Marcel V. J. Veenman
Leiden University, Leiden, The Netherlands, and
University of Amsterdam, Amsterdam, The Netherlands

Glossary

centroid method A stepwise method of factor extraction, with the first factor representing the common variance in all tests, and further extracted factors representing bipolar dimensions with lower factor loadings.
g A general factor underlying scores of tests for mental ability.
oblique rotation The rotation of factorial axes with a departure from a constant angle of 90° between two axes. Rotated factors are correlated to a certain extent.
orthogonal rotation The rotation of factorial axes where the angle between two axes is held constant at 90°.
primary factors Psychologically interpretable, rotated factors derived from factor analysis of the correlations between psychological variables.
secondary factors Factors derived from factor analysis of primary factor scores.
simple structure Rotating factorial axes in order to obtain a limited number of tests loading high on a factor, and maximizing zero loadings of other tests.

One of the recurring issues in research on intelligence concerns the question of whether intelligence is one ability or whether it represents manifold dimensions of mental abilities. During the first decade of the 20th century, C. Spearman's research stressed the importance of a single, general ability underlying the performance on mental tests. L. L. Thurstone, on the other hand, argued that intelligence represented several separated mental faculties. He obtained nine uncorrelated primary factors while using another method of factor analysis (centroid extraction and rotation to simple structure) in his analyses of various test data from 240 college students. Thurstone found no evidence in favor of a common general factor (*g*). Therefore, he argued that an individual's intellectual ability should not be represented as a single IQ index, but rather should be described in terms of a profile of factor scores on the primary mental abilities. Correlations among tests in Thurstone's battery, however, appeared to be substantial. Consequently, Thurstone eventually had to acknowledge that a factor analysis of the primary factor scores revealed a general secondary component, a conclusion that is in line with findings from J. Carroll's extensive survey and reanalysis of factor-analytic studies on the structure of intellectual abilities. It is concluded that Thurstone's bequest may be the best of both worlds: IQ does indeed represent a general intelligence factor, but a more specified profile of primary mental abilities might add to the understanding of an individual's mental capabilities.

Spearman's Early Work

In 1904, Spearman used a basic form of factor analysis in order to substantiate his assumption that only one general factor was underlying the matrix of correlations between all possible pairs of tests. According to his two-factor model of intelligence, each mental test would load on a general factor (which was referred to as *g*), as well as on a factor that was specific to that test (referred to as *s*). Spearman argued that *g* should not be equated with intelligence because he regarded *g* merely as the factor common to all mental tests. At first, he assumed that each test would further load on its own specific *s* factor. Later on, he had to admit that certain groups of mental tests might have factors other than *g* in common. These so-called group factors would reveal loadings of a limited set of tests, but certainly not all tests. For instance, Spearman identified group factors for tests of verbal or spatial ability. Eventually, Spearman's two-factor model consisted of

a general factor common to all mental tests, a limited number of group factors, and residual *s* factors. Spearman, however, still stressed the importance of the general factor *g*.

Thurstone's Theoretical Focus

Thurstone's initial position was antagonistic to Spearman's. Thurstone believed that group factors were far more important than the general factor that was advocated by Spearman. He postulated intelligence to be an assembly of mental faculties, not a generalized mental ability. He argued that the general factor was an artifact of Spearman's method of factor analysis, as Spearman failed to rotate the factorial axes after obtaining an initial solution.

Thurstone's Contribution to Factor Analysis

Thurstone was an ardent advocate of factor analysis. Together with his co-workers, he developed the centroid method of factorizing, defined principles for orthogonal and oblique rotation to simple structure, and was among the first to apply matrix algebra to factor analysis.

The first step in Thurstone's method of factor analysis was to compute correlation coefficients between all mental tests in a given test battery. This matrix of correlations was expected to contain only few negative correlations, low in magnitude, because complex mental tests were assumed to be composed of a variety of common elements or grouping factors.

The second step concerned the factorization of the correlational matrix. According to the parsimony principle, the number of factors extracted was restrained: scores on a test battery were to be accounted for by fewer factors than the number of tests. In line with this principle, the centroid method for transformation of a correlational matrix to an orthogonal factorial matrix was developed. Factors were extracted stepwise, which meant that factors were extracted one at a time, and the procedure was cyclically repeated on correlational matrices with formerly extracted factors partialed out. The first centroid factor that was extracted accounted for the highest variance in common to all tests. Hence, this factor had the highest mean of factor loadings, which were positive as a rule. Next, this first centroid factor was partialed from the correlational matrix of all tests, yielding a new matrix of residuals. Then the extraction procedure was repeated, rendering a second factor with less variance accounted for, lower factor loadings, and possibly bipolar dimensions. This procedure of repeated extraction of factors was terminated whenever the residuals approached chance level. In fact, the centroid method of factor extraction resembled principal components analysis. At the time, however, the centroid method was mathematically less demanding relative to principal components analysis, which became more readily available for the analysis of large matrices after the introduction of computers.

According to Thurstone's view, the extracted centroid factors were merely mathematical constellations without psychological salience. The orientation of centroid factorial axes represented arbitrary dimensions, and the coherence of groups of tests would only reveal itself after rotation of these axes to simple structure. One of the principles underlying simple structure was the positive manifold. Thurstone did not allow tests to load highly negative on an extracted factor, as he regarded it as unlikely that factorized mental abilities would negatively contribute to intellectual performance.

Figure 1 depicts a hypothetical rotation to simple structure. Evidently, the projection of tests (i.e., the circles) on the factorial axes of A and B is not perfect, resulting in relatively low, sometimes even negative, factor loadings. Orthogonally rotating the factorial axes to the left by 20° results in high factor loadings for a number of tests (represented by black circles) on either of the rotated axes A′ or B′, whereas the tests represented by white circles load high on neither A′ nor B′. The latter tests, however, may load substantially on another factor C or its rotated equivalent C′. Rotation to simple structure starts with a first pair of factorial axes and then resumes stepwise with further pairs of factorial axes.

Returning to Thurstone's position, the aim of rotation to simple structure essentially has been to obtain a factor matrix with a limited number of highly positive loadings on each factor, while maximizing the number of zero entries of the remaining factor loadings. If *g* existed, simple structure could not be attained, as one factor would show nonzero loadings for every test. To Thurstone it was

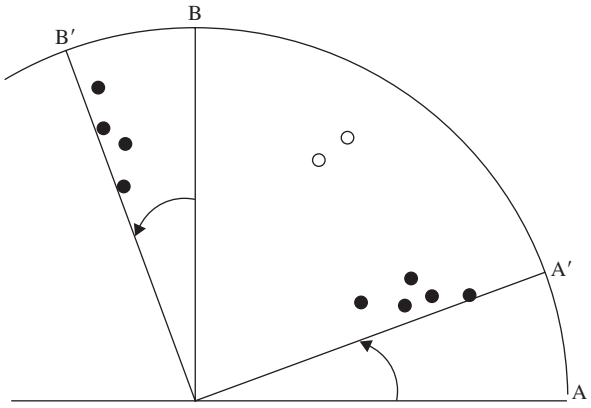


Figure 1 Orthogonal rotation to simple structure.

highly unlikely that, for instance, numerical tasks would make high demands on verbal or memory primary factors, and vice versa. Rotation to simple structure in fact endeavored to separate the sheep from the goats for each factor extracted.

At first, Thurstone relied on orthogonal rotations, which meant that the angles of all factorial axes were held constant at 90° during rotation (as depicted in Fig. 1). Orthogonally rotated factors were uncorrelated and, therefore, relatively "pure." In the 1940s, Thurstone developed the technique of oblique rotation because the primary factors emerged even more clearly from the data. With oblique rotations, he departed from the independency of factors, as he allowed for the angles between factorial axes to be less than 90°. Now the primary factors, which were intended to be entirely separated group factors, at least had some variance in common.

Primary Factors

In the late 1930s, Thurstone administered 56 mental tests to 200 students from the University of Chicago and 40 YMCA college students, who volunteered to participate in the test sessions. Mean age of participants was 19.7 years. The test battery covered a variety of mental tasks that were verbal, spatial, or numerical by nature. Tests measured fluency of production, abstraction, reasoning, or rote learning of paired associates. Some tests were adapted from existing materials, while others were especially designed and developed by Thurstone and his wife in order to cover a broad range of tasks. Many of these tests or equivalent tests are still being used in contemporary intelligence testing, e.g., figure rotation, surface development, arithmetic speed, number series, verbal analogies, syllogisms, and vocabulary. After collecting the data, a matrix of intercorrelations was first calculated from scores on the 56 tests. Almost all intercorrelations appeared to be positive, and the modal correlation coefficient was about 0.35. Next, 12 centroid factors were extracted. Orthogonal rotation to simple structure rendered nine primary factors with positive and near-zero loadings, but without substantial negative loadings. These nine factors were psychologically interpretable through the identification of common elements in tests that loaded 0.40 or more on the rotated factors. In the following paragraphs, a keyword and content description will be given for each primary factor, along with examples of pure tests that typically loaded on that factor only.

1. Spatial ability (S): The 13 tests that substantially loaded on this factor had a visual or spatial nature in common. This first primary factor should not be confused with the next factor, perceptual ability, as spatial abilities relate to mental imagery, rather than the perception of stimuli. Two representative tests were "flags," which required the rotation of nation flags, and "pursuit," which required participants to follow lines from start to end in a complex line pattern.

2. Perceptual ability (P): The nine tests that loaded on this factor to a large extent represented the facility in finding or recognizing items in a perceptual field. For instance, it involved the perception of an object that was embedded in irrelevant material. Two representative tests were "identical forms," which simply required participants to pick an object identical to the stimulus object out of an array of highly similar objects, and "word grouping," which required the categorization of easy words through the speeded perception of apparent relations among them and the identification of words that did not belong in these categories.

3. Numerical ability (N): The eight tests that highly loaded on this factor had in common a numerical nature. These tasks demanded a considerable proficiency in numerical calculation and reasoning. Pure numerical tests included those of arithmetic speed, such as "addition," "multiplication," and "division." The tests for numerical reasoning, however, also loaded on various other factors.

4. Verbal relations ability (V): The 13 tests that clearly loaded on this factor concerned the logical relations between ideas and the meaning of words. This fourth primary factor should be separated from the next verbal factor, word ability, as verbal relations pertained to the classification and association of ideas and semantics, rather than to the production of isolated words. Two characteristic tests were "inventive opposites," in which participants had to find two words with a meaning opposite to the stimulus word, and "verbal analogies," which required participants to find a relationship between two given words and to choose a target word by applying that relationship to another stimulus word.

5. Word ability (W): The six tests that loaded on this fifth primary factor were characterized by a fluency in dealing with isolated words. Both "disarranged words" and "anagrams" were relatively pure W-tests, which required participants to rearrange a jumbled sequence of characters in order to obtain a meaningful word.

6. Memory ability (M): The five tests that substantially loaded on this sixth factor concerned the recall or recognition of paired associates that had been presented shortly before. The two most unambiguous tests were "word-number," which involved the memorization and recall of paired associates of stimulus words and response numbers, and "initials," which required the memorization and recall of a list of names with initials.

7. Inductive ability (I): The five tests that loaded on this seventh factor asked participants to find a rule or principle underlying a set of stimulus items in a test. Two representative tests were "areas," which required participants to determine the total white surface area

from increasingly complex figures, and "tabular completion," which required participants to fill in missing numerical entries in a table by examining column headings. Even though "number series" loaded highest on induction, apparently this test also loaded on other factors.

8. Restriction in solution ability (R): This far less distinct factor concerned tasks that involved some sort of restriction in obtaining a solution. J. Guilford might have referred to the description of this factor as "convergent production." Seven tests loaded on this factor, but it is not entirely obvious that the two most representative tests, "sentence completion" and "mechanical movements," had task elements in common. Sentence completion required participants to add one appropriate word to an incomplete sentence, whereas mechanical movements asked participants to determine the direction of movements for interacting gear wheels.

9. Deductive ability (D): Finally, the ninth factor also appeared to be less well defined. The four tests that loaded on this last factor required participants to apply a rule to target stimuli. Relatively pure tests for D concerned the verbal syllogisms tasks of "reasoning" and "false premises." Participants had to judge whether an inference (e.g., "Mr. White is wealthy") logically followed from the given premises ("All wealthy men pay taxes. Mr. White pays taxes.").

As Thurstone rejected the existence of *g*, he also strongly opposed the use of a single IQ index as a general indicator of mental ability. He preferred a description of mental abilities in terms of an individual profile of factor scores on the primary abilities. In fact, he even tried to relate such individual mental profiles to the vocational interests of his participants. He selected certain atypical participants with extreme profiles as case studies and argued that their profile of mental abilities matched their vocational preferences. For instance, one student with a profile high on verbal relations V and perception P, but low on the problem-solving factor R, appeared to pursue a career as an actor. These highly selected case studies, however, are not entirely convincing. Indeed, Thurstone acknowledged that more than 90% of his participants were not extremely profiled. Thurstone, however, maintained that more pure, that is, factorially less complex, tests would further substantiate the simple structure of primary abilities.

Critiques on Thurstone's Work

A major critique, which has been put forward in the literature quite often, is that the sample in Thurstone's research was a highly selected group. Indeed, college and university students may be expected to score relatively high on mental tests. This appeared to be even more the case for Thurstone's sample. Thurstone compared the data of the university freshmen from his sample on a psychological examination of the American Council of Education with those of all freshmen from the University of Chicago and with the national norms, in order to prove that the distribution of ability in his sample met with requirements of normality. This comparison, however, also revealed that his freshmen participants performed much better relative to the national norms. Moreover, it showed that they even outperformed University of Chicago freshmen in general. Obtaining test data from such a highly selected group might have reduced the *g* factor due to restriction of range, and, consequently, it might have overstressed factorial loadings on primary factors.

A related issue concerns the conspicuous simplicity of many tests included in Thurstone's test battery. About half of the tests showed a distribution of scores that was skewed to the left, indicating that these tests were far too simple for the highly selected participants. Perhaps ceiling effects occurred; they might have contributed further to an overemphasis of primary factors, relative to *g*.

Another major point of criticism is that the scores on his 56 tests were almost invariably positively and considerably correlated. A modal correlation coefficient of 0.35, uncorrected for attenuation, indicated that the tests at least might have had a general factor in common. Furthermore, out of 48 tests that substantially loaded on any of the nine primary factors, 17 tests substantially loaded on two factors or more. Despite Thurstone's striving for factorially pure tests, multiple-factor loadings continued to exist. His later use of oblique rotation, which allowed for correlated factors, might also be indicative of a more general factor.

Finally, a fundamental dispute arose about whether to rotate or not. Those in favor of *g* would argue that simple structure rotations tended to eliminate or weaken the general factor. By rotation of the factorial axes, the first centroid factor was distorted and *g* was rotated out of existence. Thurstone, on the other hand, claimed that the extracted centroid factors were just mathematical orderings, without psychological relevance. However, both Spearman's and Thurstone's methods of factor analysis were mathematically justifiable.

Reconciliation of Spearman and Thurstone's Positions

Later on in 1945, Thurstone had to admit that second-order factor analysis, that is, factor analysis of the primary factor scores, yielded a general second-order factor that might represent Spearman's *g* factor. This eventually brought the antagonistic positions of Spearman and

Thurstone together. In fact, it turned out to be rather arbitrary whether to extract a general factor without rotation at first and then allow for group factors, or, conversely, to extract primary factors through rotation at first and then obtain a general secondary factor.

This arbitrariness is precisely the reason why some factor analysts have preferred confirmatory factor analysis to exploratory factor analysis. Confirmatory factor analysis requires factors and their parameters to be described beforehand, whereas no such restrictions are imposed on exploratory factor analysis. Despite the exploratory character of Thurstone's factor-analytic work, his contribution to the field has been that he took group factors seriously and that he developed methods for studying them more closely. After reading Thurstone's work, one has to acknowledge that, despite the manifestation of a higher order general component, intelligence also incorporates a number of lower order ability components of a more specific nature. In fact, the reconciliation with Spearman's position in Thurstone's later work was the overture to later hierarchical models of intelligence, such as Carroll's three-stratum structure, in which Spearman's *g* may be found at the top of the pyramidal structure and Thurstone's primary abilities may be found in the lower strata.

Recent Developments

In 1993, Carroll published his state-of-the-art book, in which he reported the results of his reanalysis of virtually all factor-analytic data on the structure of intelligence. This reanalysis covered over 450 data sets and included the data of more than 130,000 participants. Carroll used hierarchical factor analysis in order to extract primary, secondary, and eventually tertiary factors. He obtained about 70 primary factors, far more than Thurstone did. A closer inspection of Carroll's primary factors, however, revealed that Thurstone's primary abilities were distinctly represented by one or more primary factors of Carroll. Thurstone's V factor, for instance, might be equated with Carroll's verbal comprehension factors. Carroll explicitly mentioned Thurstone's inductive and deductive abilities as part of reasoning abilities. Memory is also ubiquitously present in Carroll's categorization of primary abilities. Both Thurstone's spatial and perception factors could be retraced in Carroll's category of visual perception abilities. Numerical facility was classified by Carroll as part of cognitive speed abilities. In fact, only Thurstone's R factor could not unequivocally be retraced from Carroll's categorization of primary abilities. Some of Carroll's primary abilities, on the other hand, were not represented in Thurstone's test battery, such as abilities of auditory perception and abilities of idea production (or divergent production in terms of Guilford). Obviously, the greater diversity of tests in Carroll's extensive data sets was responsible for a more fine-grained extraction of numerous primary abilities relative to Thurstone's work.

Because many of Carroll's primary factors were obtained by oblique rotation, these factors were substantially correlated. Therefore, higher order factor analysis was performed on each separate data set. In many cases, general intelligence (*G*) was extracted as a second-order factor common to all primary factors. Other second-order factors concerned fluid intelligence, crystallized intelligence, visual perception, auditory perception, cognitive speed, retrieval ability, and memory ability. Additional third-order factor analysis only yielded a general intelligence factor. Carroll concluded that his survey produced "abundant evidence" for the existence of a general intelligence factor at the secondary or tertiary order. He further asserted that mental tasks involve a variety of abilities, not only higher order abilities such as general, fluid, or crystallized intelligence, but also a number of primary abilities. This latter statement did not depart far from the conclusions Thurstone arrived at during the last decades of his scientific career.

See Also the Following Articles

Factor Analysis • Psychometrics of Intelligence • Thurstone, L.L.

Further Reading

Brody, E. B. (1992). *Intelligence. Nature, Determinants, and Consequences,* 2nd Ed. Academic Press, New York.

Carroll, J. B. (1993). *Human Cognitive Abilities. A Survey of Factor-Analytic Studies.* Cambridge University Press, Cambridge, UK.

Guilford, J. P. (1967). *The Nature of Human Intelligence.* McGraw-Hill, New York.

Thurstone, L. L. (1938). *Primary Mental Abilities.* University of Chicago Press, Chicago, IL.

Thurstone, L. L. (1947). *Multiple-Factor Analysis. A Development and Expansion of the Vectors of the Mind.* University of Chicago Press, Chicago, IL.

Thurstone, L. L.

Lyle V. Jones
University of North Carolina, Chapel Hill, North Carolina, USA

Glossary

factor analysis A statistical procedure for data reduction designed to extract a small number of components to account for interrelations among a larger number of variables.
primary mental abilities The basic mental abilities identified by L. L. Thurstone as being the components of human intelligence.
psychometrics The branch of psychology concerned with measurement.
psychophysics An area of psychology concerned with the quantitative relationships between physical stimuli and the psychological experiences of them.
simple structure In factor analysis, the stage at which factor rotation has met a set of criteria that maximize the number of variables with trivially small projections on most factors.

L. L. Thurstone was the pre-eminent 20th century advocate of a mathematical basis for scientific psychology. The influence of his work extends beyond his native United States to Europe, Canada, Central and South America, Australia, and the Far East. Effects of his contributions also transcend the bounds of psychology and remain a basis both of current measurement methods and of the future development of mathematical modeling in the social sciences.

A Peripatetic Childhood

Louis Leon Thunstrőm was born in Chicago, Illinois, on May 29, 1887. Both of his parents had emigrated from Sweden. His father had been an instructor of mathematics in the Swedish army, and later was a Lutheran minister, then a newspaper editor and publisher. His mother, born Sophie Strath, displayed a strong interest in music, and she taught both Louis and his younger sister to play the piano from an early age. Louis became an accomplished pianist, and his piano playing remained important to him throughout his life.

Louis entered elementary school in Berwyn, Illinois, transferred to a school in Centerville, Mississippi, and then between the ages of 8 and 14 attended both a public school and a private boys' school in Stockholm, Sweden, where he had moved with his family. He studied diligently to master the Swedish language so as to feel more comfortable in the company of his fellow students.

In 1901, the family returned by ship to the United States. To minimize travel costs, they brought with them only essential personal belongings. Louis announced that he would not leave Sweden without his three favorite books, a world atlas, Euclid's *Geometry*, and an encyclopedia of philosophic essays. With some reluctance, his parents acquiesced, but only if he would carry those weighty tomes in his arms, which he did.

The Thunstrőms settled in Jamestown, New York, where Louis entered high school. As a high school sophomore he sent a letter to *The Scientific American*, proposing a time-sharing arrangement for the Niagara River whereby it could be diverted to produce power for the region without serious harm to the tourist attraction at Niagara Falls. The letter became his first publication, in 1905.

When Louis became self-conscious about his imperfect mastery of the English language, his high school principal volunteered to serve as a tutor, and Louis would repeat words and sentences over and over, trying to speak English without a Swedish accent. Later he discovered that, in order to graduate from Jamestown High School, every senior had to present a five-minute talk to an assembly of several hundred students. He told the principal that he could not possibly do that, and he was excused from the requirement. At about this time,

the family officially changed the surname Thunstrőm to Thurstone, hoping that the change would promote greater acceptance of family members in their new surroundings.

A First Career in Engineering

Thurstone enrolled at Cornell University and earned a Master of Engineering degree in 1912. Prophetically, at Cornell he had attended lectures of psychologists Edward Titchener and Madison Bentley. He believed that an essential ingredient of good engineering was an understanding of how people would learn to use an engineering innovation, and he hoped that psychology could provide that understanding.

As a student at Cornell, Thurstone patented his own invention, a motion picture camera and projector that avoided flicker, a major problem with movies at that time. He actually succeeded in demonstrating a working model before Thomas A. Edison and his staff. They showed keen interest but did not adopt Thurstone's invention, citing the prohibitive cost of changing production in their plant from the Edison movie machine to Thurstone's. However, Edison invited Thurstone to become his laboratory assistant, and Thurstone accepted. Following his graduation from Cornell in 1912, he worked daily with Edison, an experience that had a lasting influence, shown later in several ways. When Edison was dissatisfied with a manuscript or a work in progress, he would discard that product and begin again, rather than edit or amend the product. Thurstone assumed that same habit: whenever not satisfied with a draft manuscript, he would abandon it and start over, rather than modify it. Thurstone was impressed by the fluency of ideas displayed by Edison; later, as Thurstone attempted to characterize human abilities, he sought to include several kinds of fluency as important components of creativity.

Thurstone left Edison's laboratory to become an instructor of engineering at the University of Minnesota, where he remained from 1912 to 1914. At Minnesota, he taught engineering courses and also enrolled in psychology classes taught by Herbert Woodrow and by J. B. Miner, continuing to follow his earlier interest at Cornell, the possibility of studying the learning function as a scientific problem.

Becoming a Psychologist

In 1914, Thurstone began graduate study in psychology at the University of Chicago. His sponsor was James Rowland Angell; he also was strongly influenced by the lectures in social psychology of sociologist George Herbert Mead. His term of residence as a student, however, was short-lived. In 1915, Walter Bingham visited the Psychology Department at Chicago to recruit assistants for the newly established Division of Applied Psychology at the Carnegie Institute of Technology in Pittsburgh, and Thurstone accepted Bingham's offer to become Bingham's assistant. While at Carnegie Tech, he completed his Ph.D. dissertation on the learning curve equation and was awarded the Ph.D. degree from the University of Chicago in 1917.

Carnegie Tech and Washington, D.C., 1915–1924

After two years as an assistant to Bingham, Thurstone became an instructor at Carnegie Tech in 1917. He then was promoted to assistant professor in 1918, to associate professor in 1919, and to professor and department head in 1920. (Thurstone had been rejected for the draft of World War I because he was underweight.) Much of his work at Carnegie was related to the development of classification tests for the U.S. Department of the Army. Among a number of publications that he completed at Carnegie was a substantial monograph on the nature of intelligence.

Thelma Gwinn had begun graduate study in Thurstone's Department of Applied Psychology in 1920, and she earned a Master of Arts degree in 1923, the year of the discontinuation of that department. Thurstone had arranged for a one-year position at the foundation-supported Institute for Government Research in Washington, D.C., and he brought Thelma Gwinn with him as his research assistant. Their assignment was to prepare materials and manuals based on new objective methods from which civil service commissions throughout the country could improve civil service examinations.

The office of the Institute for Government Research happened to be in a building also occupied by the American Council on Education (ACE), and Thurstone discussed with ACE staff members the creation of examinations that could be useful to colleges and universities for student guidance and placement. Those discussions foreshadowed more than 20 years of development and maintenance of ACE examinations for college freshmen and high school students, undertaken by Thurstone when he became a faculty member at the University of Chicago.

In 1924, L. L. Thurstone and Thelma Gwinn were married in Washington, D.C.

The University of Chicago, 1924–1952

At the University of Chicago, Thurstone was appointed Associate Professor of Psychology in 1924, then was

promoted to professor in 1928 and to Charles F. Grey Distinguished Service Professor in 1938.

First Approaches to Psychometrics

Starting in 1924, Thurstone taught a course in descriptive statistics, but his primary interest was in mental test theory. He initiated a course on test theory for which he developed the content, as there was no precedent for such a course. That led to a plethora of research publications on psychological measurement during the ensuing years. He was especially pleased with his first publication on the topic in 1925 that set forth the principle on which his theory of psychological measurement was to rest, namely, that at a given age group, a construct could be assumed to have a Gaussian distribution over individuals and thus could be described by two parameters, a mean and a measure of dispersion, thereby providing a scaling method for psychological traits. He extended this principle from its application to mental testing to the realm of psychophysics. Fifty years after the appearance of these papers on psychophysics, R. Duncan Luce presented an appraisal of their impact, especially on signal detection theory. In 1977, Luce noted that research results "clearly complicate the Thurstonian model without, however, destroying its basic spirit" (p. 487).

Thurstone went on to apply his basic theory to the assessment of attitudes and values. Prior to Thurstone, attitudes and values had been viewed as resistant to quantification. By developing objective procedures for their measurement, Thurstone placed social psychology on a quantitative scientific conceptual platform.

From 1924 to 1947, Thurstone developed annual editions of the ACE Psychological Examinations for High School Graduates and College Freshmen. He also annually published (usually jointly with Thelma Thurstone) norms for each current year of testing. (After earning the Ph.D. degree in psychology, Thelma continued to assist and then to collaborate and co-publish with her husband.)

Thurstone's office was located in the basement of the social science research building at the University of Chicago, while all other faculty members in the Department of Psychology were in other locations. Near his office, Thurstone maintained a spacious workroom, filled with tables and with books and tools used primarily for test development and related analyses. At dinner at home one evening, probably around 1930, Leon (the first name that he preferred as an adult) confided to Thelma that he thought it appropriate to establish the workroom as a psychometric laboratory, and that he would be the laboratory director. Thelma immediately commented that she thought that neither the Department Chairman nor the Dean of the college was likely to approve that. In reply, Leon agreed that she was quite right, which was why that afternoon he had attached to the door of the workroom a nameplate on which was etched "Psychometric Laboratory." No higher authority ever questioned that action, and the world-renowned Psychometric Laboratory thus was established.

Factor Analysis

The development of new annual forms of the ACE examinations was a labor-intensive undertaking, and the Thurstones were delighted when they were approached by the Community Work Education Service, the higher education branch of the Works Progress Administration (WPA) of Franklin Roosevelt's first term, and asked if they could employ some personnel. The WPA representative suggested 100 people. "No, not 100," said Thurstone, "but as many as 20 would be welcome." (Among those recruited in this way was Ledyard Tucker, who later earned his Ph.D. with L. L. Thurstone.) With an increased staff, the Thurstones developed a battery of 57 tests, which then were administered to about 300 University of Chicago freshmen who had volunteered to spend about 15 hours to take the tests during a week of vacation from classes. Test answers were scored by hand and then were subjected to detailed analyses, guided by the development of mathematical methods to facilitate the interpretation and classification of the aptitude and achievement tests.

In 1904, Spearman had postulated the existence of a common general factor of intelligence, and had shown that correlations between parts of a test could be ascribed to the action of this common factor. Thurstone, who had amassed experience with a great variety of test batteries, soon realized that one would have to consider not just one, but several common factors, and that the "partial correlations" that remained after common factors had been extracted could be attributed to "unique" parts or measurement error. Thus was born multiple factor analysis.

Today there may be insufficient understanding of the enormity of the analysis task undertaken in Thurstone's Psychometric Laboratory. There were no computers. Electrical calculators (frequently only half-automatic), slide rules, and graphical aids were used to perform the tedious item analyses and calculations. The use of tabulating machines for mathematical operations was still in its infancy. Extraction of factors had to be done by approximation (the centroid method). To interpret the nature of these factors, L. L. Thurstone invented the "simple structure" concept. By transforming (rotating) the factor loadings, usually graphically, a display was produced that had only very few of the "loadings" significantly high, while the majority were near zero. Thus, the few high loadings identified the nature of each factor. Thurstone stressed as a major advantage of simple structure the relative invariance of factorial description under alternative samplings

from a universe of tests as well as from a population of test takers. He believed that the formulation of simple structure and of the procedures to achieve it constituted his most important contributions to factor analysis.

Results from the analyses of the battery of 57 tests supported a set of "primary mental abilities" defined by Thurstone. The predominant factors were verbal, numerical, perceptual, spatial, word fluency, memory, and reasoning. The immediate consequence of the gigantic work on primary mental abilities was the development of batteries of tests to be used for the assessment of scholastic aptitude of students in grade school and high school.

In 1938, Thelma Thurstone joined the faculty of Chicago Teachers College, which gave her access to the Chicago public schools. Soon there were batteries of test forms for a broad age range. In 1946, L. L. Thurstone, Robert Burns, and Lyle Spencer formed Science Research Associates (SRA), a Chicago company that published and analyzed many tests and educational materials. At first, these were just those contributed by the Thurstones, but this quickly expanded to include other specialized tests commissioned by schools and private corporations, e.g., tests for U.S. State Department applicants, Sears Roebuck & Co. management applicants, and the National Merit Scholarship Test used nationwide until 1967 for the selection of National Merit Scholars. (Later, SRA was acquired by IBM, in response to an increasing demand for computerized educational materials.)

Psychometrics as a New Branch of Applied Psychology

In 1929, at age 29, Robert Hutchins had been named President of the University of Chicago. Among many innovations, Hutchins in the early 1930s encouraged individuals as young as 15 or 16 to enroll as university students. A related provision provided course credit by examination as an alternative to enrollment in required courses. L. L. Thurstone was appointed chief examiner, responsible for the development, administration, scoring, and reporting of examinations. Over the years, many of the individuals who served under Thurstone as examiners became prominent contributors to quantitative psychology, e.g., Dorothy Adkins, Harold Gulliksen, Paul Horst, Marion Richardson, John Stalnaker, and Dael Wolfle, among others.

In the mid-1930s, prompted by graduate students and the examiners, Thurstone helped to establish the Psychometric Society, dedicated to the support of psychology as "a quantitative rational science." The society founded the journal *Psychometrika*, which has been maintained for all ensuing years as a quarterly international journal. The first President of the Psychometric Society was L. L. Thurstone. Thelma Thurstone said that he really had labored over his 1936 presidential address, an abstract of which, published in *Science*, is still worth reading today.

The presidency of the Psychometric Society was one of a large number of honors bestowed during Thurstone's tenure at the University of Chicago. He was President of the Chicago Psychology Club, 1928–1929, President of the Midwestern Psychology Association, 1930–1931, and President of the American Psychological Association, 1932–1933. In 1938, he was elected a member of the National Academy of Sciences, one of 18 psychologists who were members at that time. Among other honors, he was a Fellow and a member of the Board of Directors of the American Statistical Association, a member of the American Philosophical Association, a Fellow of the American Academy of Arts and Sciences, and an Honorary Fellow of the British Psychological Society. In 1949, he received a career award from the American Psychological Association.

During World War II, Thurstone served on the Committee on Classification of Military Personnel of the U.S. Adjutant General's Office, and also authored psychological tests that were used for the classification of military personnel. Following the war, his research was supported by contracts both from government agencies such as the Army Quartermaster Corps and the Air Force Office of Scientific Research, and from corporate bodies such as Sears Roebuck & Co. Such outside support continued until his retirement from the University of Chicago and even beyond, after he moved to the University of North Carolina (UNC).

By 1950, the Psychometric Laboratory at the University of Chicago had become one of the university's great attractions. Visiting scholars, research fellows, and doctoral and postdoctoral students arrived from all over the United States and Europe to study with L. L. Thurstone. To all those who were at the Psychometric Laboratory at that time, the experience, both personal and professional, was unforgettable. Many of them later became leaders in the fields of psychometrics, sociology, and statistics. In Sweden, France, Germany, Switzerland, and in several universities in the United States, scholars referred to themselves as "Thurstonians" and continued to work in the Thurstone tradition. Today, many Ph.D. students of those scholars think of themselves as Thurstonian grandchildren.

The Thurstones' home in south Chicago was only two blocks from the Social Science Research building on campus, the location of Thurstone's office and laboratory. He spent long hours at the office, and typically joined faculty members from other departments for lunch at the faculty club, located midway between his home and office. At home he had installed a seminar room, fully equipped with blackboard and podium. On many Wednesday evenings, the Thurstones would entertain a guest speaker and invite as many as 30 guests to attend a seminar, to

participate in discussion, and then to enjoy cakes and coffee served by Mrs. T., as Thelma was affectionately known. Invitees typically felt privileged to be included at these events.

In 1952, Thurstone turned 65, the age of mandatory retirement at Chicago. He received offers for continued employment from the University of California, Berkeley, the University of Washington, and the University of North Carolina. All three offers were attractive, but only one included a faculty appointment for Thelma Thurstone, the offer from North Carolina, and that feature was a determining factor when that offer was accepted.

Final Years at the University of North Carolina, 1952–1955

At Chapel Hill, the Psychometric Laboratory occupied all of Nash Hall, a two-story building on the edge of campus, about two blocks from New West building, the location of the Psychology Department. With research funding from outside sources, Thurstone was able to pay salaries for laboratory personnel (including his own) and to provide stipends for graduate students who served as research assistants.

At Chicago, Lyle V. Jones had been supported on a National Research Council Fellowship in 1950–1951; he was one of several postdoctoral fellows from the United States and abroad in the Psychometric Laboratory that year. Jones became an Assistant Professor of Psychology in 1951, and also was Thurstone's successor as Director of the Psychometric Laboratory at Chicago when Thurstone departed in 1952. Some of Thurstone's research contracts were transferred to North Carolina, while others stayed at Chicago as the joint responsibility of Thurstone and Jones, who visited Chapel Hill periodically to review progress on ongoing projects.

In 1953, the Thurstones moved into their newly constructed home a few blocks from the campus in Chapel Hill. This home was built with a special seminar room off the living area, and the practice of inviting speakers and guests for evening seminars, so long maintained at Chicago, now was reinstated in this new location.

After only three years at UNC, L. L. Thurstone died at age 68 in September of 1955. Mrs. Thurstone, Professor of Education, agreed to become Acting Director of the UNC Psychometric Laboratory, but only until work on current research contracts was complete. (She remained a research associate in the laboratory for nearly four decades, and she continued to develop, revise, and publish—through SRA—reading materials for schools until a year or two before her death in 1993.)

The Thurstone legacy at Chapel Hill has continued to flourish. In 1957, Lyle V. Jones moved from Chicago to UNC to direct the Psychometric Laboratory. From that time until the present, the laboratory has maintained active programs of research, has hosted nearly 50 visiting faculty and postdoctoral fellows, and has graduated about 100 Ph.D. recipients, many of whom continue to occupy prominent positions in academia, as well as in corporate and government research agencies.

In 1967, the Psychometric Laboratory relinquished occupancy of Nash Hall to be housed in a wing of newly renovated Davie Hall with the Department of Psychology. In 1977, to celebrate its 25th anniversary and to recognize its founder, it was renamed the L. L. Thurstone Psychometric Laboratory, and it hosted the annual meeting of the Psychometric Society. Jones served as Director from 1957 to 1974 and from 1979 to 1992. John B. Carroll was Director from 1974 to 1979, and David Thissen succeeded Jones in 1992. In 2002, to mark the laboratory's 50th anniversary, the Psychometric Society again held its annual meeting on the Chapel Hill campus. Robert MacCallum became Director of the laboratory in 2003.

Currently, laboratory personnel engage in research on test theory and practice, on other facets of psychological and educational measurement, on mathematical models of human behavior, and on applied statistics. The laboratory provides both graduate and undergraduate courses on these topics, and awards the MA and Ph.D. degrees in psychometrics/quantitative psychology.

See Also the Following Articles

Factor Analysis • Psychometrics of Intelligence

Further Reading

Adkins Wood, D. (1962). *Louis Leon Thurstone.* Educational Testing Service, Princeton, NJ.

Guilford, J. P. (1957). Louis Leon Thurstone, 1987–1955. *Natl. Acad. Sci. Biograph. Memoirs* **30,** 349–382.

Gulliksen, H. (1956). A tribute to L. L. Thurstone. *Psychometrika* **21,** 309–312.

Jones, L. V. (1998). L. L. Thurstone: A vision of psychology as a quantitative rational science. In *Portraits of Pioneers in Psychology* (G. A. Kimble and M. Wertheimer, eds.), Vol. III, pp. 84–104. American Psychological Association and L. Erlbaum Associates, Washington, D.C.

Jones, L. V., and Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. *J. Appl. Psychol.* **39,** 31–36.

Luce, R. D. (1977). Thurstone's discriminal processes fifty years later. *Psychometrika* **42,** 461–489.

Thunstrőm, L. L. (1905). How to save Niagara. *Sci. Am.* **93,** 27.

Thurstone, L. L. (1924). *The Nature of Intelligence.* Harcourt Brace and Co, New York.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *J. Educ. Psychol.* **16,** 433–448.

Thurstone, L. L. (1927). A mental unit of measurement. *Psychol. Rev.* **34,** 415–423.

Thurstone, L. L. (1927). The method of paired comparisons for social values. *J. Abnorm. Soc. Psychol.* **4,** 384–400.

Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychol. Rev.* **35,** 175–197.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychol. Rev.* **36,** 222–241.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychol. Rev.* **38,** 406–427.

Thurstone, L. L. (1935). *The Vectors of Mind.* The University of Chicago Press, Chicago, IL.

Thurstone, L. L. (1937). Psychology as a quantitative rational science. *Science* **85**, 228–232.

Thurstone, L. L. (1938). *Primary Mental Abilities. Psychometric Monograph No. 1.* The University of Chicago Press, Chicago, IL.

Thurstone, L. L. (1938). *Tests for Primary Mental Abilities, Experimental Edition.* American Council on Education, Washington, D.C.

Thurstone, L. L. (1947). *Multiple-Factor Analysis.* The University of Chicago Press, Chicago, IL.

Thurstone, L. L. (1959). *The Measurement of Values.* The University of Chicago Press, Chicago, IL.

Thurstone, L. L., and Thurstone, T. G. (1942). *The Chicago Tests of Primary Mental Abilities.* American Council on Education, Washington, D.C.

Time Sampling

David G. Steel
University of Wollongong, Wollongong, NSW, Australia

Glossary

composite estimation Method of estimation that uses data for the current and previous time periods and gives different weights to matching and non-matching sample units.
longitudinal survey A survey that uses a sample in which the same units are included for several time periods.
panel survey Equivalent to a longitudinal survey.
repeated survey A survey conducted at different times with no attempt to have sample units in common.
rotating panel survey A panel survey in which a proportion of units are removed from the survey at some time periods and replaced by other units.
rotation pattern The pattern of inclusion of sample units over time.

Sampling over time enables researchers, analysts and decision makers to monitor, analyze, and understand social processes through the estimation and analysis of changes in variables of interest. In addition to the usual sample design issues considered for a sample used for one time period, the design of a time sampling scheme needs to consider the frequency of sampling and the spread and pattern of inclusion of selected units over time. A key issue is whether to use overlapping or non-overlapping samples over time. For overlapping samples, the precise pattern of overlap must be designed. Repeated, panel, and longitudinal surveys, rotating panel surveys, split panel surveys, and rolling samples are important examples of the application of time sampling. Factors that affect the design of a sample over time are the key estimates to be produced, the type and level of analyses to be carried out, cost, data quality, and reporting load. The interaction between the design of the sample in time and the other features of the design, such as stratification and cluster sampling, also needs to be decided. Time series may be produced and analyzed, which may involve seasonal adjustment and trend estimation.

Sampling and Surveys

Information obtained from samples selected using probability sampling methods can be used to provide estimates of characteristics of a population of interest and to analyze relationships between variables. Probability sampling involves methods in which members of a population have a known, non-zero probability of selection. Simple random sampling, probability proportional to size sampling, stratification, and cluster and multi-stage sampling are common probability sampling methods. Using these methods, population quantities such as means, totals, proportions, medians, and other quantiles that describe the current population can be estimated. Standard errors can also be estimated and used to make inferences, for example, by constructing confidence intervals. Relationships between variables can be analyzed, for example, by estimating linear or logistic regression coefficients. Samples are also useful because they allow estimates and analyses for subpopulations, provided subsamples of sufficient size have been selected.

Sampling is often used in surveys of human populations, collecting information on social topics such as health, income, expenditure, employment, crime, education, opinions, and attitudes. Samples of people are often obtained by selecting a sample of dwellings and including the households and people in the selected dwellings in the sample. Coverage rules are used to associate people with households and dwellings. Sampling is also used for surveys of other entities such as hospitals, schools, and businesses. Sampling of physical units, such as areas of land, can also be used.

Repeated Surveys

A particular survey may be conducted only once, or it may be repeated on several occasions on an irregular, regular, or periodic basis. Therefore, surveys can be classified as either one-off or repeated surveys. A one-off survey can provide cross-sectional estimates referring to the population at a particular time. A repeated survey also provides these estimates, but also enables estimates of changes to be calculated for the population.

The frequency of sampling depends on the purpose of the survey. Monitoring and detecting important changes will usually be a key reason for sampling in time. Common frequencies for surveys are monthly, quarterly, and annual, although even more frequent sampling may be adopted, for example, in opinion polls leading up to an election or monitoring TV ratings. How quickly changes are likely to occur and how quickly any associated decisions are needed are factors in deciding the frequency of sampling. The budget available is also a consideration. Sampling should not take place so often that the sample is registering unimportant short-term movements of no practical interest.

In any survey, the collection or interview period and reference period have to be considered. The interview period is the time period in which the sample is to be interviewed or data collected, and the reference period is the period used to define the variables about which information is collected. For example, a survey may be conducted over a particular 4-week period, collecting information on visits to the doctor for a reference period of the previous 12 months. The spread of the sample over the interview period should ideally be balanced over the important spatial dimensions used in the sample design. For example, a health survey may be conducted over an entire year to allow for seasonal effects, and the sample in each month should be of the same size and design. The design of the sample should be taken into account when deciding on the sample in each month. Ideally, the monthly sample should replicate the annual sample. In a stratified, multistage design involving the selection of primary sampling units (PSUs), the sample should include each stratum and PSU in each month. However, cost considerations may lead to each PSU being included in only one month or time period. The sample weighting used in estimation may be implemented by month if the population varies across the year or if the sample size or composition varies considerably across the year.

The reference period is a fundamental part of the definition of the variable. Having a long reference period may increase the number of episodes or incidents included in the survey, but the impact of telescoping, when events are incorrectly reported as having occurred in the reference period, and other recall errors must be taken into account. The reference period used will depend on the specific variable. For some variables, a 12-month reference period might be feasible, whereas for other variables a 1-day reference period might be appropriate. Some variables are even defined at the time of interview, for example, an opinion.

In deciding on the time periods for which data are collected a decision is being made about which time periods to sample. A sample or census of time periods may be used. For example, in a monthly survey all weeks may be included, or a single week may be used to represent the month. This aspect of the sampling needs to be considered. The possible impact of variation and cyclic patterns within the month may be relevant.

Time also has to be considered in the definition of the sampling frame from which the sample is selected, and also the definition of sampling units such as a household. Changes in the population can be important contributors to the change in the variables of interest over time. If estimates referring to the population at each time period are required, it is important that the population frame is updated to incorporate changes in the population as quickly as possible. The sample should also be updated to give new units a chance of selection and to remove defunct units whose presence may affect the sampling errors. Systems to update the sampling frame and sample therefore must be developed. In household surveys, this is sometimes done by developing a master sampling frame that is updated regularly to add new housing and to reflect other significant changes to the population. This frame also can be designed to implement a rotation pattern if necessary, by dividing the frame into rotation groups, and controlling overlap between different surveys using the same frame. For surveys of businesses and institutions, the list or register must be maintained and the sample updated. Rotation of the sample for a particular survey and overlap between the samples selected for different surveys can be controlled using permanent random number sampling.

Sampling in time may be used within a particular survey through multiphase or double sampling. An initial relatively large sample is selected, and some basic information that is relatively cheap to obtain is collected. A smaller subsample, in which more detailed information is obtained, is then selected. The initial sample is used to provide information used in stratification of the subsample or as auxiliary information in ratio or regression estimation. A particular case is when the variables in the subsample are conceptually the same as that in the initial sample, but the information is collected using more reliable methods. The subsample then provides information to adjust for the measurement or misclassification errors in the first-phase data. A multiphase design can be used to take a subsample of non-respondents to the first phase,

which is followed intensely. The resulting data can then be combined with the data obtained initially in an attempt to reduce the bias due to non-response.

Panel and Longitudinal Surveys

In a panel survey, an initial sample is selected and information is collected on several occasions. This can be done to provide estimates of change for variables for which information is collected at each occasion. A panel survey can also used to provide estimates for different variables over time. Cost savings often arise because there are higher setting-up costs for the first time a person is included in the survey than on subsequent occasions. Television rating surveys are an example of a panel survey; Internet surveys can also use this approach. Panel surveys allow longitudinal data analysis and so are also called longitudinal surveys.

A distinction can be made between a repeated survey and a longitudinal survey. In a longitudinal survey, an initial sample is selected, and then at each occasion that the survey is conducted, an attempt is made to include the members of the initial sample. The different time periods for which units are included are sometimes called waves. In a repeated survey, there is not necessarily any overlap of the sample for the different occasions. A longitudinal survey permits analysis of changes at a micro level, ultimately at the level of an individual.

Following rules need to be developed for a longitudinal survey and will be influenced by the objectives of the survey. At one extreme, people would not be followed when they leave a selected dwelling, so the panel unit is the dwelling. At the other extreme, people would be followed wherever they go, unless they die or leave the country. For cost reasons, in a cluster sample, people may be followed only if they move within a PSU.

In a household panel survey, households are often retained in the survey when they change dwellings, and individuals are retained in the sample even if they change households. Although the initial sample can provide estimates for the population existing at that time, the sample cannot provide unbiased estimates for the current population, that is, cross-sectional estimates, unless it is updated to include new entrants to the population. A decision has to be made as to whether such estimates are required. If cross-sectional estimates for each period are of interest, strategies must be adopted to keep the sample representative of the population at each time period. This implies adding to the sample so that population changes are reflected in the sample.

Even if cross-sectional estimates are not required, there may be interest in household composition or characteristics and their association with individuals. Information about the households to which sample members move should then be obtained.

Some examples of longitudinal surveys are the British Household Panel Survey (United Kingdom), the Survey of Family, Income and Employment (New Zealand), the National Longitudinal Surveys (United States), the Households, Income and Labour Dynamics in Australia Survey (Australia), and the Survey of Income and Program Participation (United States).

Rotating Panel Surveys

A longitudinal survey is a form of panel survey. Longitudinal surveys are specifically developed to permit analysis of changes at the individual level. A panel survey may use a panel at the dwelling level, so that when people or households move they are not followed, and people moving into a dwelling in the panel may be included in the survey. Rotating panel surveys also use a sample that is followed over time, but in general the main focus is on estimates at aggregate levels.

When the emphasis is on estimates for the population and possibly subpopulations, an independent sample may be used on each occasion, which is often the case when the interval between the surveys is quite large. An alternative is to try to use the same sample at each occasion, with some additions to ensure that the sample estimates refer to the current population. For regular monthly or quarterly surveys, the sample is often designed so that there is considerable overlap in the sample between successive surveys. Overlap in the sample reduces the sampling variance of estimates of change and reduces costs. Sampling variances on estimates of change in the variables of interest are reduced because the variation due to including different people is reduced. The reduction in variances depends on the correlation of the variable at the individual level over time and the degree of sample overlap. If the correlation is low, then the reduction is small and is not a major consideration. The correlation must be positive for this reduction to apply. A negative correlation will increase sampling variances; such cases are not common, but can occur.

These considerations would lead to maximizing the sample overlap at each time period at almost 100%, with the only change in the sample arising from the need to update the sample to represent people moving in and out of the population in the scope of the survey. However, such a design would lead to selected people being included in the survey indefinitely. In practice, a limit must be placed on how many times a person is surveyed to spread the reporting load and maintain response rates and the quality of the reported data. When deciding the degree of sample overlap, these considerations must be balanced.

An overlapping sample design can be implemented using a rotation pattern or design. A rotation pattern can be designed to efficiently manage the sample over time. Rotation patterns can be developed that have the same proportion of the sample in common between any two time periods the same time apart and the same proportion of sample rotated out and into the sample at each period. The rotation sample design should ensure that the cross-sectional estimates are unbiased, while reducing costs and sampling variances on important estimates of change. A rotation pattern would usually ensure that at each time point the sample is balanced according to the number of times a person has been included in the survey. This can be important because of the potential effect that the number of times a person has been included in the survey has on the data reported.

A rotation sampling design can be implemented using rotation groups and panels. The sample will consist of several rotation groups. A panel is the set of selected units that enter and leave the sample at the same time. When a panel leaves the sample it is replaced from the same rotation group. For example, in the Australian Labour Force Survey, the PSUs are allocated to eight rotation groups. In a particular month, the dwellings in one of the rotation groups are rotated out of the survey and replaced by an equivalent sample of dwellings in the same PSU.

A further aspect of the design is the level of information collected, which is the number of time periods for which information is collected on a particular occasion. For example, in a monthly survey, information may be collected from a unit for the current month and for the previous month. This approach is used in the U.S. Retail Trade Survey.

There are many different rotation patterns in use, and many that can be considered. Consider a monthly survey. The simplest rotation pattern is when a unit is included for a months. A more general class of rotation patterns is when a unit is included initially for a months, then leaves the sample for b months, and then returns to the sample for a further a months. This pattern is repeated until the unit is included for a total of c months. This can be denoted as an $a{-}b{-}a(c)$ rotation pattern. For example, the U.S. Current Population Survey uses a $4{-}8{-}4(8)$ rotation pattern, whereas the Australian Labour Force Survey uses an in-for-8 rotation pattern, which can be denoted 8(8). The Canadian Labour Force Survey uses an in-for-6 rotation pattern. These surveys use one level, so that information is collected referring to 1 month. More generally, a pattern of the form $a_1{-}b_1{-}a_2{-}b_2{-}\cdots a_m(c)$ can be considered, in which the number of months included and excluded from the survey varies.

By using sampling in time, an analysis of changes can be carried out. Consider a key variable of interest that is estimated for time period t by y_t. The simplest analysis of change is the estimate of one period change, $y_t - y_{t-1}$. In a monthly survey, this corresponds to a 1-month change. For a survey conducted annually, this corresponds to annual change. In general, the change s time periods apart can be estimated, using $y_t - y_{t-s}$. For a monthly survey, looking at 3-month and 12-month changes can be useful. Because $\mathrm{Var}(y_t - y_{t-s}) = \mathrm{Var}(y_t) + \mathrm{Var}(y_{t-s}) - 2\mathrm{Cov}(y_t, y_{t-s})$, a positive covariance between the estimates will reduce the variance, which can be achieved through sample overlap. If comparisons are made with time periods for which there are no sample units in common, then the variance of the estimate of change will be the sum of the variances, which will often be approximately twice the variance of the estimate of the level for a particular time period. These considerations result in designing the sampling so that there is overlap between the samples for time periods between which the movements are of major interest. So if there is strong interest in monthly movement, there should be high sample overlap between successive months. If there is also interest in changes 12 months apart, then consideration should be given to designs that induce sample overlap at this time-lag. However, for many variables the correlation 12 months apart may not be high enough for there to be appreciable gains.

Overlap in the sample may occur at different stages in a multistage design. In a cluster sample, even if there is no overlap at the individual level, there can be some small gains for estimates of movement by having overlap at the PSU level. Rotation is often carried out within PSUs for cost reasons, and a rotation group consists of a sample of PSUs, so that each PSU is allocated to a particular rotation group.

Averaging of estimates can be used in an attempt to produce more stable estimates when the original estimates have high sampling variances, for example, for small sub-groups or domains in the population. A particular case is estimates for small geographic areas. However, averaging over time changes the length of the time period to which the estimate refers and will hide any variation within the period over which the average is calculated. Time series methods are available to help combine data across time and space to produce small area estimates from rotating panel surveys.

For estimation of averages of estimates, positive correlation between the survey estimates involved will increase the sampling variance. It is better to average uncorrelated estimates, which can be obtained from independent or non-overlapping samples. If both averages and differences of estimates are of interest, the relative importance of each type of estimate must be considered and the impact of different options assessed on both types. To assess the impact of different rotation patterns on various estimates, some information or assumptions about the covariances involved are needed.

A longitudinal survey can be used to provide estimates of changes at aggregate levels, but these estimates refer to the population at the time of the initial sample selection, unless attempts have been made to add to the sample to make it representative of the current population. However, the main purpose of a longitudinal survey is to enable estimates of changes at the person or household level.

In a rotating panel survey, the panel aspect is often implemented at the dwelling level, which implies that people and households are not followed when they leave a selected dwelling. People and households moving into a selected dwelling are included in the survey. This approach is suitable when the main objective is to provide unbiased aggregate estimates.

In a rotating panel survey, the main focus is on aggregate estimates of change. However, any overlapping sample can also be used to analyze change at the micro level. For example, from the matched sample, a table can be produced showing the change of a variable between two time periods. An important example is when a table of change in status is produced, which is referred to as a gross flows table. It is possible to create longitudinal data from rotating panel surveys, but the length of the total time period and the time interval between observations are determined by the rotation pattern used. Also, the resulting sample of individuals for which longitudinal data are available will be biased against people who move permanently or are temporarily absent.

An alternative to a rotating panel survey is a split panel survey, which involves a panel survey supplemented on each occasion by an independent sample. Such a design permits longitudinal analysis from the panel survey for more periods than would usually be possible in a rotating panel design, but cross-sectional estimates can also be obtained from the entire sample.

Rolling Samples

In deciding on the sample design, in general the three dimensions of space, time, and variables need to be considered. A survey may be conducted continuously, but the sample size in any time period may not be sufficient to provide reliable estimates for that period, at least for subnational estimates. However, by cumulating samples over several time periods, reasonably reliable estimates may be produced. In this approach, sample overlap is detrimental. The sample design can be developed so that it is effectively a rolling sample with non-overlapping samples that over time cover many areas and eventually all areas. This approach can be useful in producing sub-national and small area estimates. A major example of this approach is the American Community Survey.

A related approach is rolling estimates. For example, in the UK Labour Force Survey, a non-overlapping sample is interviewed in each week of the quarter. Each month, estimates based on an average of the latest 13 weeks are produced.

Estimation

In a repeated survey, estimates can be calculated independently using standard sample weighting methods. When there is sample overlap implemented through a rotation sample design, it is possible to exploit the correlation structure for different rotation groups to produce estimates of levels and changes with smaller sampling variances. These methods are called composite estimators and effectively weight the common and non-overlapping samples differently.

In composite estimation, the sample for the previous time periods is used along with the sample for the current period. In its simplest form, the estimate for the current period is obtained by updating the estimate of the previous period using an estimate of the change that has occurred in which the matched and non-matched samples are given different weights. More generally, the estimates for each rotation group can be determined and combined in an efficient manner, taking into account the correlation structure of these estimates over time. However, issues of time in survey bias need to be considered.

Composite estimation methods have mostly been applied in monthly labor force surveys. Regression composite estimators have been developed and applied by several national statistics institutes to combine the benefits of composite estimation and regression estimation, which is a technique for exploiting extra information on auxiliary variables.

In assessing the potential gains from using composite estimation, attention is usually focused on the estimate of level for the most recent period and the movement between the two most recent time periods. The gains arising from composite estimation depend on the degree of overlap and the individual level correlation. The gains for estimates of levels are greatest when the degree of overlap is moderate and the correlation is high. For estimates of movement, high sample overlap is still preferred.

Methods of estimation assuming a time series structure for the population mean or total have also been developed. Estimation for panel surveys involve weighting the sample to provide population estimates. For the first wave, this will be relatively straightforward, but for subsequent waves, different weights may be required depending on the population for which estimates are required. Auxiliary information, such as population benchmarks used in estimation, may also need to be

updated if estimates for the current population are required. In developing weights for estimation, the following rules must be considered in determining a person's probability of selection, because after the first wave they can be included in the survey in more than one way.

Analysis

Producing estimates at regular intervals enables trends to be assessed. Various methods of estimating trends are available using model-based or filter-based methods. For surveys producing monthly or quarterly estimates, seasonal adjustment may be used to remove the impact of regular factors operating at different times of the year. Producing seasonally adjusted estimates can also assist in assessing the underlying direction or trends in the series. The rotation pattern chosen affects the correlation structure of the sampling errors over time, which can affect the properties of seasonally adjusted and trend estimates.

Analysis of net change using aggregate estimates may hide important gross changes occurring at the individual level, which may be revealed from longitudinal data. Longitudinal analysis can help determine the relationships between variables and look at causes by examining the temporal sequences of events. Longitudinal survey data can be used to undertake a variety of analyses, including survival analysis, event history analysis, and analysis of transition probabilities. Multilevel models that take account of the repeated nature of the data are being increasingly used. Longitudinal surveys with panels starting in different periods permit the disentangling of cohort/age/period effects.

Data Quality Issues

Including people in a survey for several occasions can affect the quality of the information reported. Non-response is a source of error in any survey, but in a longitudinal survey there is usually an accumulation of non-response over the waves, which can lead to attrition bias. There will also be cases in which a particular sample individual does not respond for one or more of the waves, affecting any analysis based on a set of data for all waves. Panel surveys have the advantages that the interviews at each wave can act as a boundary for the collection of data and reduce the impact of telescoping. However, conditioning and learning effects may occur.

See Also the Following Articles

Survey Design • Survey Questionnaire Construction • Surveys

Further Reading

Bell, W. R., and Hillmer, S. C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodol.* **16,** 195–215.

Binder, D. A. (1998). Longitudinal surveys: Why are these surveys different from all other surveys? *Survey Methodol.* **24,** 101–108.

Binder, D. A., and Dick, J. P. (1989). Modelling and estimation for repeated surveys. *Survey Methodol.* **15,** 29–45.

Binder, D. A., and Hidiroglou, M. A. (1988). Sampling in time. In *Handbook of Statistics: Volume 6: Sampling* (P. R. Krishnaiah and C. R. Rao, eds.), pp. 187–211. Elsevier Science, Amsterdam.

Chambers, R. L., and Skinner, C. J. (2003). *Analysis of Survey Data.* John Wiley & Sons, New York.

Duncan, G. J., and Kalton, G. (1987). Issues of design and analysis of surveys across time. *Intl. Stat. Rev.* **55,** 97–117.

Fuller, W. A. (1990). Analysis of repeated surveys. *Survey Methodol.* **16,** 167–180.

Holt, D., and Skinner, C. J. (1989). Components of change in repeated surveys. *Intl. Stat. Rev.* **57,** 1–18.

Kalton, G., and Citro, C. F. (1993). Panel surveys: Adding the fourth dimension. *Survey Methodol.* **19,** 205–215.

Kalton, G., Kasprzyk, D., and McMillen, D. B. (1989). Nonsampling errors in panel surveys. In *Panel Surveys* (D. Kasprzyk, G. Duncan, and G. Kalton, eds.), pp. 249–270. John Wiley & Sons, New York.

Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P. (eds.) (1989). *Panel Surveys*. John Wiley & Sons, New York.

Kish, L. (1987). *Statistical Design for Research.* John Wiley & Sons, New York.

Kish, L. (1998). Space/time variations and rolling samples. *J. Official Stat.* **14,** 31–46.

McLaren, C. H., and Steel, D. G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. *Survey Methodol.* **26,** 163–172.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *J. Bus. Econ. Stat.* **9,** 163–175.

Singh, A. C., Kennedy, B., and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodol.* **27,** 33–44.

Steel, D. (1997). Producing monthly estimates of unemployment and employment according to the international labour office definition (Disc: p33–46). *J. Roy. Stat. Soc. A* **160,** 5–33.

Time Series Analysis in Political Science

Harold D. Clarke
University of Texas, Dallas, Richardson, Texas, USA

Jim Granato
National Science Foundation, Arlington, Virginia, USA

Glossary

ARIMA model Autoregressive, integrated, moving average model of a time series variable.
autoregessive, distributed lag model A time series model with a lagged endogenous variable and one or more independent variables.
cointegrated variables Time series variables that form stationary linear combinations.
error correction model A model that specifies a long-term cointegrating relationship among two or more time series variables.
GARCH model Autoregressive, moving average model of the conditional heteroskedasticity in the stochastic error component of a model of the mean of a time series process.
Granger causality If the history of one time series variable can improve the explanation of a second variable beyond what can be explained by the history of the latter variable, the first variable "Granger causes" the second variable.
nonstationary variable A time series variable with a nonconstant mean, variance, or covariance.
vector autoregression A multiequation reduced-form model with two or more endogenous time series variables.
weak exogeneity An independent variable in a model is weakly exogenous if parameters of interest rest solely in that model and there are no cross-equation restrictions between a model for the independent variable and the model for the dependent variable.

Time series analysis in political science was invigorated in the 1970s by a combination of growing methodological sophistication and growing interest in the dynamic interplay of economic and political processes. Time series analyses in political science employ many of the same tools used by economists in their empirical work. Historically, political scientists have relied heavily on ordinary least squares and generalized least squares regression techniques for the estimation of parameters in single-equation models. However, starting in the late 1970s, Box-Jenkins ARIMA models became increasingly popular, particularly among researchers concerned with the impact of salient events (e.g., energy price shocks, policy interventions, and foreign crises or wars) on variables such as presidential approval and governing party support. Since the early 1990s, an increasing number of studies have marshalled the methodology of cointegration and error correction for investigating the dynamics of nonstationary variables. In a related development, researchers have begun to employ the concepts of fractional integration and, very recently, fractional cointegration. Perhaps reflecting their discipline's widespread lack of interest in forecasting, political scientists have paid little attention to vector autoregression (VAR) or ARCH models. An important current development is the EITM initiative, which attempts to forge strong linkages between the development of formal models and empirical testing with time series statistical methods.

Traditional Time Series Analysis

Research using time series data in political science typically has utilized many of the same regression techniques as are employed to analyze cross-sectional data. The vast majority of these traditional time series analyses

have considered single-equation models such as the following:

$$Y_t = \beta_0 + \sum \beta_{1-k} X_{1-k,t-i} + \varepsilon_t, \quad (1)$$

where Y_t is the dependent variable at time t, X_{t-i} are 1 to k independent variables at time $t-i$, β_0 is constant, β_{1-k} are the parameters associated with variables X_{1-k}, and ϵ_t is the stochastic error term $\sim N(0, \sigma^2)$.

For a model such as Eq. (1), the possible (non)stationarity of the variables is ignored, and ordinary least squares (OLS) is employed to estimate the values of the parameters β_0, β_{1-k}. The effects of the X's may be specified to occur simultaneously (i.e., at time t or with a lag i). Also, as in analyses of cross-sectional data, inferences regarding the statistical significance of the β's are made by calculating t ratios (i.e., β/s.e.). When doing diagnostic tests on such regression models, particular attention is given to the possibility that the stochastic errors (ϵ's) are correlated [i.e., $cov(\epsilon_t, \epsilon_{t-i}) \neq 0$]. Correlated errors do not bias parameter estimates but affect standard errors and, therefore, pose a threat to inference by affecting the size of the t ratios. The standard test for correlated errors has been the Durbin–Watson test, which tests only for first-order autocorrelation in the residuals of the estimated regression Eq. (1). If the null hypothesis that the residuals do not suffer from first-order autocorrelation is rejected by this test, the conventional approach is to conclude that the errors are generated by the following process:

$$\epsilon_t = \rho \epsilon_{t-1} + v_t, \quad (2)$$

where ρ captures the relationship between temporally adjacent errors, and v_t is a "well-behaved" (uncorrelated) error proces $\sim N(0, \sigma^2)$. This (assumed) relationship between the errors is treated as a "nuisance" to be "corrected." The alternative possibility, that the correlation among the residuals represents the result of model misspecification, is not considered.

The correction employed is a form of generalized least squares (GLS) that involves multiplying both sides of the model of interest by the "quasi-differencing" operator $(1-\rho L)$, where L is a backshift operator such that $L^k y_t = y_{t-k}$. This model is then subtracted from the original one. For example, for a model with a single right-hand-side variable, the result is

$$Y_t - \rho Y_{t-1} = \beta_0 - \rho\beta_0 + \beta_1 X_t - \rho\beta_1 X_{t-1} + \epsilon_t - \rho\epsilon_{t-1}. \quad (3)$$

The error process for the transformed model is $\epsilon_t - \rho\epsilon_{t-1} = v_t$, which, by assumption, is uncorrelated. Since ρ is unknown, it must be estimated from the data. Various techniques may be used for this purpose, and the resulting procedures are known as feasible GLS.

Political scientists adopting this approach to addressing the threat to inference have often failed to recognize that they have, in effect, respecified their original model in autoregressive distributed lag form and imposed a common-factor restriction $(1-\rho L)$. This may be seen by rewriting Eq. (3) as

$$(1-\rho L)Y_t = (1-\rho L)\beta_0 + (1-\rho L)X_t + (1-\rho L)\epsilon_t. \quad (4)$$

As Hendry emphasizes, the warrant for this restriction should be determined empirically, rather than simply assumed. The vast majority of time series analyses in political science have not done so. By failing to recognize that autocorrelated residuals do not necessarily imply autocorrelated errors, such analyses risk model misspecification.

Although many political scientists continue to use GLS procedures, it is increasingly common to attempt to capture the dynamics in a time series by specifying an autoregressive, distributed lag model that includes a lagged endogenous variable Y_{t-1}:

$$Y_t = \beta_0 + \gamma Y_{t-1} + \sum \beta_{1-k} X_{1-k,t-i} + \epsilon_t. \quad (5)$$

A model such as Eq. (5) may be specified initially on theoretical grounds, or after the analyst finds evidence of first-order autocorrelation in Eq. (1), a common practice is to use Eq. (5). In any event, the presence of the lagged endogenous variable Y_{t-1} means that the analyst is hypothesizing, either explicitly or implicitly, that the effects of all of the X variables are distributed through time and that all of these effects decline at exactly the same rate. That rate is γ, the coefficient on Y_{t-1}. For example, the impact of $\beta_1 X_{1t}$ in Eq. (5) is β_1 at time t, $\beta_1\gamma$ at time $t+1$, $\beta_1\gamma^2$ at time $t+2$, etc., and the long-term (asymptotic) impact of X_1 is $\beta_1/(1-\gamma)$. Clearly, the assumption that the effects of all X's evolve in exactly the same way is very strong. ARIMA intervention and transfer function models considered later relax this assumption.

ARIMA Models

ARIMA models constitute another major set of tools for analyzing time series data of interest to political scientists. ARIMA models are the product of pioneering work by Box and Jenkins. Reflecting widespread discontent with the forecasting failures of large multiequation models in the "Cowles Commission" tradition, Box and Jenkins proposed the radical alternative of forecasting a variable using only its past values. The ARIMA acronym for the models they developed has three parts: (i) AR for autoregressive, (ii) I for integrated, and (iii) MA for moving average. ARIMA models are specified as a combination of the three parts.

In a sharp departure from what had been prevalent practice in earlier regression-based approaches to time series analysis, Box and Jenkins emphasized part (ii), arguing that it was crucial to determine whether the

time series variable under consideration was stationary. Stationary variables have the following properties:

a. $E[y] = \bar{y}$; constant mean
b. $E(y - E[y]) = \sigma^2$; constant variance
c. $Cov(y_t, y_{t-k}) = \gamma_k$: constant covariance at lag k for all t.

Nonstationary variables may be the result of two different data-generation processes. In one of these processes, the nonstationarity reflects the presence of a deterministic component, e.g.,

$$y_t = \lambda T + \epsilon_t, \tag{6}$$

where λ is a parameter, ϵ_t is a stochastic shock, and T is a time counter. Historically, this type of model had been synonymous with the notion of "trend" as used by political scientists (and economists), and it was assumed that trending variables could be rendered stationary by simply regressing them on "time."

However, Box and Jenkins believed that nonstationarity in social science data was likely the result of another type of data-generating process, one that produced a "nondiscounted" accumulation of stochastic shocks. The simplest such model is the random walk:

$$y_t = 1.0 y_{t-1} + \epsilon_t. \tag{7}$$

In Eq. (7), the coefficient for y_{t-1} is 1.0, thereby ensuring that the effects of successive shocks, ϵ_t, are not discounted over time but rather continue at their full, time t, value. The result is a time series variable with an asymptotically infinite variance. Adding a constant β_0 to Eq. (7) yields a model with a deterministic trend component, $\beta_0 T$. This may be seen by recursive substitution in Eq. (7), which gives

$$y_t = \beta_0 T + y_0 + \sum_{i=1} \epsilon_i \tag{8}$$

In Eq. (8), y_0 is the initial value of y. The implication of assuming that nonstationarity is the result of such a data-generating process is that stationarity may be achieved by differencing a variable one or more times (e.g., for the random walk, $y_t - y_{t-1} = \epsilon_t$), where ϵ_t is a stationary "white nose" (uncorrelated) process. Integration, then, is the number of times a variable must be differenced to achieve stationarity. In the Box-Jenkins methodology, the initial step is to use diagnostic techniques (graphs, autocorrelation functions, and unit-root tests) to determine whether a variable is nonstationary. If the variable is nonstationary, it is differenced, and the diagnostics are repeated. If it is still nonstationary, it is differenced again, and diagnostics are performed again. After the variable is rendered stationary by differencing once or more, additional diagnostic procedures are used to identify the presence of AR and/or MA components in the data-generating process.

A pure autoregressive process is one in which a series is a function of one or more lags of itself, plus a contemporaneous stochastic error. For example, a first-order autoregressive process is

$$y_t = \phi_1 y_{t-1} + \epsilon_t, \tag{9}$$

where ϕ_1 is a parameter that will have an absolute value <1.0 if y_t is stationary. The general pth-order AR [AR(p)] model is a straightforward extension:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t. \tag{10}$$

Pure moving average processes reflect the operation of short-term shocks. Suppose ϵ_t is a purely random process with mean $\mu = 0$ and variance σ_ϵ^2. A process y_t is said to be a first-order moving average process if

$$y_t = \epsilon_t - \theta_1 \epsilon_{t-1}. \tag{11}$$

By taking a nonzero value, the parameter θ_1 indicates that a portion of a shock at time $t-1$ continues to influence y at time t. The general moving average process of order q [MA(q)] is

$$\begin{aligned} y_t &= \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q} \\ &= 1 - \theta_1 L - \theta_2 L^2 - \theta_3 L^3 - \cdots - \theta_q L^q \mathbf{XX} \epsilon_t \end{aligned} \tag{12}$$

An MA(∞) $= \theta(L) = 1 + \mathbf{XX}_{j=1}^{\infty} - \theta_j L^j$. The MA process does not depend on time and the ϵ_t's are independent. The mean is constant and the MA process is strictly stationary when $(\epsilon_t \sim N\mathbf{XX}\mu, \sigma^2\mathbf{XX})$.

Some time series exhibit mixed AR/MA behavior. Thus, the general AR(p), MA(q) model is

$$\begin{aligned} y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} \\ + \cdots + \theta_{t-q} \epsilon_{t-q}. \end{aligned} \tag{13}$$

Using the lag operator (L), the model may be written as $y_t = \theta(L)/\phi(L)\epsilon_t$, where $\phi(L)$, $\theta(L)$ are polynomials of order p and q, respectively. Taking into account the possibility of differencing (one or more times) to eliminate nonstationarity in y_t, this general ARIMA model becomes

$$(1 - L)^d y_t = \theta(L)/\phi(L)\epsilon_t, \tag{14}$$

where d is the number of differences required to achieve stationarity. ARIMA models are often described in terms of (p, d, q), where p is the number of autoregressive parameters, d is the number of differences needed for stationarity, and q is the number of moving average parameters. For example, a first-order MA model for a variable requiring one difference to attain stationarity would be a (0, 1, 1) model. The specification of ARIMA

models thus involves determination of p, d, and q, both generally and, possibly, at seasonal spans. Box and Jenkins discuss how to use autocorrelation and/or partial autocorrelation functions to determine plausible values for p, d, and q. The key point is that analytic results indicate what the nature of the autocorrelation and autocorrelation functions should be. For example, a stationary first-order autoregressive process will have an autocorrelation function that decreases at a geometric rate across successive lags and one significant correlation (a "spike" in Box-Jenkins terminology) at lag 1 in the partial autocorrelation function. Contemporary practice for determining d involves the use of unit-root tests.

After plausible values of p, d, and q are specified, parameters in the ARIMA model are estimated, and then a battery of diagnostic tests are performed. Since forecasting is a principal purpose of the construction of univariate ARIMA models, diagnostics often include out-of-sample forecasting performance. In keeping with the strong empiricist spirit that infuses ARIMA methodology, models deemed inadequate are respecified, reestimated, and rediagnosed. This procedure continues until the analyst is satisfied with the model's performance. Since more than one model may have satisfactory diagnostics, choice among rivals will be made on the basis of model selection criteria (Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)) and relative forecast accuracy.

Intervention and Transfer Function Models

ARIMA models may be augmented by the inclusion of dummy (0-1) and continuous right-hand-side variables. The former are known as interventions and the latter as transfer functions (a term from chemical engineering). In political science applications, intervention variables are typically used to measure the effects of public policy innovations (e.g., the effect of the introduction of seat belt laws on traffic fatalities); the effects of various unanticipated events such as foreign policy crises or wars on levels of presidential approval and also scandals; or the impact of economic shocks such as the OPEC oil embargo on economic growth. Examples of continuous variables include levels of consumer confidence in models of public support for governing political parties and their leaders and interest rates in models of unemployment rates. These intervention and transfer function components can be hypothesized to exert either abrupt or gradual effects and, in the case of interventions, the effects can be hypothesized to be either permanent or temporary. Moreover, unlike the ADL model Eq. (5) the ARIMA intervention and transfer function models may specify different dynamics of the effects of various right-hand-side variables.

Specification of the effects of continuous variables in transfer function models is often accomplished by "cross-correlating" the dependent and independent variables. Cross-correlations are the correlations between Y_t and X_{t-i}, where $i = 0, 1, \ldots, p$. The aim is to detect at what lag X might affect Y. Typically, the variables are "prewhitened" (i.e., filtered) to purge them of possible spurious correlations before cross-correlations are computed. Although useful when theory is weak or absent, cross-correlations are not required when the analyst has hypotheses about when effects should occur. Since cross-correlations are bivariate measures of strength of relationship, they always should be viewed as heuristic devices in the model specification process.

An example of a (0, 1, 1) ARIMA model with one intervention and one transfer function component is

$$(1-L)Y_t = \frac{\omega_1}{(1-\delta_1 L)}(1-L)I_t + \frac{\omega_2}{(1-\delta_2)}(1-L)X_t + \epsilon_t - \theta\epsilon_{t-1}, \quad (15)$$

where I_t is an intervention (a 0-1 dummy variable) hypothesized to affect Y_t immediately (at time t). By scoring I_t as 0 for each period except t, the effect is temporary, declining at rate δ_1. Thus, the impact of I is ω_1 at time t, $\omega_1\delta_1$ at time $t+1$, $\omega_1\delta_1^2$ at time $t+2$, etc. In contrast, the impact of X_t begins as ω_2 at time t and increases to $\omega_2 + \omega_2\delta_2$ at time $t+1$, to $\omega_2 + \omega_2\delta_2 + \omega_2\delta_2^2$ at time $t+2$, etc. X_t's long-term effect on Y is $\omega_2/(1-\delta_2)$. A key point is that δ_1 and δ_2 may have different values. Note that in this example model, Y_t and X_t are $I(1)$ variables that have been "first differenced" (differenced once) to achieve stationarity. I_t is also differenced—a requirement to maintain the hypothesized nature of the effect of the intervention (temporary or permanent) if the dependent variable is differenced. After a model such as Eq. (15) is estimated, various diagostic procedures are employed to check its adequacy. As is the case for univariate ARIMA models, respecification, reestimation, and rediagnosis may be required.

Political scientists typically have used ARIMA intervention and transfer function models to test hypotheses about the nature and strength of various independent variables on a dependent variable of interest. In particular, there is a large literature in the United States about the impact of economic conditions and political events on the dynamics of presidential approval. Comparable bodies of research in Great Britain and other mature democracies examine the effects of such variables on the evolution of support for governing political parties. Although most of these studies have relied on the traditional time series analysis procedures described in the preceding section, some have used ARIMA modeling techniques. Contrary

to the original intentions of Box and Jenkins, only a few political scientists have used these techniques for forecasting purposes.

Cointegration and Error Correction

During the past two decades, political scientists have become increasingly aware of the threats to inference posed by nonstationary variables. In most cases, the reaction to this "spurious regressions" threat has been to difference variables suspected of nonstationarity and then to use the traditional procedures described previously. However, a growing number of analysts have recognized that such analyses ignore possible long-term relationships among the variables of interest. Following Engle and Granger, they have attempted to study such long-term relationships by using the concepts of cointegration and error correction. Two nonstationary variables are cointegrated when there exists a stationary linear combination of the variables. Cointegration among three or more nonstationary variables is defined the same way, although it is possible that the variables will form multiple cointegrating variables. Note that cointegration cannot be assumed; rather, it is an empirical question. There are two types of tests for cointegration—one proposed by Engle and Granger and the second by Johansen. Most political science applications have used the former and avoided the potential problem of multiple cointegrating vectors by focusing on only two variables. The Engle-Granger approach involves two steps: (i) Test the variables under consideration for nonstationarity, and (ii) if the variables are nonstationary, regress one of the variables on the other(s) and test the residuals for nonstationarity. If the residuals are stationary, the variables are cointegrated. As demonstrated by Engle and Granger's "representation theorem," cointegrating variables can be modeled in error correction form.

Error correction models are theoretically attractive because they enable one to study both the short- and long-term relationships among a set of variables. Error correction models have the following form:

$$(1-L)Y_t = \beta_0 + \beta_1(1-L)X_t + \alpha(Y_{t-1} + \lambda_1 X_{t-1}) + \epsilon_t, \quad (16)$$

where Y_t and X_t are nonstationary variables and have been rendered stationary by first differencing. β_1 captures the short-term effect of a change in X on a change in Y. Y and X are cointegrated, and the expression $(Y_{t-1} - \lambda X_{t-1})$ is the "error correction mechanism" that captures the long-term relationship between these variables. The strength of this cointegrating relationship is indicated by the adjustment parameter α. For a cointegrating system such as that depicted in Eq. (16), it is expected that α will carry a negative sign and be less than 1.0 in absolute value. The magnitude of α is theoretically interesting because it tells the speed with which a shock to the system is reequilibriated by the cointegrating relationship between Y and X. For example, if $\alpha = -0.5$, a shock at time t will be eroded at the rate of 50% in each subsequent period. In the Engle–Granger methodology, the error correction mechanism is measured as the residuals from the cointegrating regression of Y_t on X_t. Thus, there is a two-step estimation process. Step 1 is estimating the cointegrating regression, and step 2 is estimating the error correction model. Both analyses can be performed using OLS procedures. Recently, analysts have recommended that gains in statistical efficiency will be obtained by estimating Eq. (16) in a one-step process. If Y and X do not cointegrate, α will not be significantly different from zero.

Error correction models are attractive because they address the threat of spurious regressions while simultaneously enabling the analyst to study long-term relationships among a set of variables. However, these models are not a panacea. As in traditional time series regression models with lagged endogenous variables, the effects of all of the X variables are specified such that they have a common dynamic captured by the α parameter. As noted previously, this restriction may (often) be theoretically implausible. Also, important questions regarding the exogeneity of the X's need to be addressed to warrant confidence in the parameter estimates. These questions are not particular to error correction models, but they naturally arise in the context of demonstrating cointegrating relationships that ground the development of error correction models.

Exogeneity

With rare exceptions, the time series models estimated by political scientists are single-equation specifications. The concept of exogeneity as it has been developed by econometricians during the past two decades is crucial for evaluating inferences based on these models. The word "inference," in the econometric sense, refers to a single regression coefficient. In the 1940s and 1950s, the Cowles Commission discussed issues of identification and exogeneity by distinguishing between variables that were predetermined and those that were strictly exogenous within the context of a particular structural model or system of simulaneous equations. In subsequent work on the conditions for valid inference, Engle *et al*. developed weak exogeneity requirements that shift the focus to

parameters of interest as opposed to variables. These concepts may be defined as follows:

Predetermined: A variable is independent of the contemporaneous and future disturbances in the equation in which it appears.

Strict exogeneity: A variable is independent of contemporaneous, future, and past disturbances in an equation.

Weak exogeneity: A variable is weakly exogenous if the parameters of interest rest solely in that (conditional) model and the parameters of interest are variation free. There are no cross-equation restrictions between a (marginal) model for the process generating the independent variable and the conditional model for the dependent variable.

If the parameters of interest are variation free, the models are independent, and knowledge of a specific parameter in the marginal model provides no information about the range of values for parameters in the conditional model. Hence, no information is lost if the marginal model is ignored and only the conditional model—a single equation—is estimated. Tests for weak exogeneity are thus crucial for establishing the credibility of single-equation models such as the error correction models considered previously. In the context of an error correction model with one right-hand-side variable, a three-step process may be employed to test for weak exogeneity. First, specify a model for the marginal process (i.e., for the X). Second, add the error correction mechanism to this model. If X is weakly exogeneous to Y, the error correction mechanism should be insignificant in the model for X. Third, the residuals of the model for X (absent the error correction mechanism) should be insignificant when added as a predictor in the error correction model for Y.

Granger Causality

Granger developed a widely used definition of causality that is frequently employed by political scientists interested in the intertemporal flow of effects between two variables X and Y. Y is said to "Granger cause" X if information about the history of Y improves one's ability to predict the behavior of X, above what can be achieved when only information about the history of X is used for this purpose. Thus, if Y does not Granger cause X, Y is strictly exogenous to X. Granger causality tests can be performed within a Box-Jenkins ARIMA modeling framework or in the context of a traditional OLS regression analysis. For example, consider the following:

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_k X_{t-p} + \beta_1 Y_{t-1} + \cdots + B_p Y_{t-p} + u_t. \quad (17)$$

Estimating this model and performing a block F test for the joint significance of the Y's tests if Y Granger causes X. It is important to recognize that Granger causality tests are tests for strong, not weak, exogeneity. Granger causality tests are thus not helpful in deciding if it is permissible to draw inferences about the parameters in a single-equation model. However, Granger causality tests are useful for deciding if a single equation model of Y where X is a right-hand-side variable will be useful for forecasting purposes. If Y does Granger cause X, then this needs to be taken into account. The intuition here is that the single-equation model for Y does not take account of the feedback from Y to X revealed by test results indicating that the former variable Granger causes the latter. An important general point is that analysts should not confuse tests for Granger causality with tests for weak exogeneity (i.e., parameter stability tests). Both are useful, but they are useful for different purposes.

(Non)Stationarity

Both traditional regression-based approaches to time series analysis and ARIMA modeling procedures require that the variables being analyzed be stationary. (Non)stationarity is also a principal consideration in analyses of cointegration. Until the mid-1980s, nonstationarity was assessed by two methods. First, researchers displayed their data in graphic form and looked to see if the series had an upward or downward trend. Second, following Box and Jenkins, analysts computed an autocorrelation function (ACF) for a series and inspected the results. An ACF is a series of correlations between a variable (Y_t) and itself at successive lags (i.e., Y_{t-1}, $Y_{t-2}, \ldots, Y_{t-k}$). The classic signature for a nonstationary series is a set of very stong correlations that decay very slowly as the lag length increases. If such a pattern was detected in the ACF, the series was judged to be nonstationary. Clearly, neither the graphic nor the ACF procedures constituted formal statistical tests.

Since the late 1970s, econometricians have developed a wide variety of formal statistical tests for nonstationarity. The most widely used such test is that of Dickey and Fuller. The basic Dickey-Fuller test statistic is computed from an OLS regression for the following model:

$$(1-L)Y_t = (\beta_1 - 1)Y_{t-1} + \epsilon_t. \quad (18)$$

The null hypothesis for the test is that the series is nonstationary, the basic null hypothesis being that it is generated by a random walk. If so, $\beta_1 - 1 = 0$. Additional parameters for a constant and/or a deterministic trend may be added to Eq. (18), depending on what specific assumptions are made about the data-generating process. In any event, rejection of the null hypothesis

prompts the inference that the series is stationary. A key point is that t distributions for the Dickey–Fuller test are nonstandard and, hence, special critical values must be employed. Also, as in any regression, inference is problematic if the ϵ's have nonzero correlations. If diagnostics suggest the presence of such correlations, lags of the dependent variable (which is differenced) are included in a respecified model, and the model parameters are reestimated. The resulting test is called an augmented Dickey–Fuller test.

Two points regarding (non)stationarity tests bear emphasis. First, structural breaks in otherwise stationary processes may falsely prompt the conclusion that a series is nonstationary in the sense of being generated by some type of stochastic trend data-generating process. Second, unit-root tests have low power against near-integrated and fractionally integrated alternatives. The result is that failure to reject the null hypothesis may lead the analyst to falsely conclude that β_1 in Eq. (18) equals 1.0, and the data-generating process is random walk. Recognition of this possibility has led some researchers to argue that alternative procedures should be employed to determine the order of integration.

Fractionally Integrated Processes

Work on "long-memory" processes relaxes the assumption that the differencing parameter d must have integer values. Analyses are done within the framework of an ARFIMA (autoregressive, fractionally integrated, moving average) model:

$$(1-L)^d Y_t = \frac{\theta(L)}{\phi(L)}\epsilon_t, \tag{19}$$

where Y_t is a time series variable, ϵ_t is a stochastic error, ϕ signifies an autoregressive parameter(s), θ signifies a moving average parameter(s), and d is the differencing parameter. The model thus resembles the conventional ARIMA model except that the differencing parameter d can take values along the real line from -0.5 to 1.0. Nonzero values of d signify a fractionally integrated process. When $0 < d < 0.5$, the process is stationary. However, when $0.5 \leq d < 1.0$, the process is nonstationary, but ultimately mean reverting. A key point is that d can be estimated from the data, and associated standard errors can be used for hypothesis testing purposes. As with conventional ARIMA models, rival ARFIMA models can be evaluated on the basis of various diagnostics, including model selection criteria such as the AIC and BIC. Work on fractionally integrated processes has increased in recent years, and some studies have generalized the concept of cointegration to consider fractionally cointegrated systems. In political science, a finding with potential theoretical significance is that many often-analyzed time series, including presidential approval, party identification, and economic evaluations, are nonstationary, fractionally integrated processes. Clarke and Lebo cite relevant literature.

Vector Autoregression

In reaction to the forecasting failures of traditional multiequation Cowles Commission models, Sims proposed a technique called VAR. Sims focused his critique on the "incredible" assumptions made in the specification of traditional models. Lacking adequate theoretical or empirical underpinnings, these assumptions were principally matters of convenience to ensure that model parameters could be identified. Traditional models were a system of structural equations that expressed endogenous variables as being (in part at least) a function of the current value of other endogenous variables. Sims proposed a reduced-form alternative whereby all endogenous variables are functions of their own lagged values as well as lagged values of other endogenous variables. Motivated by a desire to improve forecasting performance rather than to test economic theory, Sims contended that such a reduced-form specification could capture dynamic interrelationships among variables of interest. All that was needed was to "round up the usual suspects" (i.e., variables that theory and experience suggest are relevant to the forecasting exercise at hand) and include them in a VAR.

The following is an illustrative two-variable example of a VAR:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \alpha_1 X_{t-1} \\ &\quad + \cdots + \alpha_p X_{t-p} + \varepsilon_t \\ X_t &= \gamma_0 + \gamma_1 X_{t-1} + \cdots + \gamma_p X_{t-p} + \lambda_1 Y_{t-1} \\ &\quad + \cdots + \lambda_t Y_{t-p} + \xi_t. \end{aligned} \tag{20}$$

Error terms in such a VAR model (ε_t, ξ_t) are assumed to be $\sim N(0, \sigma^2)$ and serially uncorrelated. Thus, each equation in the model may be estimated using OLS. If ε_t and ξ_t are contemporaneously correlated, seemingly unrelated regression provides no gains in efficiency because identical sets of predictor variables are in each equation. In addition, although the presence of interrelated lagged regressors produces collinearity, this is not a problem because there is no interest in inference on individual parameters. Assuming variables in the system are stationary, the only question concerns the appropriate lag length, p, for the variables in the system, and various procedures (block F tests and model selection criteria) may be employed for this purpose. Since the stationarity assumption may not obtain, econometricians have debated whether nonstationary variables should be used in a VAR. Enders reviews the

issue, usefully linking his discussion to Johansen's methodology of cointegration tests.

In Sim's original formulation, VAR is proposed as a forecasting tool. However, analysts also use VAR methods to investigate relationships among a set of interacting variables. To this end, a VAR is expressed as a vector moving average (VMA) system. A VMA representation allows one to trace the impact of a shock to one variable through time on other variables in the system via impulse response functions and forecast error variance decomposition. Such "innovation accounting" exercises are sensitive to assumptions about the overall flow of causality through the variables in the system. Varying the order of variables in the system enables the analyst to determine the sensitivity of innovation accounting results to such assumptions.

Despite increasing sophistication in the use of time series methods, and widespread availability of suitable software, political scientists have been slow to adopt VAR methods for applied work. In part, this reflects the discipline's continuing emphasis on inference rather than prediction. However, forecasting exercises also have eschewed VARS in favor of simple structural models or, occasionally, univariate ARIMA models.

ARCH Models

Political scientists are typically interested in modeling the mean of a time series variable. However, some analysts, (e.g., political economists studying exchange rates) are concerned with the volatility of a series. For this purpose a class of ARCH (autoregressive conditional heteroskedasticity) models pioneered by Engle are very useful. ARCH models express the conditional variance of stochastic errors as an ARMA process:

$$\epsilon_t = v_t \sqrt{\bar{h}_t} \tag{21}$$

$$h_t = \alpha_0 + \sum_{i=1}^{q} a_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i} \tag{22}$$

Here, v_t is $\sim N(0, 1)$, and so the conditional variance of ϵ_t is $E_{t-1}\epsilon_t^2 = h_t$. This basic GARCH (generalized ARCH) model has been modified in several ways. For example, ARCH-M models specify that the mean of a series is a function of its conditional variance (h_t). Threshold ARCH and exponential GARCH models account for asymmetries in the effects of positive and negative shocks, and integrated GARCH models consider situations of strong persistence in the conditional volatility of a series. ARCH models may also be extended to study how various events and conditions affect the conditional volatility of a series. Although most political science applications of ARCH models are found in work by political economists, some analysts are beginning to use these models to study volatility in presidential approval ratings and support for political parties.

The Future of Time Series Analysis: Linkages to Theory

Although there are many applications of time series methods in political science, it is likely that these techniques will become increasingly popular in the future. In part, this development will reflect the continued growth in time series databases relevant to political science research. Equally important is the growing recognition of the need for a closer articulation of theory and method. In this regard, there have been sporadic attempts, albeit no systematic intellectual movement, to link time series techniques directly to behavioral theories. To date, these efforts have occurred in other disciplines, such as economics. This situation is changing. Although there will always be a need for pure time series statistical tools, there is also a new appreciation among political scientists of the utility of linking time series techniques to formal models—what has been called the empirical implications of theoretical models (EITM) (see http://www.nsf.gov/sbe/ses/polisci/eitm_report/start.htm). Using EITM, political scientists can take a set of plausible facts or axioms, model them in a rigorous mathematical manner, and identify causal relations that explain empirical regularities over time. Of course, these transparent linkages between theory and testing procedures do not mean a theory is correct. Instead, political scientists who link formal models with time series techniques would satisfy a minimal requirement that theory and test are related. This linkage bears on important issues regarding falsification and the accumulation of scientific knowledge.

See Also the Following Article

Fixed-Effects Models

Further Reading

Box, G. E. P., and Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*, Rev. Ed. Holden Day, Oakland, CA.

Charemza, W., and Deadman, D. F. (1997). *New Directions in Econometric Practice*, 2nd Ed. Edward Elgar, Aldershot, UK.

Clarke, H. D., and Lebo, M. (2002). Fractional (co)integration and governing party support in Britain. *Br. J. Polit. Sci.* **33,** 283–301.

DeBoef, S., and Granato, J. (2000). Testing cointegrating relationships with near-integrated data. *Polit. Anal.* **8,** 99–117.

Enders, W. (2004). *Applied Econometric Time Series*, 2nd Ed. Wiley, New York.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50,** 987–1007.

Engle, R. F., Hendry, D., and Richard, J.-F. (1983). Exogeneity. *Econometrica* **51,** 277–304.

Engle, R. F., and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica* **55,** 251–276.

Granato, J., and Scioli, F. (2004). Puzzles, proverbs, and omega matrices: The scientific and social significance of empirical implications of theoretical models (EITM). *Perspect. Politics* **2,** 313–323.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37,** 24–36.

Gujarati, D. (2003). *Basic Econometrics*, 4th Ed. McGraw-Hill/Irwin, New York.

Hendry, D. (1995). *Dynamic Econometrics.* Oxford University Press, Oxford, UK.

Johansen, S. (1991). Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica* **59,** 1551–1580.

Sargent, T. (1981). Interpreting economic time series. *J. Polit. Econ.* **89,** 213–248.

Sims, C. (1980). Macroceonomics and reality. *Econometrica* **48,** 1–49.

Time-Series–Cross-Section Data

Nathaniel Beck
New York University, New York, New York, USA

Glossary

contemporaneous correlation of the errors When errors for observations in a unit are correlated with errors for other units observed at the same time.
dynamic model States that effects of covariates occur over time rather than instantaneously.
feasible generalized least squares (FGLS) Generalized least squares in which parameters of the error process are estimated.
panel correct standard errors (PCSEs) Ordinary least squares standard errors corrected for panel heteroskedasticity and contemporaneous correlation of the errors.
panel heteroskedasticity Heteroskedasticity in which variance differs from unit to unit, but is constant within units.
pooling The assumption that all units follow the same specification with identical parameters.
spatial effects Effects when values for other units enter the specification for a given unit.
time-series–cross-section (TSCS) data Data observed at regular intervals on fixed units.

Time-series–cross-section data consist of repeated observations at fixed intervals on a group of fixed units in which the observed units comprise, in general, the entire population of such units. A common application is to the political economy of advanced industrial societies for which there are annual measures of political and economic phenomena in approximately 20 nations, but there are many other applications in political science, economics, and sociology.

Introduction

What Are Time-Series–Cross-Section Data?

There are many types of data for which there are repeated observations on the same units over time. Time-series–cross-section (TSCS) data are one type of such data. The units in TSCS data are of interest per se; in other types of data, the units studied are a sample from a population, and interest centers on inferences to that larger population.

Thus, in a study of how politics affects economic growth in advanced industrial societies, all advanced industrial societies are observed. There are no issues of inference to a larger population of nations. In resampling experiments, a new "Germany" will not be drawn from a large population of nations; the various unobserved characteristics of "Germany" remain constant over any sampling (thought) experiments. Many thought experiments will be of the form: "What would have happened if Germany had only right-wing governments?" As shall be seen, this is different from panel data based on sample surveys of individuals, in which there is no interest whatsoever in inferences conditional on a specific individual.

The repeated observations in TSCS data are at fixed intervals. For many political economy applications, this interval is annual, but it could be quarterly, monthly, or daily. The fixed intervals allow the analyst to know that, for example, the observation for unit 3 at time period 6 is perhaps related to the observation for unit 7 at time period 6. This allows analysts to model the dynamics of TSCS data as they would for simple time-series data, and allows for many of the methods and insights of the time-series analyst to be used by the TSCS analyst.

Time-series insights for TSCS data are only relevant if enough time points are observed. While there is no hard and fast rule as to how many time points need be observed, analysts clearly cannot use time-series insights and methods if each unit is observed only two or three times. TSCS data thus consist of repeated observations at fixed intervals on fixed units, with enough repeated observations to make time-series insight relevant.

Examples of TSCS Data

The paradigmatic political economy application, as exemplified by the work of Franzese, Garrett, and Iversen, examines economic policies or outcomes in 15–20 advanced industrial societies (members of the OECD) in the post-World War II period (often 1960–1990), using annual data. The dependent variable is either an economic outcome (e.g., growth, unemployment, or inflation) or a policy (e.g., budget deficit or rate of money growth), with the covariates being both economic and political indicators. The political indicators usually measure the strength of left parties in the government, the organization of labor, and various political rules in the country (such as the electoral system).

Economists undertake similar analyses. Pesaran, Shin, and Smith, for example, examined the consumption function in 24 advanced industrial societies observed annually over 32 years. The dependent variable was the level of national consumption, with independent variables being the level of national income and other economic indicators. In this model, interest centers on both short-term and long-run effects.

Related Types of Data

TSCS data appear similar to other types of data commonly used in the social and biomedical sciences; the notation for these data sets often appears identical. But there are fundamental differences between TSCS and related data sets, both theoretical and in terms of practical issues of estimation. In particular, panel data, which appear to be notationally equivalent to TSCS data, are analyzed very differently from TSCS data.

Panel data consist of repeated observations, at fixed intervals, on a set of units. The units are sampled from a larger population, and interest centers on inference to that population. Typically, the number of repeated observations is not large. The paradigmatic panel study is the Panel Study of Income Dynamics, which (in simplified version) surveys a large number of individuals each month for approximately one year. Thousands of respondents are sampled, but interest centers on adult Americans. In principle, subjects can be interviewed a large number of times, but in practice, they are only interviewed a few times, with two or three interviews being most typical.

It should also be noted that the TSCS model is related to Zellner's "seemingly unrelated regressions" (SUR) model. In that model, time-series data for a number of units are observed. While the parameters of each time-series model differ, it is assumed that the error processes for each series for the same time period are related. As shall be seen, this is just one variant of a TSCS model.

Both panel and TSCS data are special cases of hierarchical data. Here, data are observed on units that are tied together as subunits. The paradigmatic example is data on students who are tied together in classes that are tied together in schools; the students studied are randomly drawn from a larger population of interest. Both TSCS and panel data impose more structure on the data, since the same subject is observed at repeated intervals. While the notation in hierarchical data may indicate subject 1 in class 1 and subject 1 in class 2, there is no relationship between those subjects.

So far, it has been assumed that the models have a continuous dependent variable. But TSCS data can also have a binary (or other discrete) dependent variable. The common application is the study of conflict, in which nations (or pairs of nations) are observed annually; each year it is recorded whether a dyad was in conflict, with covariates being either continuous or dichotomous. Such data are particularly difficult to model and present many interesting estimation issues that go beyond the continuous case.

Plan of the Article

TSCS data can be seen as presenting either estimation issues or modeling issues. The former derive from an older tradition, and many estimation issues can be simplified with a more modern approach. Current research focuses on modeling issues, particularly the modeling of spatiality and heterogeneity.

The next section discusses the notation used and overviews basic TSCS models, while showing how they may be differentiated from their close cousins; the following section emphasizes the importance of preliminary graphical analysis. Then some estimation issues related to the cross-sectional and temporal properties of the data are treated; attention is paid to both "old fashioned" and more modern approaches. The last two sections cover the direct modeling of some properties of TSCS data and some current issues.

Notation and Related Models

The common notation and generic specification used in this article are set forth here, with further discussion of some assumptions implicit in that notation. Because the notations do not adequately distinguish TSCS from panel data, this issue is pursued further.

The Basic Model

It is assumed throughout that there is a single dependent variable that is explained by a vector of covariates, as well as, possibly, by the past history of that dependent variable and various possible combinations of the error process. Reciprocal causation is thus ruled out, and the covariates

are assumed to be exogenous to the variable that is being explained.

Let $y_{i,t}$ indicate the dependent variable, which is indexed by both unit i and time period t, where it is assumed that all observations $y_{i,t}$ and $y_{i',t}$ refer to the same time period and all observations $y_{i,t}$ and $y_{i,t'}$ refer to the same unit. For convenience, given the paradigmatic application to political economy, i can be thought of as a country and t as a year. Assume we have N countries observed for T years.

It is assumed that the relation between $y_{i,t}$ and the vector of covariates $\mathbf{x}_{i,t}$ is linear, although this assumption can be weakened in the usual ways of introducing nonlinearities while maintaining an additive model. Notationally, we have

$$y_{i,t} = \mathbf{x}_{i,t}\beta + \epsilon_{i,t}; \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \qquad (1)$$

where ϵ is an error process that may have complicated properties.

While the error process in Eq. (1) may allow for temporal observations to be related, Eq. (1), is "static," in that the exogenous variables have an immediate effect on y that lasts exactly one period. The covariates may be lagged without problem, but they still affect y for only one period.

Relationship to Other Models

TSCS vs Panel Data Models

TSCS models are often confused with panel data models because the notation for panel data models is identical to Eq. (1). But the differences between the two types of data, and the methods and models used for estimation, are large. The reason is that in TSCS models, whereas N, the number of units, is taken as fixed, T, the number of observations per unit, may be thought of as growing larger and larger. All asymptotic theory for TSCS data is in T, with N fixed. Although N may be large, in practice it is not; the most common political economy studies have $N = 20$ and very few studies have more than 100 units.

The situation is exactly the opposite for panel data models. In these models, T is taken as fixed, whereas N is usually large and asymptotics are in N. Thus, a common panel data structure may have thousands of respondents interviewed three or four times. Whereas the number of respondents sampled can be thought of as growing larger and larger, the number of interviews per respondent cannot change, and, in practice, is small (almost always under 10, with 3 being the most common number of "waves" of a panel).

This has enormous consequences both for how the two types of data are modeled and for the estimation of these models. In TSCS data, with usually 20 or more observations per unit, it is possible to think of richer time-series models for each unit (though clearly not of the richness of standard single time-series models, in which hundreds of time-series observations, often at very high frequency, are available). But in panels with only very few observations per unit, it is clearly impossible to model the temporal structure of the data in any detail. While TSCS and panel data models have interrelated observations for each unit, it is difficult to do more than provide some simple fixes for those interrelationships in panel data.

Because the units in panel data consist of a random sample of individuals from a larger population, it is usually assumed that observations of different individuals are independent of each other. But for TSCS data, there is great interest in modeling the interrelationship between observations of the different units. Panel data modelers usually assume no spatial effects (relationships across units) in their data.

Finally, panel data analysts assume that they observe a sample of units drawn from a larger population of units, with inferences to that larger population being of interest. TSCS analysts, on the other hand, usually observe the entire population of units, and so have no need to worry about inferences to a larger population.

The differences in asymptotics also have implications for how diversity between units is modeled. The simplest assumption is that each unit in Eq. (1) has its own intercept, a_i, adjoined to the specification, which leads to

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\beta} + a_i + \epsilon_{i,t}; \quad i = 1, \ldots, N; \quad t = 1, \ldots, T. \quad (2)$$

In panel data, fixed effects cause enormous estimation difficulties related to the incidental parameters problem first discussed by Neymann and Scott. In brief, this problem is that the number of incidental parameters, a_i, grows asymptotically as N grows. The a_i, therefore, cannot be consistently estimated, because they are estimated using only three (or T) observations. The difficulties of fixed effects models for panel data have led to a whole series of complicated models in which the a_i are taken as random draws (random effects). Note, however, that for TSCS data, asymptotics are in T with N being fixed. Thus, there will always be N dummy variables in Eq. (2) and hence a_i can be consistently estimated, since each is estimated with T observations that can be thought of as growing larger and larger. One of the biggest specification and estimation problems for panel data is thus a simple issue for TSCS data.

Note that even if fixed effects were statistically feasible for panel data, it would still be of little interest to panel analysts because they care about the population, not the observed units. Noting that some particular unit, say K, is, on average, a_K higher than a baseline unit is therefore of no interest—the estimated fixed effects tell the panel analyst nothing about a population. TSCS analysts, however, find great interest in the estimates of the fixed effects. Say K refers to Germany; then a_K tells the analyst, for example, the average growth rate of Germany relative to a reference unit.

There are many other differences between TSCS and panel data models, all brought on by the small T and big N for panel data and the reverse for TSCS data.

Seemingly Unrelated Regressions

The Zellner seemingly unrelated regressions (SUR) model assumes that there are N time-series models, indexed by i, in which each time-series is observed over the same time period, $1, \ldots, T$. Thus, time-series models can be thought of for various sectors of the economy, observed quarterly over some time period. While the time series for each sector can be estimated separately (assuming that T is large enough), Zellner's insight was that each observation at any time point, t, is related to every other observation at t. This information could be exploited to improve estimation.

Formally, the SUR model assumes

$$y_{i,t} = \mathbf{x}_{i,t}\boldsymbol{\beta}_i + \epsilon_{i,t}; \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \qquad (3)$$

where the specifications for the different units may include different covariates (in notation, some elements of β_i may be constrained to be zero). As will be seen in the next section, the error process for the SUR model is similar to some of the error processes for TSCS models. The assumption of a single $\boldsymbol{\beta}$ relating the covariates to y in Eq. (1) is one of complete homogeneity of the units, the assumption of complete "pooling." The SUR model is thus a model of complete heterogeneity of the units, or "non-pooling." This distinction has many consequences.

It should be noted that the typical SUR application has a relatively large T because the time-series model for each unit must be estimated separately; $T > 100$ is not uncommon. Also, in many applications, the number of units is quite small (sometimes only 3). This must be borne in mind, since it has implications for estimation.

Other Assumptions

It is assumed that the data set is "rectangular," that is, that all countries are observed for the same time period $1, \ldots, T$, with the same beginning and ending dates, and no missing data in the interior of this period. It is completely trivial, though notationally cumbersome, to allow each country to have its own starting and ending period and to allow different countries to be observed for different lengths of time. It is critical that time period t in $y_{i,t}$ and $y_{i',t}$ refer to the same calendar period (year).

Preliminary Data Analysis

Although it is tempting to turn directly to econometric estimation, researchers should, as always, begin by examining the data using common summary statistics and graphical methods. But in addition to standard data inspection methods, TSCS data presents some unique possibilities. Given that the number of units is small, and that the researcher is knowledgeable about those units, box plots of the dependent variable, disaggregated by unit, can be most informative. The unit box plots are only informative if the number of observations per unit, T, is sufficiently large, but this should be the case for most data sets that will be analyzed using TSCS methods.

Researchers can examine the box plots of the dependent variable to see whether there are gross differences in the median value for different units; the width of each box can also be examined to see whether the variance of the dependent variable differs markedly by unit. The plot can also show whether there is some unit that differs radically from the preponderance of units. It can also show whether there is sufficient intra-unit variation to make TSCS analysis meaningful. Finally, the unit box plots can be most helpful in allowing the analyst to find and deal with outliers.

Figure 1 shows a box plot for TSCS data used by Franzese in his 2002 study of the political determinants of the government deficit (as a proportion of GDP) in developed democracies in the post-World War II era (with 3 outliers winsorized). First, it can be seen that all countries exihibit temporal variation in their deficit, indicating that the deficit is not purely a function of stable cross-sectional variables. The medians by country, although different, do not indicate that any country is dramatically different from the others. Finally, while the variation of deficit by country is not constant, there does not appear to be an enormous amount of country-to-country variation in the variation of deficits over time. But analysts might expect to find some panel heteroskedasticity when estimating models of government deficit. Plots like Fig. 1 should always be undertaken before more technical analysis is attempted. Researchers can also examine the impact of time on the dependent variable by producing box plots of the dependent variable disaggregated by year. Having done this, the researcher can then turn to estimation of the relevant models. Once a model is estimated, the model residuals can be graphed using a similar disaggregated box plot. Such a residual plot is invaluable in the refinement process.

Estimation Issues: Cross-Sectional

If the errors in Eq. (1) satisfy the Gauss-Markov assumptions, then it is optimal to estimate it by ordinary least squares (OLS). The Gauss-Markov assumptions about the errors are

$$\mathrm{Var}(\epsilon_{i,t}) = \sigma^2, \qquad (4)$$

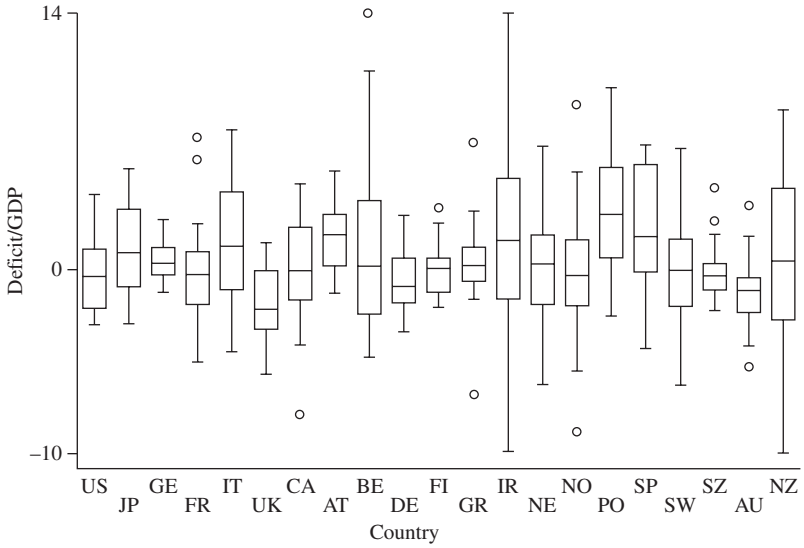


Figure 1 A box plot of government deficit (as percent of GDP) by country in post-World War II developed democracies.

$$\text{Cov}(\epsilon_{i,t}\epsilon_{i',t'}) = 0 \quad \text{if} \quad i \neq i', t \neq t'. \tag{5}$$

It is assumed throughout that the models contain a constant term so that $E(\epsilon_{i,t}) = 0$. It is also assumed throughout that the covariates are exogenous.

The traditional approach to TSCS data is based on the idea that the properties of TSCS data cast doubt on the Gauss-Markov assumptions in several ways, but that, given weaker assumptions about the error structure, Eq. (1) can be optimally estimated by feasible generalized least squares (FGLS). This approach is traditional in that it does not question the basic specification, Eq. (1). Rather, it simply regards the TSCS properties of the error process as a nuisance that causes difficulties for OLS. A more modern approach allows for the various features of TSCS data to enter the specification relating the covariates to the dependent variable.

More General Error Processes and FGLS

While the errors in TSCS models may violate the Gauss-Markov assumptions for all the reasons that error processes in any model may be complicated, there are some features of TSCS data that make the Gauss-Markov assumptions particularly suspect. These can be categorized as panel heteroskedasticity, contemporaneous correlation of the errors, and serial correlation of the errors. For this section, assume that there are no dynamics, so work with Eq. (1) will assume temporally independent errors. (This assumption is relaxed in the next section—the methods discussed there can easily be conjoined with the recommended approach from this section, and so there is no loss in postponing the issue of dynamics.) It is also possible to correct the standard errors for violations of the Gauss-Markov assumptions. This section concludes with a discussion of panel correct standard errors (PCSE) and a recommended methodology for dealing with violations of the Gauss-Markov assumptions that are related to cross-sectional issues. (The nomenclature using the term panel has unfortunately become standard; these are all TSCS, not panel data, issues.)

Panel Heteroskedasticity

The simplest complication of the error process retains independence across observations but allows for the error variance to vary from unit to unit. Panel heteroskedasticity thus maintains the independence of observations of the Gauss-Markov assumptions but relaxes the assumption in Eq. (4) by

$$\text{Var}(\epsilon_{i,t}) = \sigma_i^2. \tag{6}$$

Note that this form of heteroskedasticity allows for the spatial structure of TSCS data and is more restrictive than the general forms of heteroskedasticity studied in general linear models.

The presence of panel heteroskedastic errors means that OLS is no longer optimal and that the standard errors reported by OLS are no longer accurate. It is easy to test for panel heteroskedasticity via a likelihood ratio test. The null hypothesis for this test is that all of the σ_i^2 in Eq. (6) are identical, with the alternative hypothesis that at least one differs from the others. Let $\hat{\sigma}^2$ and $\hat{\sigma}_i^2$ be the maximum likelihood estimates of the homoskedastic and

panel heteroskedastic σ^2 in Eqs. (1) and (6) (these are estimated by the relevant sums of squared errors divided by either $N \times T$ or T). The likelihood ratio statistic is then

$$T\left(N \ln\hat{\sigma}^2 - \sum_{i=1}^{N} \ln\hat{\sigma}_i^2\right), \tag{7}$$

which is asymptotically distributed as chi-squared with $N-1$ degrees of freedom.

While the likelihood ratio test (or other common variants based on Lagrange multiplier methods) is useful for testing the null hypothesis of panel homoskedasticity, researchers should do more than simply perform this test. With TSCS data, it is possible to meaningfully estimate each of the σ_i^2. (It is not possible to pursue this investigation for simple cross-sectional data.) These $\hat{\sigma}_i^2$ can then be examined to see, for example, if only one particular unit has a particularly large error variance, that is, it does not fit the basic specification well. Such inspection of the estimated unit variances should always be combined with a more formal hypothesis testing strategy. Such inspection is invaluable if researchers are to make an informed tradeoff of the costs and benefits of alternative estimation strategies for panel heteroskedastic data.

If there is panel heteroskedasticity, Eq. (1) can be estimated by FGLS. This is equivalent to panel-weighted least squares, with observations for each unit being weighted by the inverse of the square root of $\hat{\sigma}_i^2$. This has the usual asymptotically optimal properties of FGLS.

Even if the likelihood ratio test indicates that the null hypothesis of panel homoskedasticity can be rejected, researchers should be careful in using FGLS that downweights observations on units that do not fit Eq. (1) well, while giving units that fit greater weight. Consequently, FGLS can easily mislead researchers into concluding that results based heavily on units for which Eq. (1) fits well apply to all units. Researchers should thus have a strong prior belief that Eq. (1) holds equally for all units before using FGLS to correct for panel heteroskedasticity. As shall be seen, it is easy to correct the OLS standard errors for panel heteroskedasticity without causing the problems inherent in panel-weighted least squares.

Contemporaneously Correlated Errors

Much attention has focused on estimating models in which the errors show both panel heteroskedasticity and contemporaneous correlation of the errors, that is,

$$\text{Cov}(\epsilon_{i,t}\epsilon_{i',t'}) = \begin{cases} \sigma_{i,j} & \text{if } t = t' \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

This assumes that at any given time, the error process for any given unit is related to the error process for other units, but that this linkage only occurs contemporaneously. The amount of correlation between units is unspecified, so there are many $[N(N-1)/2]$ extra parameters in this model.

Parks suggested that models with panel heteroskedastic and contemporaneously correlated errors could also be estimated by FGLS. While the estimation is complicated, the basic idea is that Eq. (1) is first estimated by OLS (which is consistent). The OLS residuals are then used to build up estimates of the contemporaneous covariance matrix and heteroskedasticity as in Eqs. (6) and (8), and finally observations are transformed by the inverse of the Cholesky decomposition of that estimated matrix. For this procedure to work, it is required that $T > N$.

But even if $T > N$, FGLS estimates $[N(N-1)]/2$ additional parameters that are not accounted for in the FGLS standard errors (because FGLS assumes that the error process parameters are known, not estimated). Monte Carlo evidence indicates that the FGLS standard errors in this case may be off by 50% or more unless T is much greater than N. For typical TSCS data, the FGLS correction for contemporaneously correlated errors should thus not be used.

As with panel heteroskedasticity, analysts can also examine the estimated contemporaneous correlation of the errors, computed from the OLS residuals. These estimates can be used to see if OLS might be highly inefficient. But even if the estimated contemporaneous correlations are high, the poor properties of the FGLS correction for contemporaneously correlated errors indicate that it should not be used. But high contemporaneous correlation of the residuals indicates that researchers should try to respecify their model.

Fortunately, it is still possible to use OLS, which is consistent, and then provide panel correct standard errors that are accurate indicators of the variability of OLS estimates even in the presence of panel heteroskedasticity and contemporaneous correlation of the errors. This allows researchers to avoid the serious problem that is common with TSCS data, incorrect OLS standard errors, without having to turn to the very problematic FGLS approaches.

Panel Correct Standard Errors

Beck and Katz showed that it is easy to correct the OLS standard errors for problems of panel heteroskedasticity and contemporaneous correlation of the errors. Unlike other corrections for heteroskedasticity, such as that of White, this method relies less heavily on asymptotic results, since with TSCS data it is possible to estimate the covariance matrix of the errors using T replicates of the OLS residuals.

The basic insight behind panel correct standard errors (PCSEs) is that, however complicated the error process,

the true sampling variance of the OLS estimators is given by

$$\text{Cov}\left(\hat{\boldsymbol{\beta}}\right) = (\mathbf{X}'\mathbf{X})^{-1}\{\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\}(\mathbf{X}'\mathbf{X})^{-1}, \tag{9}$$

where $\mathbf{X}$ is the data matrix built of all covariates (with data stacked by time period so that the first row of data is for unit 1, time period 1, followed by unit 2, time period 1, etc.) and $\boldsymbol{\Omega}$ is the true covariance matrix of the errors.

Under the Gauss-Markov assumptions, $\boldsymbol{\Omega}$ reduces to $\sigma^2 I$, which then yields the usual OLS standard errors. For TSCS data with contemporaneously correlated and panel heteroskedastic errors, however, $\boldsymbol{\Omega}$ does not reduce to this simple form. Rather, $\boldsymbol{\Omega}$ is a block diagonal matrix, with the contemporaneous covariances forming each block (and by assumption, each block is identical). Letting $\mathbf{E}$ denote the $T \times N$ matrix of the OLS residuals, $\boldsymbol{\Omega}$ can be estimated by

$$\hat{\boldsymbol{\Omega}} = \frac{(\mathbf{E}'\mathbf{E})}{T} \otimes \mathbf{I}_T, \tag{10}$$

where $\otimes$ is the Kronecker product. This estimate can then be substituted into Eq. (9) to produce a correct estimate of the variance-covariance matrix of the OLS estimates. Monte Carlo evidence indicates that standard errors computed from this matrix are quite accurate, usually being within a few percent of the true variation. Equally importantly, this accuracy is obtained with the typical Ts seen in common TSCS data sets (that is, PCSEs are reasonably accurate with 20 time points per unit, and are almost perfect with 50 time points per unit).

It should be stressed that the PCSEs correctly indicate sampling variation of the OLS $\hat{\boldsymbol{\beta}}$s. Because the $\hat{\boldsymbol{\beta}}$s are not themselves modified, PCSEs clearly do not remedy the possible inefficiency of OLS. If analysis indicates that this inefficiency is serious, then alternative methods must be sought. But, in practice, the inefficiency of OLS is often not severe, and it is possible to improve the basic specification [Eq. (1)] to model the causes of heteroskedasticity and contemporaneous correlation of the errors directly, leaving the resulting OLS estimates reasonably efficient.

Thus, current practice is to estimate TSCS models that may show panel heteroskedasticity and contemporaneous correlation of the errors using OLS and PCSEs. While OLS is not optimal, the FGLS "improvements" work out very poorly in practice. PCSEs, on the other hand, guard against incorrect estimates of sampling variability at essentially no cost.

Estimation Issues: Dynamics

As in single time-series models, it is unlikely that the observations in TSCS models will be temporally independent and show no dynamics. It is often the case that TSCS data sets are observed for a time period long enough that analysts can seriously think about modeling dynamics. Nevertheless, it is also often the case that such data sets are observed for a shorter time period, and at a lower frequency, than single time-series data sets. As a result, many of the refinements possible for the analysis of a long single time series are not available to most TSCS analysts. This section focuses on simple (first-order) models; analysts with richer time-series data can use most of the more sophisticated time-series methods as their data allow. This section examines only dynamic issues; as shall be seen, it is easy to combine the treatment of dynamics with the recommended PCSEs.

Serially Correlated Errors

The old-fashioned approach, corresponding to the FGLS approaches of the previous section, is to assume that dynamics manifest themselves as serially correlated errors that are an estimation nuisance. In this approach, analysts allow for first-order serial correlation of the errors and then use FGLS to correct for this problem. Serially correlated errors are assumed to follow

$$\epsilon_{i,t} = \rho\epsilon_{i,t-1} + \nu_{i,t}, \tag{11}$$

where the ν are a "white noise" (independent and identically distributed) process, and some suitable assumption must be made about the first observation for each unit.

A variety of tests exists for serially correlated errors. The Lagrange multiplier (LM) test is the easiest and also fits best with what is done below. The null hypothesis for the test is that $\rho = 0$ against the alternative that it is not zero. The LM test regresses the OLS residuals [of Eq. (1)] on their first-order lags and all of the covariates in the model. The null can be assessed by either examining the t-statistic on the coefficient of the lagged residual or by examining the statistic $N \times T \times R^2$, which has a chi-squared distribution with 1 degree of freedom and where R^2 is the uncentered squared correlation of the "auxiliary" regression.

If the null hypothesis of serial independence of the errors is rejected, standard FGLS methods can be used. These consist of running OLS on Eq. (1), computing the correlation of the residuals and lagged residuals and then taking a pseudo-difference of each observation (the observation minus $\hat{\rho}$ of the lag of that observation, where $\hat{\rho}$ is the estimated correlation of the errors). In this procedure, the first observation for each unit is lost. This procedure can be improved by using the Prais-Winsten method, which retains the first observation for each unit (suitably transformed) while taking pseudo-differences of all observations other than the first one for each unit. Because there are N first observations and T is usually

not large, the Prais-Winsten procedure is superior to the more commonly used pseudo-differences.

Lagged Dependent Variable Models

Although the FGLS correction for serially correlated errors has good statistical properties, it has largely been supplanted in single time-series analysis by more modern treatments. The serial correlated error approach sees the dynamics as a nuisance that impedes estimation; more modern approaches add the dynamics to the basic specification. This corrects some theoretical oddities in the serially correlated errors approach.

Note that for the paradigmatic political example, the serially correlated errors model assumes that the effect of unmeasured variables on growth persists over time (declining geometrically), but the measured variables (the covariates) have only an immediate and non-dynamic impact. This is odd, as the covariates are presumably of more theoretical interest than the unmeasured variables in the error term.

While there are a variety of specifications used in modern time-series analysis, the annual data and smallish T in typical TSCS data limit the richness of time-series models that can be used. It is therefore common to use a model that simply adjoins the first lag of y to the model, yielding

$$y_{i,t} = \phi y_{i,t-1} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \epsilon_{i,t}; \quad i = 1, \ldots, N; \quad t = 1, \ldots, T. \tag{12}$$

Note that, as with the serially correlated errors model, a change in the error term has a persistent exponentially declining impact on y. But, in contradistinction to that model, a change in a covariate also has an effect that sets in over time (in the same exponential manner). This model also forces the analyst to explicitly recognize dynamics rather than thinking of those dynamics as simply a nuisance impeding estimation. [Eq. (12) could also include lags of exogenous variables, yielding a richer dynamic model without increasing estimation complexity. But the researcher should not expect that, with 20 or 30 years of annual data, it will be possible to distinguish between alternative dynamic models, so in many cases the simple lagged dependent variable model will be adequate.] This is not to say that Eq. (12) is always superior to the serially correlated errors model, but rather that, in general, it seems to be a better approach to modeling TSCS dynamics.

If the errors in Eq. (12) are serially uncorrelated, then it can be correctly estimated by OLS. If the errors are serially correlated, then OLS is inconsistent. Fortunately, it is easy to test for serially correlated errors using the same LM test as for the static model (adjoining the lag of y to the list of regressors in the auxiliary regression). The test statistic for whether the errors in Eq. (12) are uncorrelated is identical to that for Eq. (1).

It is interesting that for panel data, Eq. (12) with fixed effects causes very serious econometric problems. This is because with small T, the OLS estimate of ρ is biased downward, with the degree of bias inversely proportional to T. But this is much less of a problem in TSCS data, because T is almost always large enough to make this bias trivial.

It should also be noted that it is assumed that [in Eq. (12)] $\|\phi\| < 1$, that is, the model is stationary. Much recent research on single time series has focused on the nonstationary case, but little is known about nonstationary TSCS data. In many applications, including the paradigmatic political model, the dependent variable is a rate of growth, and hence very likely to be stationary. However, there will be models with a nonstationary dependent variable. Because the asymptotics of TSCS data are different than for single time series, the standard tests for nonstationarity may not be correct. Analysts estimating models with ϕ close to 1 should clearly worry about this issue. It may well be that a TSCS version of the "error correction" model espoused by Hendry *et al.* may be the appropriate model for nonstationary or near-nonstationary TSCS data.

This model assumes that in the short run there is some relationship between the covariates and y, which is modeled by $\Delta y_{i,t} = \Delta\mathbf{x}_{i,t}\boldsymbol{\beta} + \epsilon_{i,t}$. In addition, there is a long-run equilibrium between y and $\mathbf{x}$ modeled by $y_{i,e} = \mathbf{x}_{i,e}\boldsymbol{\gamma}$. When the system is out of equilibrium, y adjusts to that equilibrium at a rate of ϕ percent per year. Note that this is the single equation form of the error correction setup, in which the system returns to equilibrium by adjustments in y only. The error correction model is

$$\Delta y_{i,t} = \Delta\mathbf{x}_{i,t}\boldsymbol{\beta} - \phi\left(y_{i,t-1} - \mathbf{x}_{i,t-1}\boldsymbol{\gamma}\right) + \epsilon_{i,t}. \tag{13}$$

This can be estimated by OLS as long as the ϵ are stationary. While the typical N and T of TSCS data make the distribution of a Dickey-Fuller type statistic problematic, it will often be the case that the estimated serial correlation of the residuals is either so close to 1 or so far from 1 that nonstationarity (or stationarity) is obvious, regardless of any exact distribution of a test statistic. For the remainder of this article, stationarity is assumed, either of Eq. (12) or (13). Since Eq. (13) can be rewritten as a more complicated form of Eq. (12) (with levels and differences), for simplicity this article works only with Eq. (12), but everything should hold for the error correction formulation.

If Eq. (12) shows serially correlated errors, then the researcher must resort to complicated instrumental variable methods to estimate it. But if the LM test indicates that the null of uncorrelated errors cannot be rejected, OLS is the optimal estimation method. In practice, it is difficult to distinguish between a static equation with

serially correlated errors and the lagged dependent variable model. So, in many cases, the null hypothesis of serially independent errors will not be rejected and hence OLS can be used to estimate Eq. (12).

Instrumental variable estimation is itself problematic. Eq. (12) with serially correlated errors, can be rewritten as a model with uncorrelated errors but with more lags of both y and $\mathbf{x}$. Thus, if Eq. (12) shows serially correlated errors, it is surely worth examining (via LM tests) whether specifications with more lags of y or $\mathbf{x}$ result in serially uncorrelated errors.

This is particularly helpful to TSCS analysts, since they can then estimate dynamic models but use PCSEs, which are computed in the same manner as the static model. Consequently, while the old-fashioned procedure consists of testing for various error complications and then using FGLS to correct for these, the more modern approach estimates Eq. (12) with OLS and PCSEs and then tests whether the residuals are correlated, noting in practice that they seldom are. This simple estimation strategy allows analysts to focus on the important issue: the basic substantive specification. While all the usual issues surrounding choice of specification are present, TSCS data also present some unique issues that are discussed in the next section. But TSCS analysts will, of course, engage in all of the same specification searches and tests that are the stock and trade of regression analysts.

Specifying TSCS Models

Much of the debate in TSCS analysis has focused on how to estimate such models when the error structure is complicated. But, as has been seen, a simple estimation strategy is available. This leaves analysts to focus on the choice of specification. Here two strategies are considered: modeling heterogeneity and spatial issues. These are all done in the context of Eq. (12).

Heterogeneity

As noted, Eq. (12) implies complete pooling, meaning that all units have exactly the same relationship between the covariates and the dependent variable. (In this subsection, only heterogeneity in the mean function, that is, the relationship of the covariates to the expected value of the dependent variable, is considered.) Not only are the same covariates used for each unit, but also the coefficients on those covariates are assumed to be identical from unit to unit as well. (Note that for SUR, it is typically assumed that different covariates affect different units, since this is what gives SUR its power. But for TSCS models, it is almost invariably assumed that the covariates for all units are identical. This assumption is used here.)

The opposite of pooling is the completely unpooled model, in which each unit has its own set of coefficients. In dynamic form, this is

$$y_{i,t} = \phi_i y_{i,t-1} + \mathbf{x}_{i,t}\boldsymbol{\beta}_i + \epsilon_{i,t}; \;\; i = 1, \ldots, N; \;\; t = 1, \ldots, T. \tag{14}$$

If this specification is correct, then it is optimal to use unit-by-unit OLS. That is, it is optimal to run N separate OLS regressions, each on T observations (as always, assuming that specification tests show that this model is appropriate for each unit). This assumes that there is no relationship between the units and no possible information gains by imposing some degree of homogeneity on the coefficients (across the units). Analysts thus seem to be forced to choose between the assumption that the units are completely homogeneous or completely heterogeneous.

Assessing Heterogeneity

Because Eq. (12) is nested inside Eq. (14), it is easy to test which specification the data prefer. Taking as the null hypothesis that Eq. (12) is correct, a standard F-test comparing the sums of squared errors from the two specifications provides the best test between these two specifications.

But while this test is statistically straightforward, interpreting its results is less so. With a very large T, the researcher may end up rejecting the null of complete pooling even when the units are by and large homogeneous. There is less danger in accepting the null hypothesis due to a small T, since then unit-by-unit OLS will not be a feasible strategy.

It is also the case that some coefficients may be homogeneous while others may be heterogeneous. Eq. (14) uses up a substantial number of degrees of freedom. For example, it may be considered likely that ϕ, the speed of dynamic adjustment, is similar from unit to unit. Or we may be willing to allow the coefficients on controls that are not of theoretical interest to be homogeneous. Allowing for some coefficients to be homogeneous across units, while others are allowed to vary freely, can save many degrees of freedom and allow for efficient use of the data for the Ts typically seen in TSCS data. The researcher can easily modify the standard F-test to allow for the imposition of unit homogeneity in some coefficients.

The F-test approach may also mask another problem when most units are similar, but one is rather different from the others. With a large N, this heterogeneity may be missed; alternatively, one outlying unit may lead to the rejection of the pooled model, whereas the appropriate strategy is to estimate a pooled model on all but the outlying unit.

It is easy to assess whether there are one or a few outlying units via cross-validation. TSCS data present the analyst with a natural cross-validation strategy. The fully pooled model can be estimated leaving out one unit at a time, with the dependent variable for that omitted unit then being "predicted" based on the estimates that omitted that unit. While cross-validation is most commonly used for model selection, here the interest is in whether most units show low "prediction" error, with only one or a few showing high prediction error. If the units that have high prediction error are also those that appear different on theoretical grounds (e.g., geography, type of political or economic system), then the most sensible strategy would be to estimate the pooled model omitting the few units with high prediction error (of course, noting the change in "sample").

Estimation with Heterogeneity

There are attempts to steer an intermediate course between assuming complete pooling and no pooling whatsoever. These attempts all build on the classic work of Swamy, who proposed a "random coefficients model" (RCM) akin to Eq. (14), but in which the coefficients are joined by being draws from a normal distribution. He adjoined to Eq. (14) the assumption that $\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}$, the amount of unit-to-unit variation in the unit coefficients, is a hyperparameter to be estimated. Swamy and Hsiao have proposed various FGLS methods for estimating this model. Smith, and more recently Western, have proposed estimating this model in a Bayesian or empirical Bayes context; recent advances in Markov chain Monte Carlo methods now make this feasible.

The model appears attractive. One particularly attractive feature is that the analyst can modify the assumption about how the coefficients are generated to allow them to vary systematically by unit specific covariates (z_i), which measure items that do not vary over time, such as the political system. This results in $\boldsymbol{\beta}_i \sim N(\boldsymbol{\beta} + \mathbf{z}\gamma, \boldsymbol{\Gamma})$, where both γ and $\boldsymbol{\Gamma}$ are parameters to be estimated.

Although this model appears to be an attractive position in between two implausible extremes, and the RCM has proven extremely useful in hierarchical modeling, its usefulness in TSCS data is less clear. It does appear clear that the FGLS estimator proposed by Swamy and Hsiao does not work well in practice. This is because in order to ensure the positive definiteness of a covariance matrix, it assumes that sampling variance is zero. With the typical T of a TSCS data set, this is often incorrect, and leads to very poor performance of the Swamy/Hsiao estimator. This is unfortunate, as this is the RCM estimator seen in most computer packages.

Monte Carlo evidence indicates that the fully pooled model estimated via OLS does quite well unless there is either a very large T or a large amount of diversity in the unit $\boldsymbol{\beta}_i$. With a large T, unit-by-unit OLS [Eq. (14)] performs quite well, though such an estimator performs poorly for moderately sized Ts often seen in TSCS data. Although empirical Bayes estimators also perform well for a smallish T or relatively homogeneous $\boldsymbol{\beta}_i$, pooled OLS is better. For large T, unit-by-unit OLS is as good as the more complicated techniques. While more research is necessary, at this moment it appears that researchers can simply do an F-test to discriminate between Eqs. (12) and (14) and then do either fully pooled OLS or unit-by-unit OLS. The Monte Carlo evidence indicates that fully pooled OLS might be superior, even when the F-test marginally prefers the unpooled model. Choosing between the two methods thus requires a bit of art, and choice should be informed by theory as well as simple statistical tests, cross-validation, and other examinations of the data.

Spatial Variables

The units in TSCS data are typically spatially related. Note that analysts using FGLS were worried about whether the error process of one unit was related to the error process of other units. Spatial econometrics is a very complex arena by itself; most analyses deal with issues where interrelated units are observed only once. Spatial issues are simpler to analyze in the TSCS context.

FGLS Estimation

For concreteness, consider the analysis of economic growth as a function of economic and political variables in the 20 or so advanced industrial societies. The contemporaneously correlated errors approach assumes that the "errors" in the growth equation for one country are related to the error terms for other countries. Note that the problem with using FGLS to correct for this is that the correlation matrix of the errors was left free, leaving FGLS to estimate an inordinately large number of parameters in that matrix. Spatial analysis can dramatically improve on this.

Spatial analysts assume that the relationship between countries is proportional to some measure of distance. Whereas geographers focus on physical distance, political economists might assume that the interrelationship between the economies in two countries varies with the amount of trade they engage in. Letting D_{ij} be a measure of the distance between two countries (whether geographic or economic), the spatial approach replaces the non-zero covariances in Eq. (8) by

$$\mathrm{Cov}(\epsilon_{i,t}\epsilon_{i',t'}) = \begin{cases} \nu D_{i,j} & \text{if } t = t' \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

This then reduces the problems of FGLS drastically. FGLS estimation of the unparameterized correlated errors model has approximately $N^2/2$ estimated parameters that are not accounted for in the FGLS equations. If this error correlation is parameterized using spatial

notions, the number of estimated parameters that FGLS does not account for is reduced to 1. Such an approach would make FGLS a very reasonable choice of estimator for treating spatial correlation in the error process. At that point, it would be better to estimate this model via maximum likelihood, since the analyst would then obtain an estimate and standard error for the spatial correlation parameters, ν. To test the null of no spatial correlation of the errors, the analyst would then simply test the null $\nu = 0$. It should be noted that while this approach is difficult for standard cross-sectional spatial data, it is much simpler in the TSCS case, because in the latter case T replicates of the error process are observed, whereas in the former case only one such replicate is observed.

Adding Spatial Variables to the Specification

Allowing for spatially correlated errors is an improvement on the traditional FGLS estimator for contemporaneously correlated errors. But spatial notions can also be used in a more modern context, where spatial variables are included in the specification. The assumption of spatially correlated errors is equivalent to the assumption that unmeasured variables that affect growth in one country also affect growth in other countries. In this model, only measured variables pertaining to the unit affect the growth of that unit; unmeasured variables pertaining to all units affects the growth of each unit. This is odd because it is presumably the measured variables that are of interest.

It thus seems reasonable that the economic growth of one country is affected by the growth experienced by its neighbors, with the effect declining with distance (either geographic or economic). Then the specification called the "spatial lag," that is, the spatially weighted sum of growth in all of the other countries can be considered. Because it seems reasonable that neighbors have an effect only with some time lag, it might reasonably be expected that this spatially lagged variable should also be lagged one year. Note that this makes estimation simple. If, however, the contemporaneous spatial lag was included, this would return to the same very complicated estimation problem of the cross-sectional spatial analyst, in which there are T observations on an N-variate vector instead of $N \times T$ observations on scalar dependent variables.

This leads to a spatial and temporally dynamic specification:

$$y_{i,t} = \phi y_{i,t-1} + \lambda \sum_{j \neq 1} w_j y_{j,t-1} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \epsilon_{i,t}; \quad i = 1, \ldots, N; \quad t = 1, \ldots, T, \tag{16}$$

where the spatial weights, w_j, are defined *a priori* by the analyst. If the errors are temporally and spatially independent, this can be estimated by OLS and PCSEs. Such a procedure has been used informally by many TSCS modelers, who might, for example, include the lag of the trade-weighted average growth in all partner nations in their study. It is, of course, better to model spatial effects explicitly. This is an ongoing area of research in which much more study needs to be done, particularly of the appropriateness of using the temporally lagged spatial lag in the dynamic specification.

Current Issues

TSCS data present many interesting complications. The old-fashioned approach is to view these complications as nuisances that cause problems for OLS. The more modern approach is to add the TSCS features to the specification, explicitly modeling both dynamic and spatial features of the data. In many cases, this can be done via OLS with panel-corrected standard errors, and so leads to a relatively simple estimation problem.

So far, only TSCS data with a continuous dependent variable have been discussed. But there is often TSCS with a binary (or other discrete) dependent variable; the paradigmatic example is whether a pair of nations in conflict or not. While discussing this type of data is beyond the scope of this article, it should be noted that this is an extremely active area of research. There are at present two leading approaches for this type of data: correcting the error process in a generalized linear model setup, leading to Liang and Zeger's generalized estimating equation, and Beck, Katz and Tucker's event history approach (which is intimately related to the Markov switching model). But new breakthroughs in computational methods (Markov chain Monte Carlo) may allow for more direct modeling of binary TSCS data. At present, there is no definitive recommendation on how to model such data in general. But for the continuous case, OLS combined with panel corrected standard errors of a model that specifies the interesting features of TSCS data is the generally recommended method.

See Also the Following Articles

Fixed-Effects Models • Longitudinal Studies, Panel • Ordinary Least Squares (OLS) • Time Series Analysis in Political Science • Yule, George Udny

Further Reading

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic, New York.

Beck, N., and Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *Am. Pol. Sci. Rev.* **89,** 634–647.

Beck, N., Katz, J. N., and Tucker, R. (1998). Taking time seriously: Time-series–cross-section analysis with a binary dependent variable. *Am. J. Pol. Sci.* **42,** 1260–1288.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press, New York.

Franzese, R. J. (2002). *The Political Economy of Macroeconomic Policy in Developed Democracies.* Cambridge University Press, New York.

Hendry, D., Pagan, A., and Sargan, J. D. (1984). Dynamic specification. In *Handbook of Econometrics* (Z. Griliches and M. Intriligator, eds.), Vol. 2, Ch. 18. North-Holland, Amsterdam.

Hsiao, C. (1986). *Analysis of Panel Data.* Cambridge University Press, New York.

Parks, R. (1967). Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated. *J. Am. Stat. Assn.* **62,** 500–509.

Swamy, P. A. V. B. (1971). *Statistical Inference in Random Coefficient Models.* Springer Verlag, New York.

Western, B. (1998). Causal heterogeneity in comparative research: A Bayesian hierarchical modelling approach. *Am. J. Pol. Sci.* **42,** 1233–1259.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *J. Am. Stat. Assn.* **57,** 500–509.

Time–Space Modeling

Donald G. Janelle
University of California, Santa Barbara, Santa Barbara, California, USA

Glossary

agent-based spatial models Models that simulate an agent's decisions and relationships to other agents through space and time based on the modeler's designated rules of agent behavior.

cellular automata A procedure for simulating dynamically changes in the characteristics of cells, arrayed usually as a grid to represent an environment (e.g., demographic transition in an urban neighborhood).

human extensibility The capability of a person or institution to exert influence beyond their current location and time (e.g., via telecommunications or through application of economic and political power).

time geography The study of the temporal dimensions of human spatial behavior and of the embeddedness of time in geographic patterns.

time–space compression The accelerated growth of events in a person's life and the intensified uses of space through time.

time–space convergence/divergence The rate (e.g., minutes per year) at which places (e.g., cities) move closer together or further apart as a result of technological changes or congestion factors that impact on the travel time or communication time between them.

time–space distanciation The spatiotemporal outreach of a social system to form larger organizational entities and agents of change.

time–space path A line of movement behavior of a person through time (usually represented as a single dimension of clock time on the vertical axis of a graph) and over space (usually represented on the horizontal axis as a single dimension of distance or as a geographic surface of two dimensions). The time–space path is a central feature of Torsten Hägerstrand's time geography model of society. Depending on the temporal scale, it may be a daily path or a lifeline.

time–space-prism A space that delimits the movement possibilities and activity choices of a person, usually based over a defined time period such as 24 hours.

This article discusses concepts and research methods that seek to merge time and space into an integrated time–space depiction of process. It reviews modeling frameworks used to represent time–space structures of human behavior, including the application of new tools that seek a more dynamic depiction of change and development. Most of the examples are drawn from the disciplines of human geography, regional science, and related fields.

Representing Time–Space Processes

Many significant societal issues are best understood and dealt with as time–space in nature. Examples include diffusion processes such as epidemics, which may be guided by complex patterns of human behavior, spatial structures of human contacts, facilities for movement, as well as the intrinsic temporal parameters of biological transfer among hosts and susceptible populations. Similarly, crime events in cities may exhibit temporal sequence, which may not be independent of spatial patterns and structures.

Although the term *process* implies some notion of continuity of change over time and space simultaneously, researchers have focused largely on one dimension or the other as opposed to treating time–space operationally as a single integrated framework. This is due, in part, to the scarcity of appropriate data resources and to the immaturity of methodological means to treat process in an explicit manner. Thus, for example, time series data on economic indicators are usually investigated for trends through time while ignoring patterns of change in such indicators over space. The temporal dimension is usually represented at highly

aggregate levels in space (e.g., the nation). The availability of data (especially from the census) has conditioned many images of landscape change to a series of snapshots 10 years apart. Examples include population density patterns or household income distributions within cities. Since the census and most customized social surveys typically focus on discrete points in time, social scientists have not had access to continuous data for study of processes in their full temporal–spatial contexts.

Early attempts to understand the changing character of geographic space over time often skirt data constraints by using spatial concepts to envision landscapes at different points in time. The identification of historical phases in the development of urban settlement systems is an example, typified by the idealized stage model proposed by Taaffe *et al*. in 1963 to account for the sequence and extent of transportation development and urban growth in West Africa. Although such approaches fall short of embracing a full understanding of how the world works, they are useful for generating hypotheses and suggesting approaches to empirical verification, thereby enhancing prospects for predictions of spatial patterns over time.

An interesting example of a time–space perspective to modeling is represented by the popular computer game, SimCity (Maxis Software, Division of Electronic Arts). SimCity simulates the development of an urban landscape based on a complex set of interdependent decisions by the player. This involves allocation of land resources to given activities in an attempt to satisfy both individual and community needs, the establishment of employment centers, and the raising of tax revenue. In turn, these decisions are embedded with both positive and negative spillover effects at neighborhood and regional levels—seen in terms of environmental impact, traffic flows, land values, criminal activity, and public assessment. Unforeseen events related to the economy, political sentiment, and natural forces add interest and test the robustness and resiliency of the city as it evolves in time and space. The graphic displays in SimCity capture the dynamics of change in space and time, allowing players to alter the scale of representation, explore different scenarios of social and economic change or land use design, and visualize information about the city's development in different data formats (e.g., as maps, trend lines, and tables).

Within a framework of simplified rules and impressive graphic display, SimCity captures much of what might be embraced in a time–space perspective on social and behavioral change. Game results may be explored in greater depth to describe and predict the evolution of networks and infrastructure over time, the diffusion of land uses, changes in population density, and implications for individual behavior. Although the modeling of game processes reflects understanding of the gaming environment, it would be presumptuous to claim equivalent specificity in the time and space attributes of social change. Data constraints and confidentiality issues in the treatment of information about people and firms usually preclude the same level of detail for modeling or analyzing human systems.

Scientific efforts to describe human behavior and organizational systems, explore relationships among variables across space and time, and predict outcomes as time–space patterns are constrained by the underlying complexities of the processes at work, the adequacy of theory and measurement tools, and the nature of standard data resources at the disposal of researchers. While recognizing these limitations, it is useful to highlight some of the methodologies that have been applied. Of the approaches most commonly used, diffusion and migration modeling and spatial point processes are discussed elsewhere in the encyclopedia. Time–geography, cellular automata simulations, and agent-based spatial models are introduced in the following section. This is followed by a review of time–space concepts that have yet to be incorporated in existing modeling frameworks and by discussion of technical developments that may alter standard methodologies for modeling time–space processes.

Time–Space Modeling Approaches

Statistical analysis, based mostly on regression techniques, is widely used to model trends in variable relationships across time in hopes of explaining values of a dependent variable by a set of independent variables. This methodology relies on historical data and is most appropriate in circumstances in which theoretical understanding within a knowledge domain is weak. In accounting for spatial variation in dependent variables at a point in time or for analysis of changes over a period of time across the spatial units of observation, these methods raise concerns regarding the dependency of results according to variability in sizes and shapes of spatial units, gaps in time between data points, and periods of measurement. Additional concerns about time series autocorrelation and spatial autocorrelation are dealt with successfully with econometric methods and local indicators of spatial association. Nonetheless, difficulties in treating simultaneously the assessment of temporal and spatial variation continue to limit applications of statistical methods in modeling.

As an alternative to statistical modeling, dynamic models may use simultaneous equations to describe systems based on theories that purport to capture functional and causal linkages in the flows of resources and changes over time. In these models, system dynamics are describable by a set of state variables (stocks), controls over rates of flows, and parameters of the system. A systems approach can be used to explore "what if" scenarios that give insight to system responses to changes in policies and different parameter values. The aggregate analytical

approach to dynamic modeling lends itself to predictive uses. However, needs for analytical or numerically intensive solutions, conditions of equilibrium, and other simplifying assumptions have tended to constrain the complexity and applicability of these models. These kinds of models have been used to explore the functioning of environmental systems and have also seen application in attempts to model the dynamics of regional economies.

A general critique of systems modeling is embraced in structuration theory, which argues for a more recursive link between independent human agencies and "the system" and for greater emphasis on the contextuality of human actions. Increasingly, researchers have turned to microlevel methods and concepts for describing and simulating time–space processes and to methods that allow for representation of human intention in making decisions. The following sections treat some of the approaches that are explicitly time–spatial in their formulation and that hold promise for representation of human agency. These include time geography and simulations based on cellular automata and agent-based modeling.

Time Geography

Torsten Hägerstrand's time geography model of society treats individual activity behavior explicitly as a time–space process and views individual behavior as the building block of larger social systems. In the time geography perspective, the activities and movement paths (or time–space paths) of individuals are subject to a set of constraints. These constraints link the individual to a broader system of social ties (couplings), controls over the use of space and facilities (authority), and differential access to the means of overcoming distance and to requirements for meeting personal needs for rest and nourishment (capabilities). Hägerstrand's motivation for casting human behavior in this way was to provide a method to assess the impact of policy decisions on the freedom of action that people have within regional or urban settings. The modeling approach recognizes that human activity choices of any kind have finite durations, are linked to prior activities, and condition future choices. At any given point along a time–space path, a person has a limited degree of freedom to make choices—represented by Hägerstrand as a time–space prism. The prism defines the outer boundaries of locations that a person can access given his or her level of mobility (speed of movement) and the amount of time at his or her disposal between any designated set of activities. Lenntorp provides the classic application of this approach, modeling the possibilities for residents to engage in different activities by means of public transportation in Karlstad, Sweden. Burns used it to model tradeoffs between time and space in the assessment of options for urban transportation planning. Although the time geography modeling approach has been slow in seeing wide application, researchers have linked it through geographic information systems (GIS) to the analysis and representation of time–space travel surveys and have used it to characterize human behavior as dynamic geographies and as a basis for modeling accessibility. Time geography modeling lends itself to linkages with other modeling approaches, including event history analysis, cohort studies of demographic and behavioral change, and agent-based spatial modeling.

Cellular Automata

Cellular models, including cellular automata (CA), provide simple representations of dynamic systems and have been especially popular for simulating environmental change at a range of geographical and temporal scales. CA are conceived typically as a grid of cells, with the character (state) of cells subject to a set of transition rules about the rates and likelihood of change over time. These rules usually include the effects of neighboring cells on each given cell (e.g., as in a contagion process, such as gentrification within the central region of cities). Implementation of the rules is used to generate a dynamic computer display of change over time. Thus, the elemental components of a CA model include the lattice of cells, a set of discrete states for cells, a specified neighborhood of surrounding cells, transition rules for cell transfer from one state to another, and a time step (e.g., 1 year) for updating the state status of all cells. Cell dynamics may be constrained further by introduction of external factors (e.g., by some measure of performance for the national economy) and by relaxation of assumptions about the homogeneity of cells. Thus, a CA model on urban land use transition might weight certain cells for housing value based on the expected amenity value of a cell's physical setting and the quality of neighboring houses. An early use of the CA approach was Tobler's dynamic depiction of land use change in the Detroit region. Clarke *et al*. offer a another example, and Batty *et al*. provide a general overview of the CA approach in the context of urban dynamics. Among the weaknesses of cellular models is the inability to reflect the role of complex decision making by human agents. However, cellular frameworks may be combined with Markov models, in which transition rules are treated probabilistically and may be conditioned by temporal lags in cell response. In addition, they can be combined within the framework of agent-based simulations that focus explicitly on the behavior of agents and the interdependency of such behavior on cell characteristics.

Agent-Based Spatial Modeling

Agent-based modeling (ABM) provides an attractive methodology for linking social and behavioral theory

within the explicit time–space context of the behavioral environment. The agents represent decision makers who act and interact with others in an environment according to rules of behavior and defined bases of motivation. The environments may be based on replication of an established behavioral setting (e.g., a market) and could be set forth within the framework of a CA model. Some of the underlying assumptions of these models (that the systems are dynamic and evolving and not in equilibrium) are congruent with real-world patterns of change. Within this framework, microlevel individual decisions may have effects at other scales; for example, consumer selection of housing within an area of a city for personal reasons could impact the composite urban structure of land values, demand for classrooms, commercial potential, and the spatial distribution of social needs both locally and regionally. Because these diverse impacts relate to the behavior of multiple decision makers, these models are frequently defined as multiagent systems (MAS). Gilbert and Troitzsch present a thorough introduction to agent modeling. ABM/MAS offer a way to imitate behavior within a dynamic modeling framework. Examples of applications in the time–space context include the modeling of pedestrian flows in business districts, intraurban household migration, and social segregation processes.

The ability to use "representative" data as opposed to empirically derived data sets and the ability to input theoretical constructs into the model of decision making by agents are intrinsically attractive given the difficulty in securing adequate data for exploration of many social science research questions. However, this flexibility also raises issues regarding the validity of models and the verification of model outputs and their interpretation. The use of realistic parameters, possibly derived from empirical investigation, is one way that researchers have attempted to deal with this issue. CA and ABM provide a means of assessing the current state of knowledge regarding time–space processes, they help reveal gaps in our understanding of behavior, and they have the potential to demonstrate how processes at one scale (the microscale) can impact in unexpected ways on patterns of human organization and resource use at more macroscales. Through systematic exploration of parameter values to determine a model's sensitivities to a range of plausible behaviors, it may be possible to enhance the acceptance of predictions based on data that are representative of real-world situations.

Merging Time and Space in Social–Behavioral Research

Social science theory treats time–space as an embedded structure of human social and economic systems. Four general concepts about such systems address explicitly their spatiotemporal attributes. For the most part, time–space modeling approaches have not attempted to include these concepts, but they represent significant challenges in attempts to capture and interpret the dynamic structures of human environments. These concepts are time–space convergence, human extensibility, time–space distanciation, and time–space compression. Collectively, these ideas relate to processes that alter the significance of space in the functioning of social and economic systems. Brief discussions of each of these concepts and their relationships to one another expose issues that have not been fully explored in time–space modeling methodologies.

Time–Space Convergence

Time–space convergence and the related concept of time–space divergence treat space as a product of human efforts to reduce the travel time and travel cost between places. For example, as a consequence of improved rail and highway infrastructure, Boston and New York City converged on each other at an average rate of 26 minutes per year between 1800 and 1960. Such measures may help in the interpretation of social and economic responses to reduced constraints on human interactions, but they also mask complexities that are inherent to convergence processes. For instance, measures will differ depending on access to modes of movement and to social class differences that may deny use of preferred (faster) modes to some. Thus, there exist multiple representations of time–space convergence for any single place or region. Among the multiple of places that make up urban systems, the variability of convergence rates characterizes a non-Euclidean geometry that complicates any simple visualization (e.g., a map) of results. Such formidable difficulties in establishing an empirical framework have no doubt impeded research on convergence/divergence processes.

Human Extensibility

Human extensibility is the reciprocal of time–space convergence. It describes how the relaxation of constraints on movement allows people to extend their presence beyond their current locations. The nature and degree of extensibility are unique to each individual or to general categories of people divided by social cleavages across income, class, and other attributes. Empirical work in this area has been extended by Paul Adams based on detailed surveys of daily behavior of individuals. The emergence of the Internet and of multiple global communication systems has sparked renewed interest in this concept. General issues relate to the connectedness of people within communities and to disconnections among communities versus broader

levels of regional and global consciousness. New methodologies are needed to expand empirical understanding of the extensibility processes in different behavioral situations.

Time–Space Distanciation

Time–space distanciation is a related process that merges time with space. Advanced by sociologist Anthony Giddens in relation to his structuration theory, this concept illustrates how convergence and extensibility processes go beyond the context of individuals and places to shape the emergent organization of entire social systems, one example being the intensified globalization of economic activity.

Time–Space Compression

Time–space compression is another dimension to the collapse of space through time. David Harvey pioneered this idea based on the intensification of events per unit of time and per unit of space. It links a process of accumulation in the Marxian sense with the daily lives of people, with the appreciation of some places over others, and with the resulting enlargement in disparities among regions. This notion adds the dimension of experiential meaning associated with the annihilation of space through time.

The concepts of convergence, extensibility, distanciation, and compression are all central to understanding the time–space context of human activities and the systems in which they take place. They offer insight on how the individual fits within the broader system of time–space structures. The challenge for researchers is to incorporate these ideas within more formal modeling frameworks, be they agent-based models, time geography, or dynamic systems approaches.

Developments in Spatiotemporal Representation and Modeling

Two developments stand out with regard to their likely future impact on advances in time–space modeling for the social and behavioral sciences: the emergence of new data visualization technologies and the growth of on-demand, real-time information systems that use technologies that know where they are at any point in time [i.e., location-based services (LBSs) that link individuals within a framework of georeferenced information about the world around them].

Visualization

Advances in graphical visualization methods offer ways to explore representations of time–space processes. Video capture of simulations or graphical animations of changing geographical patterns (dynamic maps) are expected to refine capabilities for time–space modeling. For instance, slowing down or speeding up processes may facilitate pattern detection. Animations at different resolutions of time and space may expose process linkages at multiple scales. Other options include the ability to explore possible causal linkages among processes both as lagged and nonlagged time series across different spatial scales. Embedding these capabilities within a GIS offers yet additional scope to application of analytical modeling methods. Peuquet explores some of the computational and data modeling issues that relate to efforts of integrating space and time within GIS software environments. In addition, new tools of interactive spatial data analysis are becoming more sensitive to the needs of time–space analysis.

Continuous Data Capture in Space and Time

Real-time capture of information at the level of the individual person or for specific locations is based on the locational awareness of new communication technologies that are linked with geographically referenced information systems. The merging of the laptop computer, personal digital assistant, Internet, database, and telephone capabilities into personal, wearable, wireless information utilities increases the scope for new types of human behavior and for new research methodologies in the social and behavioral sciences. It is an open question as to what new forms of behavior might emerge or what LBS functions government or business might provide. Nonetheless, it is likely that LBS technologies will influence the spatial and temporal organization of society, offering opportunities for new forms of retailing, work activity, governance, emergency care delivery, services for the blind and deaf, criminal behavior, and search-and-find methods.

From the perspective of science, these technologies offer options for sensing processes as they occur, sampling human behavior in a time–space framework, carrying out dynamic calculations and mappings, and analyzing information as it is collected. The capabilities to retrieve information based on where one is at any given time or to target information to respondents based on where they are open possibilities for dynamic modeling of time–space attributes of social and behavioral processes. These new tools may substantially alter the methods of primary data capture in field research. The real-time recording of time–space activity diaries for respondents provides a basis for testing the theoretical foundations of time geography modeling. Linking of such data with computational capabilities could yield new methods for evaluating ABM simulations. Of course, new data mining

and visualization tools will be needed to exploit these potentials. At the same time, these capabilities may burden the ethical foundations of behavioral research with issues regarding the protection of individual privacy and autonomy from unwanted surveillance and from the time–space profiling of individuals or regions. Time–space modeling is entering an exciting era of potential and challenges.

See Also the Following Articles

Spatial Autocorrelation • Spatial Econometrics • Spatial Pattern Analysis

Further Reading

Adams, P. C. (2000). Application of a CAD-based accessibility model. In *Information, Place and Cyberspace: Issues in Accessibility* (D. G. Janelle and D. C. Hodge, eds.), pp. 217–239. Springer-Verlag, Berlin.

Anselin, L. (1995). Local indicators of spatial association LISA. *Geographical Anal.* **27,** 93–115.

Bailey, T. C., and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis.* Longman, Essex, UK.

Barentsen, W., and Nijkamp, P. (1989). Modelling non-linear processes in time and space. In *Advances in Spatial Theory and Dynamics* (A. E. Andersson, D. F. Batten, B. Johansson, and P. Nijkamp, eds.), pp. 175–192. Elsevier, Amsterdam.

Batty, M., Xie, Y., and Sun, Z. (1999). Modelling urban dynamics through GIS-based cellular automata. *Comput. Environ. Urban Systems* **23,** 205–233.

Burns, L. D. (1979). *Transportation, Temporal, and Spatial Components of Accessibility.* Lexington Books, Lexington, MA.

Clarke, K. C., Hoppen, S., and Gaydos, L. (1997). A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environ. Planning B Planning Design* **24**(2), 247–262.

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration.* Polity, Cambridge, UK.

Gilbert, N., and Troitzsch, K. G. (1999). *Simulation for the Social Scientist.* Open University Press, Buckingham, UK.

Hägerstrand, T. (1973). The domain of human geography. In *Directions in Geography* (R. J. Chorley, ed.), pp. 67–87. Methuen, London.

Harvey, D. (1990). Between space and time: Reflections on the geographical imagination. *Ann. Assoc. Am. Geographers* **80,** 418–434.

Janelle, D. G. (1969). Spatial reorganization: A model and concept. *Ann. Assoc. Am. Geographers* **59,** 348–364.

Janelle, D. G. (1973). Measuring human extensibility in a shrinking world. *J. Geogr.* **72**(5), 8–15.

Kwan, M.-P. (2000). Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. *Transport. Res. C* **8,** 185–203.

Lenntorp, B. (1976). Paths in space–time environments: A time–geographic study of movement possibilities of individuals. *Lund Stud. Geogr.* **44**.

Miller, H. J., and Wu, Y. H. (2000). GIS for measuring space–time accessibility in transportation planning and analysis. *GeoInformatica* **4**(2), 141–159.

Peuquet, D. J. (2002). *Representations of Space and Time.* Guilford, New York.

Taaffe, E. J., Morrill, R. L., and Gould, P. R. (1963). Transportation expansion in underdeveloped countries: A comparative analysis. *Geographical Rev.* **53,** 503–529.

Tobler, W. (1979). Cellular geography. In *Philosophy in Geography* (S. Gale and G. Olsson, eds.), pp. 379–386. Reidel, Dordrecht, The Netherlands.

Total Survey Error

Tom W. Smith
University of Chicago, Chicago, Illinois, USA

Glossary

bias/systematic error An error that creates a difference between the measured and true overall mean values.
context effects Differences in measurements resulting from the order and content of prior items.
interviewer effects Differences in measurements resulting from differences in the characteristics or interviewing behaviors of those conducting an interview.
mode effects Differences in measurements resulting from method of administration (e.g., in person, by telephone, self-completion).
nonresponse Failure to respond to or participate in a survey (unit or survey nonresponse) or to answer a question in a survey (item nonresponse).
response option effects Differences in measurements resulting from the content, number, or order of the answer categories to closed-ended questions.
variance/variable error A random error with no expected effect on the overall mean values.

Total survey error is the sum of two components: (1) variance, or variable error, which is random and has no expected impact on mean values, and (2) bias, or systematic error, which is directional and alters mean estimates. Variable error consists of sampling and nonsampling (collection and processing) errors. Collection errors are further broken down into errors associated with mode, instrument, interviewer, and respondent. Processing errors consist of errors related to coding, data entry, data transfer, and documentation. Systematic errors also consist of sampling and nonsampling errors. Sampling errors may relate to the sample frame, selection, or statistical inference. Nonsampling errors consist of nonobservational and observational errors. The former results from either noncoverage or nonresponse and the latter results from errors in collection, processing, and analysis. As with variable errors, collection errors are related to mode, instrument, interviewer, and respondent, and processing errors are related to coding, data entry, data transfer, and documentation. Analysis errors can be conceptual, statistical, or presentational.

Introduction

Total survey error sums up all of the myriad ways in which measurement can be wrong. As Judith Lessler has noted, total survey error is "the difference between its actual (true) value for the full target population and the value estimated from the survey." Total survey error comes in two varieties: (1) variance, or variable error, which is random and has no expected impact on mean values, and (2) bias, or systematic error, which is directional and alters mean estimates. Total survey error is the sum of these two components.

The concept of total survey error goes back at least to the 1940s; the term was in general use by the 1960s and is now frequently invoked in general discussions of survey error. As Fig. 1 illustrates, total survey error has many components. Each component must be considered to understand the total error structure of a survey.

Variable Error

Looking at variable error first, the most well-understood and frequently discussed source of error is sampling variance, which is the variability in estimates that results from using a random subset of observations to represent all units in the population. Sampling error is usually overemphasized because it is the only component of total survey error that can be readily quantified into confidence intervals and levels of statistical significance. It is also usually

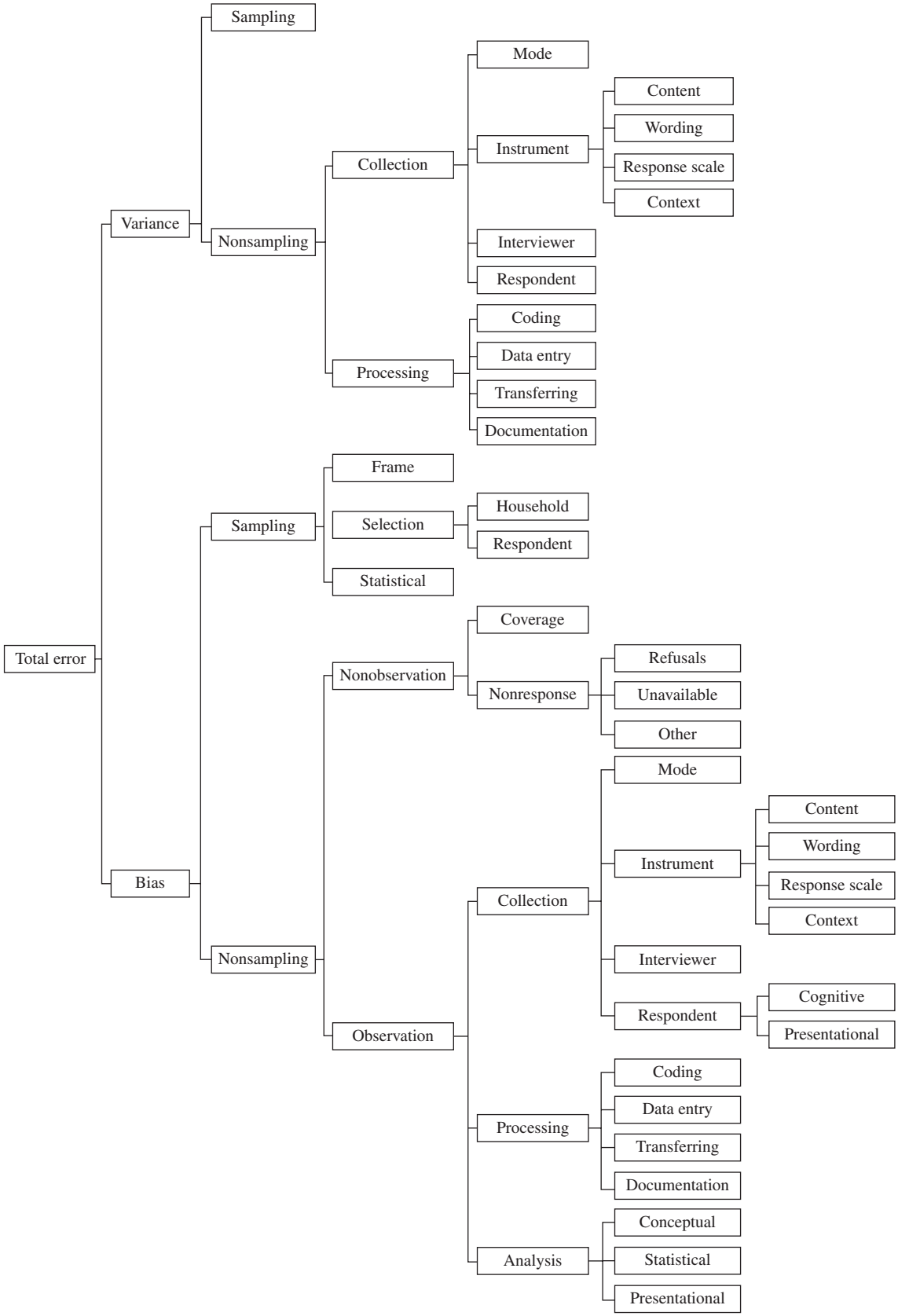


Figure 1 Total survey error.

misreported because estimates of sampling error commonly ignore both the design effects associated with nonsimple random samples and the variable-by-variable differences in sampling variance associated with marginal distributions and clustering.

The other major component of variable error is nonsampling variability, which consists of collection and processing errors. The main aspects of collection error concern differences resulting from mode of administration, instrument format, interviewer, and respondent. Mode of administration differs in three major ways: format (audio, visual, mixed), presenter (interviewer administered, self-administered, mixed), and technology (none/in person, paper, telephone, computer, mixed). Each of these elements can be combined to form a particular method of surveying. For example, using an audio format, interviewer-administered survey with no technology would be the traditional, face-to-face interview—the paper-and-pencil interview (PAPI). Another method would be the visual format, self-administered, and computer-driven interview—the CASI (computer-assisted, self-interview). Each of the many mode variants has unique measurement aspects and error profiles, depending on differences in factors such as the literacy and eyesight and hearing capabilities of the respondents, load capacity, and personal vs. nonpersonal interaction.

In many ways, the survey instrument (i.e., the questionnaire) is the heart of the survey. First, content (i.e., what type of information is being collected) is crucial. For example, a better developed index with more items will produce a more reliable measurement, compared to a less reliable scale. Second, the wording of specific items will affect measurement reliability (but, as discussed later, question wording differences are more likely to be associated with systematic than with variable error). Third, the choice of response scale affects variability. For example, much rounding occurs in many count estimates, and when a "101° feeling thermometer" scale is used, almost all responses are limited to "temperatures" (numerical rankings or ratings) divisible by 10 or 25 (e.g. 40 or 75). Finally, the ordering of questions creates context effects by which earlier items affect the responses to later items. This is usually related to bias, but when order is randomized to minimize context effects (as is often done in computer-assisted interviews), then the order effects essentially become part of the survey variance.

Interviewers also introduce random error. They may unintentionally go to the wrong household, select the wrong respondent in the household, use the wrong version of the questionnaire, etc. Also, almost 30% of the time, interviewers do not ask a question exactly as it is written. The frequency of interviewer mistakes depends in large part on the level of training, experience, and supervision. Of course, error also comes from respondents. They will mishear or misread questions and misspeak or mismark responses. These inadvertent errors will increase when respondents' interest and motivation are low, when there are cognitive, language, or other impediments, and in other circumstances (e.g., when there are time pressures or distractions during the interview).

Processing errors occur when data are being transferred or transformed. Such errors can be minimized by instituting and enforcing quality-control procedures. Coding of open-ended material can be enhanced by the development of detailed, comprehensive coding schemes and close supervision. In particular, items can be coded more than once by different coders and intercoder reliability measures can be calculated. Similarly, data entry errors can be measured and minimized by double-entry verification. Likewise, when transferring data, such as transcribing recorded interviews or converting from data collection programs to data analysis programs, careful cross-checking is needed. Finally, solid documentation is needed to avoid mistakes. Question wordings may be paraphrased or misreported, codes may be mislabeled, and wild punches or unexplained values may show up in the final data set.

Bias, or Systematic Error

Turning to bias, or systematic error, there is also a sampling component. First, the sample frame (i.e., the list or enumeration of elements in the population) may either omit or double count units. For example, the U.S. Census both misses people (especially African-Americans and immigrants) and counts others twice (especially people with more than one residence), and samples based on the census reflect these limitations. Second, certain housing units, such as new dwellings, secondary units (e.g., basement apartments in what appears to be a single-family dwelling), and remote dwellings, tend to be missed in the field. Likewise, within housing units, certain individuals, such as boarders, tend to be underrepresented and some respondent selection methods fail to work in an unbiased manner (e.g., the last/next birthday method overrepresents those who answer the sample-screening questions). Third, various statistical sampling errors occur. Routinely, the power of samples is overestimated because design effects are not taken into consideration. Also, systematic sampling can turn out to be correlated with various attributes of the target population. For example, in one study, both the experimental form and respondent selection were linked by systematic sampling in such a way that older household members were disproportionately assigned to one experimental version of the questionnaire, thus failing to randomize respondents to both experimental forms.

Nonsampling error comes from both nonobservational and observational errors. The first type of

nonobservational error is coverage error, in which a distinct segment of the target population is not included in sample. For example, in the United States, preelection random-digit-dialing (RDD) polls want to generalize to the voting population, but systematically exclude all voters not living in households with telephones. Likewise, samples of businesses often underrepresent smaller firms. The second type of nonobservational error consists of nonresponse (units are included in the sample, but are not successfully interviewed). Nonresponse has three main causes: refusal to participate, failure to contact because people are away from home (e.g., working or on vacation), and all other reasons (such as illness and mental and/or physical handicaps).

Observational error includes collection, processing, and analysis errors. As with variable error, collection error is related to mode, instrument, interviewer, and respondent. Mode affects population coverage. Underrepresentation of the deaf and poor occurs in telephone surveys, and of the blind and illiterate, in mail surveys. Mode also affects the volume and quality of information gathered. Open-ended questions get shorter, less complete answers on telephone surveys, compared to in-person interviews. Bias also is associated with the instrument. Content, or the range of information covered, obviously determines what is collected. One example of content error is when questions presenting only one side of an issue are included, such as is commonly done in what is known as advocacy polling. A second example is specification error, in which one or more essential variable is omitted so that models cannot be adequately constructed and are therefore misspecified.

Various problematic aspects of question wordings can distort questions. These include questions that are too long and complex, are double-barreled, include double negatives, use loaded terms, and contain words that are not widely understood. For example, the following item on the Holocaust is both complex and uses a double negative: "As you know, the term 'holocaust' usually refers to the killing of millions of Jews in Nazi death camps during World War II. Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?" After being presented with this statement in a national U.S. RDD poll in 1992, 22% of respondents said it was possible that the Holocaust never happened, 65% said that it was impossible that it never happened, and 12% were unsure. Subsequent research, however, demonstrated that many people had been confused by the wording and that Holocaust doubters were actually about 2% of the population, not 22%. Error from question wording also occurs when terms are not understood in a consistent manner.

The response scales offered also create problems. Some formats, such as magnitude measurement scaling, are difficult to follow, leaving many, especially the least educated, unable to express an opinion. Even widely used and simple scales can cause error. The 10-point scalometer has no clear midpoint and many people wrongly select point 5 on the 1–10 scale in a failed attempt to place themselves in the middle. Context, or the order of items in a survey, also influences responses in a number of quite different ways. Prior questions may activate certain topics and make them more accessible (and thus more influential) when later questions are asked. Or they may create a contrast effect under which the prior content is excluded from later consideration under a nonrepetition rule. A norm of evenhandedness may be created that makes people answer later questions in a manner consistent with earlier questions. For example, during the Cold War, Americans, after being asked if American reporters should be allowed to report the news in Russia, were much more likely to say that Russian reporters should be allowed to cover stories in the United States, compared to when the questions about Russian reporters were asked first. Even survey introductions can influence the data quality of the subsequent questions.

Although social science scholars hope that interviewers merely collect information, in actuality, interviewers also affect what information is reported. First, the mere presence of an interviewer usually magnifies social desirability effects, so that there is more underreporting of sensitive behaviors to interviewers than when self-completion is used. Second, basic characteristics of interviewers influence responses. For example, Whites express more support for racial equality and integration when interviewed by Blacks than when interviewed by Whites. Third, interviewers may have points of view that they convey to respondents, leading interviewers to interpret responses, especially to open-ended questions, in light of their beliefs.

Much collection error originates from respondents. Some problems are cognitive. Even given the best of intentions, people are fallible sources. Reports of past behaviors may be distorted due to forgetting the incidents or misdating them. Minor events will often be forgotten, and major events will frequently be recalled as occurring more recently than was actually the case. Of course, respondents do not always have the best of intentions. People tend to underreport behaviors that reflect badly on themselves (e.g., drug use and criminal records) and to overreport positive behaviors (e.g., voting and giving to charities).

Systematic error occurs during the processing of data. One source of error relates to the different ways in which data may be coded. A study of social change in Detroit initially found large changes in respondents' answers to the same open-ended question asked and coded several decades apart. However, when the original open-ended responses from the earlier survey were recoded by the same coders who coded the latter survey, the differences virtually disappeared, indicating that the change had been

in coding protocols and execution, not in the attitudes of Detroiters. Although data-entry errors are more often random, they can seriously bias results. For example, at one point in time, no residents of Hartford, Connecticut were being called for jury duty; it was discovered that the new database of residents had been formatted such that the "d" in "Hartford" fell in a field indicating that the listee was dead. Errors can also occur when data are transferred. Examples include incorrect recoding, misnamed variables, and misspecified data field locations. Sometimes loss can occur without any error being introduced. For example, 20 vocabulary items were asked on a Gallup survey in the 1950s and a summary scale was created. The summary scale data still survive, but the 20 individual variables have been lost. Later surveys included 10 of the vocabulary items, but they cannot be compared to the 20-item summary scale.

Wrong or incomplete documentation can lead to error. For example, documentation on the 1967 Political Participation Study (PPS) indicated that one of the group memberships asked about was "church-affiliated groups." Therefore, when the group membership battery was later used in the General Social Surveys (GSSs), religious groups were one of the 16 groups presented to respondents. However, it was later discovered that church-affiliated groups had not been explicitly asked about on the earlier survey, but that the designation had been pulled out of an "other-specify" item. Because the GSS explicitly asked about religious groups, it got many more mentions than had appeared in the PPS; this was merely an artifact of different data collection procedures that resulted from unclear documentation.

Most discussions of total survey error stop at the data-processing stage. But data do not speak for themselves. Data "speak" when they are analyzed, and the analysis is reported by researchers. Considerable error is often introduced at this final stage. Models may be misspecified, not only by leaving crucial variables out of the survey, but also by omitting such variables from the analysis, even when they are collected. All sorts of statistical and computational errors occur during analysis. For example, in one analysis of a model explaining levels of gun violence, a 1 percentage point increase from a base incidence level of about 1% was misdescribed as a 1% increase, rather than as a 100% increase. Even when a quantitative analysis is done impeccably, distortion can occur in the write-up. Common problems include the use of jargon, unclear writing, the overemphasis and exaggeration of results, inaccurate descriptions, and incomplete documentation. Although each of the many sources of total survey error can be discussed individually, they constantly interact with one another in complex ways. For example, poorly trained interviewers are more likely to make mistakes with complex questionnaires, the race of the interviewer can interact with the race of respondents to create response effects, long, burdensome questionnaires are more likely to create fatigue among elderly respondents, and response scales using full rankings are harder to do over the phone than in person. In fact, no stage of a survey is really separate from the other stages, and most survey error results from, or is shaped by, interactions between the various components of a survey.

Conclusion

Conducting surveys often seems easy—and it is easy, if the job is done poorly. The pseudo-motto of one survey firm, "Price, Speed, Quality: Pick Two," makes the choice clear. Surveys are complex measurement tools that depend on a combination of many skills, including mathematics, creative writing, logistics, accounting, motivation, persuasion, and theorizing. Total survey error is both a means of keeping the survey researcher aware of the myriad ways in which error invades the data and a research agenda for identifying and minimizing those errors.

What is needed is what was called for in 1984 by the National Academy of Sciences' Panel on Survey Measurement of Subjective Phenomena, "a systematic long-term study of all phases of the survey process . . . [to] develop error profiles and models of total survey error." But at present, our knowledge of the error structure in general is limited, and the relative contribution of various error components in a particular survey is mostly incalculable. Today, the survey researcher's task and challenge is less that of an engineer (applying well-established formulas to well-measured physical parameters), and more that of an artisan (applying knowledge based on both science and experience to incompletely understood factors). The survey researcher draws on the existing scientific literature, and undertakes careful development work (e.g., pretesting) to make the best possible decisions about study design, sample size, question wording, interviewer training, and all the other components of surveys, without having complete information on the exact nature and relative magnitude of the error associated with each component of a survey. In finding the right balance, the survey researcher must avoid what Robert Groves called the "tyranny of the measurable." As Groves noted in *Survey Errors and Survey Costs*, "Errors that elude simple empirical estimation are often ignored in survey statistics practice. . . . The art of survey design consists of judging the importance of unmeasurable sources of error relative to the measured."

See Also the Following Articles

Experimenter Effects • Interviews • Non-Response Bias • Omitted Variable Bias

Further Reading

Andersen, R., Kasper, J., and Frankel, M. R. (1979). *Total Survey Error: Applications to Improve Health Surveys.* Jossey-Bass, San Francisco, CA.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (eds.) (1991). *Measurement Errors in Survey*. John Wiley & Sons, New York.

Binson, D., and Catania, J. A. (1998). Respondents' understanding of the words used in sexual behavior questions. *Public Opin. Q.* **62,** 190–208.

Brown, R. V. (1967). Evaluation of total survey error. *Statistician* **17,** 335–356.

Couper, M. P. (1997). Survey introductions and data quality. *Public Opin. Q.* **61,** 317–338.

de Leeuw, E. D., Hox, J. J., and Snijkers, G. (1995). The effects of computer-assisted interviewing on data quality: A review. *J. Market Res. Soc.* **37,** 325–343.

Deming, W. E. (1944). On errors in surveys. *Am. Sociol. Rev.* **9,** 359–369.

Duncan, O. D., Schuman, H., and Duncan, B. (1973). *Social Change in a Metropolitan Community.* Russell Sage, New York.

Groves, R. M. (1989). *Survey Errors and Survey Costs.* John Wiley & Sons, New York.

Kish, L. (1965). *Survey Sampling.* John Wiley and Sons, New York.

Lavrakas, P. (1993). *Telephone Survey Methodology.* Sage, Newbury Park, CA.

Lee, S.-Y. (1997). Understanding total survey error and statistical models according to the type of survey error. *Korean J. Sociol.* **31,** 223–257.

Lessler, J. (1984). Measurement error in surveys. In *Surveying Subjective Phenomena* (C. F. Turner and E. Martin, eds.), Vol. 2. pp. 405–440. Russell Sage, New York.

Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opin. Q.* **63,** 220–236.

Schuman, H., and Presser, S. (1981). *Questions and Answers.* Academic Press, New York.

Schwarz, N., Groves, R. M., and Schuman, H. (1995). *Survey Methods.* Survey Methodology Program Working Paper No. 30. Institute for Social Research, University of Michigan, Ann Arbor.

Smith, T. W. (1989). Random probes of GSS questions. *Int. J. Public Opin. Res.* **1,** 305–325.

Smith, T. W. (1994). An analysis of response patterns to the ten-point scalometer. In *American Statistical Association 1993, Proceedings of the Section on Survey Research Methods,"* ASA.

Smith, T. W. (1995). Holocaust denial: What the survey data reveal. *Working Papers on Contemporary Anti-Semitism."* American Jewish Committee, New York.

Tourangeau, R., and Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opin. Q.* **60,** 275–304.

Turner, C. F., and Martin, E. (eds.) (1984). *Surveying Subjective Phenomena*. Russell Sage, New York.

Tourism Statistics

Stephen L. J. Smith
University of Waterloo, Waterloo, Ontario, Canada

Antonio Massieu
World Tourism Organization, Madrid, Spain

Glossary

domestic tourism Tourism that occurs within the visitor's country of residence.
international tourism Tourism that occurs when a visitor travels to a country other than that of his/her usual residence.
tourism The set of activities engaged in by persons temporarily away from their usual environment for a period of not more than one year, and for a broad range of leisure, business, religious, health, and personal reasons, excluding the pursuit of remuneration from within the place visited or long-term change of residence.
tourism product A good or service that would be produced only in a substantially reduced volume or would virtually cease to be produced at all in the absence of tourism.
tourism industry An industry that produces a tourism product.
visitor A person temporarily away from his/her usual environment for a period of not more than one year whose primary purpose of travel is not the pursuit of remuneration from within the place(s) visited. The person may be engaged in any of the broad range of activities listed above in the definition for tourism. If the person does not stay away from his/her usual residence over night, the visitor is a "same-day visitor"; if he/she stays away overnight, the person is a "tourist."

The development of consistent measures of tourism has challenged statisticians and economists since the 1930s. The challenges arise, in part, from the nature of tourism as an economic activity. Although tourism is often referred to as an industry, it is fundamentally different than conventional industries and these differences complicate the measurement of tourism. Further, the development of measures of tourism consistent among nations has required extensive negotiations among national statistical agencies and international organizations to reach a consensus on definitions of tourism and related concepts. These concepts then had to be operationalized through new analytical tools. International agreement on core definitions and measurement techniques has now been achieved in principle. The tasks facing tourism statisticians are to apply and extend the concepts and tools that have been developed.

The Challenge of Measuring Tourism

The Journey to a Common Definition of Tourism

As many social phenomena, tourism can be defined in numerous ways. The definitions not only reflect the different uses to which their authors wish to put their definition, they also reflect fundamentally different attitudes to tourism. For 200 years, some social commentators have imbued "tourism" with an invidious connotation, as something unworthy of those with refined tastes. However, most contemporary social scientists work with less judgmental perspectives.

One of the more common perspectives can be characterized as "demand-side" because it focuses on tourism as a human experience. Pearce, for example, wrote "tourism may be thought of as the relationships and phenomena arising out of the journeys and temporary stays of people travelling primarily for leisure or recreational purposes." Leiper proposed a more expansive

demand-side definition: "[t]ourism comprises the ideas and opinions people hold that shape their decision about going on trips, about where to go (and where not to go) and what to do or not do, about how to relate to other tourists, locals, and service personnel. And it is all about behavioral manifestations of those ideas and opinions."

McIntosh, Goeldner, and Ritchie incorporated elements of the supply side (the businesses serving tourists) when they defined tourism as "the science, art, and business of attracting and transporting visitors, accommodating them, and graciously catering to their needs and want." Jafar proposed an even more intellectually comprehensive definition when he argued that "[t]ourism is the study of man away from his usual habitat, of the industry that responds to his needs, and of the impacts that both he and the industry have on the host's socio-cultural, economic, and physical environments."

Cohen side-stepped the task of defining tourism by defining a tourist: "Tourism is a fuzzy concept—the boundaries between the universe of tourist and non-tourist roles are vague . . . [a] 'tourist' is a voluntary, temporary traveller, travelling in the expectation of pleasure from the novelty and change experienced on a relatively long and non-recurrent round-trip."

These examples illustrate the divergence in academic approaches to defining tourism. Unfortunately, not only do these examples fail to represent a consensus on the nature of tourism, they do not provide a solid conceptual foundation on which to build reliable and accurate measurement tools. However, statisticians in national agencies and international organizations have also been debating the nature of tourism, particularly in the context of international measurement. Their interests arise from two related needs: (1) the need to accurately and consistently measure the magnitude of international tourism flows and (2) the need to analyze the structure and magnitude of tourism in national economies.

The work of these experts covers more than a half century, beginning in 1937 with the Committee of Statistical Experts of the League of Nations. The committee defined "international tourist" as anyone visiting a country other than his/her usual residence for more than 24 hours, with the exception of workers, migrants, commuters, students, and en route travellers (e.g., persons from one country passing through an airport in another country on their way to a third country). Little was done with this definition for the next two decades, however, as a result of the demise of the League of Nations and World War II.

After the war, international tourism grew rapidly, driven by a curiosity among many people to see more of the world, rising discretionary incomes, and improved modes of commercial transportation. In 1950, the International Union of Official Travel Organizations (IUOTO)—the precursor to the World Tourism Organization—expanded the 1937 definition by including students on tours. IUOTO also proposed a definition for "international excursionists" (an individual visiting a country other than his/her residence for pleasure on a same-day trip) and "transit travellers" (travellers en route through a second country on their way to a third). A few years later, in 1953, the UN Statistical Commission modified the IUOTO definition by specifying a maximum duration of six months for a tourism trip. A decade later, the 1963 UN Conference on International Travel and Tourism drew a distinction between "tourists" (those who stayed away for 24 hours or more) and "excursionists" or "day visitors" (for persons who stayed away from for less than 24 hours). The combination of "tourists" and "excursionist" were collectively called "visitors." In 1978, a conference hosted by the WTO, the U.N. Conference of Trade and Development, the Conference of European Statisticians, the East Caribbean Common Market, and the Caribbean Community ratified these basic definitions, although they called for the use of "excursionist" to cover both same-day visitors and "in-transit travellers."

The IUOTO definition, in this modified form, became an operational standard for over a decade. In practice, however, it did not adequately address other core concepts and had little impact on promoting harmonization of tourism statistics among nations. The next major step forward occurred in 1991 at WTO-sponsored conference in Ottawa, Canada, on tourism statistics. The Ottawa Conference achieved agreement on a number of important definitions and extended the maximum time for a trip to be considered as a tourism trip to one year. It also provided the foundation for harmonizing international tourism statistics and the development of an analytical framework to make the measurement of tourism consistent with those of other industries (the framework was to become known as the Tourism Satellite Account or TSA).

Another WTO conference on tourism statistics was held at Nice, France, in 1999 to report on progress in measuring the contribution of tourism to national economies. The delegates from 160 nations at that conference endorsed a number of key conventions proposed by the WTO related to TSAs. There were some technical differences, though, between the WTO recommendations and a set of conventions proposed by the OECD. The WTO and OECD resolved these differences in the months following the 1999 conference and, in 2000, the U.N. Statistical Commission approved a joint submission by the WTO, OECD, and Eurostat that represents a new international standard in tourism statistics.

Why Is Tourism Difficult to Measure?

Tourism is something that people do rather than something that businesses produce. This is an important

distinction when one tries to measure the magnitude of tourism as an industry because industries are defined by the products they make. For example, the motor vehicle industry is the set of businesses that produce motor vehicles; the corn industry is composed of farms that grow corn. In the case of tourism, though, there is no single characteristic product that could be classified as "the tourism product." Instead, tourism is associated with a variety of commodities, primarily services. Measurement of tourism necessarily involves measurement of the production and use of many fundamentally different commodities produced by many different types of businesses.

As noted, those commodities tend to be services, and the measurement of the production and consumption of services is difficult. Services are intangible; they cannot be stored and inventoried. The production of services typically is intrinsically tied to their consumption. For example, the provision of overnight accommodation in a hotel to a guest does not occur until a guest checks in and spend the night; and if that room-night is not sold, it is lost forever.

Not only does tourism cover a wide range of products, the businesses providing those products tend to be small or medium-sized enterprises that are widely dispersed throughout rural areas, small towns, and cities. Collectively, the tourist clientele of those businesses make billions of individual transactions—most fairly small—over the course of a year. Moreover, most of these businesses typically serve both visitors and nonvisitors. The development of a statistical infrastructure that can provide accurate and reliable coverage of this large, diffuse, and complex environment clearly is challenging.

Definition of Tourism and Related Concepts

Tourism and Its Forms

The history of efforts by national and international statistical agencies to reach agreement on a definition of tourism culminated in the World Tourism's Organization definition (endorsed by the U.N. Statistical Commission in 1993):

> *The set of activities engaged in by persons temporarily away from their usual environment for a period of not more than one year, and for a broad range of leisure, business, religious, health, and personal reasons, excluding the pursuit of remuneration from within the place visited or long-term change of residence.*

This represents the starting point for collecting, reporting, and analyzing tourism statistics. Tourism occurs in six different forms, three that can be considered fundamental and three that are combinations of the fundamental forms.

Domestic: A visit by a resident of a country totally within the border of that country.

Inbound: A visit by a resident of another country into a specified country.

Outbound: A visit by a resident of one country to another country.

Internal: The combination of domestic and inbound tourism.

National: The combination of domestic and outbound tourism.

International: The combination of inbound and outbound tourism.

Related Concepts

Two other concepts are core to measuring tourism: tourism product and tourism industry. The basic notion of product and industry are adapted from those used by Systems of National Accounts (SNA), and the U.N.'s International Standard Industrial Classification system and the Central Product Code. The reasons for this are to ensure the definitions and measures of tourism are consistent with those used by conventional industries. Although tourism is a demand-side concept—something that people do rather than something businesses produce—it still has to be conceptualized in a way that linkages can be made to the broader international systems of industrial statistics.

Tourism Product: A good or service that would be produced only in a substantially reduced volume or would virtually cease to be produced at all in the absence of tourism.

Tourism Industry: An industry that produces tourism products.

The application of these definitions is not as straightforward as they may appear. Tourism products are used by both visitors and nonvisitors. For example, not everyone who stays in a hotel, travels by airplane, rents a car, or dines in a restaurant is a visitor. Moreover, tourism products can be produced by nontourism industries. Some retail department stores or grocery stores also operate restaurants or dining counters. Certain tourism industries produce tourism services that are not their characteristic product: hotels provide food service operations and may sell tour packages. Airlines and rail companies provide food services.

Some tourism industries sell nontourism products. Hotels sell fax and long-distance telephone services to their guests as well as dry cleaning or laundry services. Restaurants may sell clothing and nonfood gift items. Some nontourism industries produce commodities routinely purchased by visitors: travel books, food and

beverages from retail stores, sunscreen lotion, crafts, and clothing.

In brief, tourism industries produce both tourism and nontourism products. Nontourism industries produce tourism and nontourism products. Visitors consume nontourism products and nonvisitors consume tourism products. The measurement of tourism requires data sources and analytical tools that allow analysts to separate out the portions of the tourism-related production and consumption of tourism products and nontourism products from the nontourism activity. The analytical framework by which this happens is known as the Tourism Satellite Account.

Tourism Satellite Accounts (TSA)

Objectives and Logic of TSAs

A Satellite Account is a term developed by the United Nations to refer to an extension of the SNA (hence, a "satellite" of the SNA) to measure the size of economic sectors that are not defined as industries in national accounts. Tourism, for example, is an amalgam of industries such as transportation, accommodation, food and beverage services, recreation and entertainment, and travel agencies.

Tourism is a unique phenomenon because it is defined in terms of a certain type of consumer, a "visitor." Visitors buy goods and services, both tourism and nontourism. The key from a measurement standpoint is to associate visitors' purchases to the total supply of these goods and services within a country.

However, visitor consumption is not restricted to a set of predefined goods and services produced by a predefined set of industries. What makes tourism special is not what is purchased but the temporary situation in which the consumer finds himself/herself. He/she is outside his/her usual environment, traveling for a purpose other than the exercise of an activity remunerated from within the place(s) visited. This is the characteristic that distinguishes visitors from other consumers.

The TSA is a new statistical instrument that brings together these diverse aspects of tourism by providing a tourism dimension to the framework of an SNA. It permits the separate measurement of the demand and supply sides of tourism within an integrated system (the SNA) that describes the production and demand aspects of the whole economy.

Structure of TSAs

Tables

The methodological design for the elaboration of the TSA is a set of definitions and classifications integrated into tables and organized in a logical, consistent way. It allows the examination of the whole economic magnitude of tourism in both its aspects of demand and supply. In principle, there are 10 key tables that constitute a TSA:

Tables 1, 2, and 3: Visitor final consumption expenditure in cash, by product, and form (as described above) of tourism.

Table 4: Internal tourism consumption, by product and form of tourism.

Table 5: Production accounts of tourism industries and other industries.

Table 6: Domestic supply and internal tourism consumption, by product.

Table 7: Employment in tourism industries.

Table 8: Gross fixed capital formation of tourism industries and other industries.

Table 9: Tourism collective consumption, by functions and level of government.

Table 10: Nonmonetary indicators such as employment and numbers of visitors.

Aggregates

TSAs support the measurement of so-called "aggregates" or macroeconomic indicators such as tourism consumption, tourism value-added, tourism GDP, tourism employment, and tourism gross fixed capital formation. These aggregates are not the most important feature of the TSA, whose primary objective is to provide detailed and analytical information on all aspects of tourism, particularly the composition of visitor consumption, the productive activities most concerned by the activities of visitors, and relationships between tourism and other productive activities. Nevertheless, these aggregates are often politically important because they are measures of the quantitative importance of tourism in a country. The importance of credible measures of the magnitude of tourism in a national economy should not be underestimated.

Uses of TSA

The development of TSAs has been fueled by the recognition that they can:

- Improve knowledge of tourism's importance relative to overall economic activity in a given country;
- Provide an instrument for designing more efficient policies relating to tourism and its potential for job creation; and
- Create awareness among the players directly and indirectly involved with tourism of the economic importance of this activity and, by extension, the role of tourism in all industries producing the myriad goods and services demanded by visitors.

These needs are met by the TSA insofar as it provides (for an entire country and over a period of one year) an articulated framework of economic information that serves the interests of both political decision-makers and entrepreneurial decision-makers. Australia, Canada, Dominican Republic, Ecuador, Mexico, Spain, Sweden, and United States are countries with an established TSA.

Types and Sources of Tourism Statistics

Governments, entrepreneurs, and analysts usually do not have adequate information for designing sound tourism policies and business strategies, and for evaluating their effectiveness and efficiency. This chronic shortage of information on the role of tourism in national economies worldwide is partly due to the "horizontal" nature of tourism—the fact that tourism is an aspect of many different industries. The challenge of compiling and disseminating information from so many different types of businesses requires many different quantitative and qualitative sources.

The demand for tourism statistics continues to grow worldwide. There is also increasing demand for reliability, accuracy, precision, and timeliness in tourism statistics. Some actual or potential statistical sources are tourism-driven, such as tourism expenditure surveys, arrival and departure counts at national borders, records of overnight stays, and occupancy rates in accommodation enterprises. However, the full measurement of tourism economic impacts and the development of TSAs requires data from general economic statistical sources as well as administrative sources such as air traffic regulation authorities and tax records.

Internationally comparable tourism data should be based on certain common key indicators related to the different forms of tourism and certain tourism industries.

Inbound Tourism

Arrivals are a basic measure and are not necessarily equal to the number of different persons traveling. When a person visits the same country several times a year, each visit by the same person is counted as a separate arrival. If a person visits several countries during the course of a single trip, his/her arrival in each country is recorded separately. Arrivals associated with inbound tourism equals arrivals by international visitors to the economic territory of the country of reference and include both tourists and same-day nonresident visitors.

Data on arrivals may be obtained from different sources. In some cases, data are obtained from border statistics derived from administrative records (police, immigration, traffic counts, and other type of controls) as well as border surveys. In other cases, data are obtained from registrations at tourism accommodation establishments.

Statistics on *overnight stays* refer to the number of nights spent by nonresident tourists in hotels and similar establishments, or in all types of tourism accommodation establishments. If one person travels to a country and spends five nights there, that makes five tourist overnight stays (or person-nights).

Average length of stay refers to the average number of nights spent by tourists (overnight visitors) in all types of tourism accommodation establishments.

Tourism expenditure data are obtained from the item "travel receipts" of the Balance of Payments (BOP) of each country and corresponds to the expenditure of nonresident visitors (tourists and same-day visitors) within the economic activity of the country of reference. BOP data typically are collected either by a central bank or a nation's official statistical agency.

Domestic Tourism

Domestic *overnight stays* is the number of nights by resident tourists in hotels and similar establishments, or in all types of tourism accommodation establishments, and may be obtained either by household surveys or from records of accommodation establishments. *Same-day visits* typically are estimated from household surveys.

Outbound Tourism

Departures associated with outbound tourism correspond to the departures of resident tourists outside the economic territory of the country of reference.

Tourism expenditure data in other countries are obtained from the item "travel expenditure" of the BOP of a country and correspond to the "expenditure of resident visitors (tourists and same-day visitors)" outside the economic territory of the country of reference. Alternative sources, in principle, include border surveys and currency control forms. However, the current level of practice with respect to these sources does not support international comparison of data; only BOP data are comparable.

Tourism Industries

One of the most basic statistics related to tourism industries is the *number of establishments* within each tourism industry. This information may be compiled from business registration or tax data. Many different characteristics of businesses in each tourism industry are possible, but some of the most common are those related to the

accommodation industry. The *number of rooms and bedplaces* data refers to the capacity in hotels and similar establishments for providing temporary accommodation to visitors.

Occupancy rate refers to the relationship between available capacity and the extent to which it is used. This rate may refer either to use of rooms or of beds. Occupancy rate is based on the number of overnight stays of both resident and nonresident tourists.

Employment estimates are difficult to develop because of the high percentage of part-time or seasonal jobs in tourism industries. Several different measures can be developed. "Full-time equivalent jobs" (FTEs) are a statistical compilation of part-time and temporary jobs as well as full-time and permanent jobs into a statistical estimate of full-time equivalent jobs. Thus, FTE employment numbers are lower than the number of persons actually employed in tourism. *Jobs* or *positions* refer to specific positions within an employer, which often are associated with a particular job description. A job/position may be held by more than one person over the course of a year as a result of employee turnover. *Employees* refers to the number of different individuals that are employed in a tourism industry over a year, regardless of the number of hours per week or weeks per year worked. Employment data may be obtained from surveys or censuses of enterprises or statistically estimated from payroll data and average pay rates by occupation.

See Also the Following Articles

Geography • Social Economics • Transportation Research

Further Reading

Smith, S. L. J. (1998). Tourism as an industry: Debates and concepts. In *The Economic Geography of Tourism* (D. Ioannides and K. Debbage, eds.), pp. 31–52. Routledge, London.

Smith, S. L. J. (2000). New developments in measuring tourism as an area of economic activity. In *Trends in Outdoor Recreation and Tourism* (W. Gartner and D. Lime, eds.), pp. 225–234. CAB International, London.

United Nations and World Tourism Organization (1994). *Recommendations on Tourism Statistics.* United Nations, New York.

United Nations, World Tourism Organization, Organisation for Economic Co-operation Development, and Commission of the European Communities (Eurostat) (2001). *Tourism Satellite Account: Recommended Methodological Framework.* United Nations, New York.

World Tourism Organization (2001). *Basic References on Tourism Statistics*. (www.world-tourism.org/statistics/tsa_project/basic_references/index-en.htm).

World Tourism Organization (2002). *TSA in Depth: Analysing Tourism as an Economic Activity*. (www.world-tourism.org/statistics/tsa_project/TSA_in_depth/index.htm).

World Tourism Organization and International Air Transport Association (2002). *General Guidelines for Using Data on International Air-Passenger Traffic for Tourism Analysis.* World Tourism Organization, Madrid.

Transportation Research

Ryuichi Kitamura
Kyoto University, Kyoto, Japan

Glossary

centroid An area representing a zone that is treated as a node connected to transportation networks. All trips originating from a zone are assumed to start from its centroid and end at the centroid of the destination zone.
destination The terminal end of a trip.
origin The beginning end of a trip.
origin-to-destination trip table A table that shows the frequency of trips made in the area, typically per day, between each pair of origin and destination zones.
trip The movement made to engage in activities at another location in urban area. An individual's day may contain several trips, e.g., a morning commute trip from home to work, a trip from the office to a nearby restaurant and a trip back to the office during the lunch break, a trip to return home after work, an evening trip to a movie theater, and a trip back home.
trip chain A closed sequence formed by the series of trips made by an individual that starts and ends at the same place.

Transportation research is concerned with the movement of both people and goods. This article, however, focuses on the measurement of people's movements in time and space, primarily within an urban area. First, the history of travel surveys is briefly summarized. Early surveys of car trips, traditional large-scale household surveys, and enhancements recently made to them are discussed next, followed by sections concerned with reporting errors in travel surveys, application of time-use survey methods that is believed to reduce trip reporting error, and multi-day surveys. The article concludes with a discussion on the application of information technology to travel surveys.

Measuring Travel Patterns

Survey methods to measure the characteristics of trips made by individuals have evolved in several stages. Surveys in the first stage were concerned only with trips made by the automobile. The unit of sampling was the trip, and no consideration was given to the fact that an individual often makes several trips in a day and their attributes are interrelated. In the second stage, large-scale household travel surveys were conducted to collect information on all trips made by household members (typically age 5 and older) on the survey day. In the third stage, various improvements were made to survey instruments and administration methods to reduce non-response, under-reporting of trips, and other reporting errors in household travel surveys. Also initiated in this stage were extension of the survey period to multiple days, application of methodologies in time-use surveys, and the use of computer-aided telephone interviews (CATI) to retrieve survey responses. Currently, the development of transportation survey methods may be viewed as in its fourth stage, in which global positioning systems (GPS), mobile communications systems, and other new technologies are being applied to obtain more accurate and complete information of individuals' trajectories in time and space.

Many different types of surveys exist. For example, traffic detectors permanently installed on roadways continuously provide traffic counts. Or a traffic counter with a rubber tube may be placed on the roadside to obtain traffic counts data at different locations on a rotational basis. These may be viewed as forms of surveys. Another form of survey is the transit on-board survey in which transit passengers are sampled. Certain techniques are used to obtain data on the use of parking facilities or to obtain the distribution of vehicle speeds.

Roadside Origin-Destination Studies

Passenger travel demand forecasting methods in their early forms can be found prior to World War II. In a 1927 study in Cleveland, for example, linear extrapolation was used to produce forecasts, and a 1926 Boston study adopted a gravity model that depicted the flow of people between a pair of geographical zones analogous to Newton's law of gravity.

Because motorization created a vast need for road building, early studies focused on auto trips. Roadside interviews were used to obtain information on origin-destination, purpose, and vehicle occupancy (the number of individuals in the vehicle). Most typically, vehicles were stopped by the police at a survey station and questions were asked to obtain this information. Alternatively, a questionnaire to be completed and mailed back was handed out to the driver. In yet another method, the license plate numbers of vehicles passing a survey station were recorded, and a questionnaire was mailed to the registered owner of each vehicle identified.

The information used for transportation planning in these early days primarily came from the origin-to-destination trip table developed from the data produced by such surveys. Later, however, the need to study the factors that influenced travel behavior was recognized, which led to the development of large-scale household travel surveys and full-fledged travel demand forecasting procedures after World War II.

Household Travel Surveys

Individuals often have means of travel other than the automobile, and planners may aim to develop a transportation system in which the automobile and public transit are well balanced. This calls for the knowledge of each individual's travel, regardless of the mode used.

The U.S. Bureau of Public Roads (BPR) published a manual for home interview travel surveys in 1944, and household-based origin-destination travel surveys were conducted in seven urban areas. Also in the 1940s, the approach of developing transportation plans based on the relationships among transportation, socio-demographics, and land use was recognized. In the San Juan study initiated in 1948, for example, models were developed to estimate the number of trips generated in an area based on land use information. Following this, major home interview surveys were conducted in Detroit and Chicago in 1955 and 1956, respectively.

In these surveys, households were sampled on the basis of residence, and were approached without advance notice. Random sampling or cluster sampling was in general adopted. In the interview, information on the trips of the day before was collected from all household members at least five years old. High sampling rates (approximately 5%) were adopted in the early studies. Information on the household and its members as well as the attributes of trips made by the household members on the survey day was collected in these travel surveys. Typical member attributes on which information was collected included age, sex, employment, and driver's license holding. Household characteristics of interest were household size (the number of household members living together), housing type and ownership, the number of vehicles owned or available, and household income.

In planning analysis, the data from a travel survey are often supplemented with data on land use and transportation networks. The former comprise population, population density, the number of employees by industry, the number of housing units by type, average household size, average number of vehicles per household, median income, and other indicators of the characteristics of respective geographical zones. Transportation network data typically comprise zone-to-zone (or centroid-to-centroid) travel times obtained using network models for all relevant pairs of zones, often evaluated for different periods (e.g., morning peak, off-peak, afternoon peak, and evening). For public transit, network data contain attributes of transit trips, e.g., transit fare, waiting time, and number of transfers.

Because transportation planning at the time was primarily concerned with road building, travel surveys were primarily concerned with motorized trips, i.e., those made by automobile, taxi, bus, or rail. For example, the 1955 and 1965 household travel surveys in the Detroit metropolitan area excluded walk and bicycle trips; it was only in the 1980 survey that these non-motorized trips were addressed.

It is important to note that the treatment of absent households can result in systematic bias if they are not properly handled. Although non-response due to absence at the time of contact, either by visiting or by placing a telephone call, may not introduce any systematic bias in general, this is not the case with a travel survey. That household members are absent implies that they are making trips, which is the subject of the survey. If absent households were not included in the sample, that would imply a tendency of excluding households with higher propensities to make trips. It is therefore critically important that repeated call backs are made to ensure that the absent households are included in the sample.

Enhanced Household Travel Surveys

Although home interview surveys are a highly desirable means of survey administration, they are costly and are

becoming increasingly difficult to conduct, as urban residents have become less cooperative in recent years. Other problems being recognized are trip underreporting and response inaccuracies. Since approximately 1980, new types of survey instruments and alternative methods of survey administration have been experimented with. Notable is the use of CATI; adoption of travel diaries, memory joggers, and other instruments for improved quality of trip reporting; adoption of multi-day survey periods; and application of time-use survey methods. Around the same time, the sample size had decreased to 2000–10,000 households, with an average in the beginning of 1990s of about 2500 households.

The method of contact had also changed. Instead of the retrospective surveys with cold contact as used in the earlier days, prospective respondents were first recruited by the telephone. The typical procedure can be described as follows:

1. Recruiting by the telephone or mail.
2. Notification of the survey date; delivery of a letter describing the purpose of the survey, travel diaries, memory joggers, and other survey instruments.
3. Retrieval of responses through CATI.

The questionnaires used in earlier self-administered travel surveys were similar to the worksheets used by interviewers to record responses in home interviews. More recent questionnaires are more "respondent friendly," often taking on the form of a pocket-size diary.

Despite these efforts, several problems persist. Although there are reports of improved trip reporting due to the improved questionnaire design, underreporting of trips is far from being eliminated. The effectiveness of travel diaries or memory joggers is not conclusive. Moreover, CATI presents its own problems with household travel surveys, because retrieving trip information from all household members requires quite lengthy interviews for larger households, often requiring several telephone interviews. This leads to systematic non-response when larger households are underrepresented in the sample. Furthermore, because it is not always possible to interview every household member, especially for a large household, travel information tends to be offered through proxies, compromising its quality. Consequently, responses tend to be less accurate, and chances of trip underreporting tend to increase for larger households or for household members who tend not to be at home. The effectiveness of alternative administration methods, including mail surveys (mail-out mail-back, or drop-off mail-back), is being evaluated.

Another problem is the difficulty when using random digit dialing in determining whether a telephone number is eligible when there is no answer. Determining whether a number is eligible is crucial to avoid systematic non-response biases, but it increases the cost of survey. In addition, telephone-based travel surveys are subject to general problems, including the presence of households without a telephone, call screening by an answering machine, and the more recent practice of screening calls using the caller ID.

On the other hand, one of the advantages of telephone-based surveys is easy call backs, which aid in reducing systematic non-response biases. Another advantage is that CATI facilitates the dialogue between the respondent and the interviewer. This is important in travel surveys because the definition of trip is quite involved and the definition of trip purposes can be confusing; a well-trained interviewer can aid the respondent in providing travel information and thus contribute to its quality. CATI is also effective when the survey questionnaire involves complex branching or customization. The latter is often adopted in stated-preferences surveys where the respondent is asked to indicate his or her preference under a hypothetical scenario. Responses to earlier questions in an interview may be used to customize the scenario and make it seem more realistic for the respondent.

Trip Reporting Errors

It has been found that short trips, walk and bicycle trips, trips that neither start from nor end at the home base (called non-home-based trips), trips made for work-related businesses, and trips back to home (home trips) tend to be underreported with the survey instruments and administration methods that have been adopted in household travel surveys. On the other hand, commute trips to work (work trips) and trips made by motorized modes tend to be reported well.

A comparison of the entries in an activity diary and the answers to a travel survey questionnaire has indicated that short movements and brief excursions out of the home tend not to be recognized as trips. In addition, the definition of the trip is difficult to convey, and the respondent may not clearly understand which trips should be reported in the survey.

For example, consider a case in which a commuter does grocery shopping at a supermarket on the way back home from work. In this case, the journey from work to home is composed of two trips: a shopping trip from the workplace to the supermarket, and a trip from the supermarket to home. Consider, next, a case in which the commuter picks up a newspaper from a vending machine located at a shopping center. It is unlikely in this case that the respondent will break up the movement from work to home into two segments and report them as two trips. But should the movement be reported as two trips? Unfortunately, the answer is not obvious.

How the respondent is expected to report such movements as jogging, walking the dog, or pleasure driving is also ambiguous. In these movements, there are no clear destinations, making them difficult to report in the survey. Again, how these movements should be reported is often not well spelled out in the survey. Finally, and to make the matter more confusing, movements within the same premise, e.g., a shopping mall or a college campus, are not considered as trips. In this case, the respondent is not supposed to report movements as trips at all.

Given a reported trip, the next question is how accurately its attributes are reported. The most noticeable reporting errors are associated with departure and arrival times, which tend to be rounded to the nearest quarters of an hour, i.e., 00, 15, 30, or 45 minutes. For example, in a nationwide survey conducted in the United States in 1990, 36.2% of trips were reported to have started at exact hours, 27.8% at 30 minutes, 9.2% at 15 minutes, and 8.7% at 15 minutes past hours. Only 18.1% of the trips were reported to have started at other minutes. The duration of a trip tends to be rounded in a similar manner. On the other hand, a study in which the movement of vehicles was recorded using GPS devices shows that trip starting times are distributed uniformly from 00 to 59 minutes.

Another attribute of the trip that is difficult to report is the destination location. The respondent does not always know his destination by its street address. Rather, it may be reported as "McDonald on Anderson Street." When such a description does not offer sufficient information to geo-code the location, the respondent must be contacted to retrieve adequate levels of information.

Application of Time-Use Survey Methods

Time-use surveys have been conducted since the early 20th century to collect information on individuals' activities. A standard scheme for activity classification has been developed, and a large-scale international study of time use was carried out in the late 1960s. The first time-use study in the transportation planning area dates back to the mid-1940s, but it was in the 1990s that time-use survey methods were adopted in travel surveys with the purposes of improving the accuracy of trip reporting and obtaining data that may be used to probe into the mechanism of trip making.

Travel surveys that adopted time-use survey methods differ from time-use surveys themselves in that information on in-home activities is not always collected exhaustively. For example, in one travel survey, the respondent was requested to report only those in-home activities that exceeded 30 minutes in duration. Another travel survey, on the other hand, asked the respondent to report those in-home activities that substitute for out-of-home activities (e.g., watching television at home instead of going to a movie theater).

The anticipation that trip underreporting can be reduced by applying time-use survey methods is based on the belief that activities are easier to recall than trips. Namely, "What did you do next?" is easier to answer accurately than "Where did you go next?" because activities are continuous in time and are therefore easier to trace back and recall, while trips are intermittent and difficult to recall exhaustively.

In one example that illustrates the extent of trip underreporting, results of two time-use surveys, conducted in the Netherlands and California, were tabulated. The two surveys adopted different schemes of data collection. The survey in the Netherlands was based on a time-interval method in which the representative activity was asked for each of 144 15-minute intervals of the day. The survey in California was based on activity episodes, with the respondent prompted to report activity by activity as they were pursued. Despite the difference in their data collection methods, these two surveys offered consistent time-use figures for such basic activities as sleeping and having meals. The episode-based California data contained more reported trips (3.046) than the interval-based Dutch data (2.484). The average number of unreported trips, estimated based on activity location codes, was 2.130 for California and 2.455 for the Netherlands, with the total number of trips estimated as 5.176 and 4.939, respectively. It is evident that trips tend to be underreported, and that a substantial number of unreported trips can be captured using time-use data.

Multi-Day Surveys and Other Applications

Some trips, like commute trips, are repeated regularly day to day; others are not. People do not repeat exactly the same travel pattern every day. Knowing daily variations in an individual's travel patterns is important in some contexts. For example, the statement that "20% of commuters use public transport" may imply that every commuter has a 20% chance of using public transit on a given day, or that a fixed 20% of the commuter population uses public transit every day. One-day data do not offer enough information to determine which is the case. But this question is important if one wishes to determine who are the beneficiaries of subsidies for public transit, or to determine the target of effective marketing to promote public transit use.

Multi-day surveys are needed to determine how variable are day-to-day travel patterns. Likewise, multi-week surveys are needed to evaluate the variability in travel

from week to week, and probably panel surveys to assess seasonal variations. Most travel surveys are concerned with just one day, which is often taken as a "typical" weekday (Tuesday, Wednesday, and Thursday) in an autumn month. There have been only several multi-day surveys for survey durations of up to one week, and a few surveys with survey durations of several weeks, reported in the literature on travel behavior analysis. The primary reason for the rarity of multi-day surveys is presumably the increased respondent burden, and the decline in response accuracy as a consequence. For example, the number of trips reported by the respondent gradually declines toward the end of the survey period in a weekly survey, then jumped up on the very last day of the survey.

Day-to-day variations in daily travel create problems when one wishes to measure change in travel patterns over time, e.g., in order to evaluate the effect on travel of a planning measure such as the opening of a bypass. Day-to-day variations in travel patterns make this comparison less reliable. Higher levels of precision may be achieved by increasing the length of the survey period in the before and after periods, by increasing the sample size, or both. Little research has been done on this subject in the travel behavior research field.

There are cases in which random sampling is impractical or inefficient as a means of collecting data on travel behavior. For example, suppose one wishes to study how people choose between private automobiles and public transit for commuting, but the fraction of commuters who use public transit is very small. In such a case, random sampling of households would yield only a small number of transit users, increasing the total number of samples required to achieve a desired level of accuracy. An alternative is to over-sample transit users by adopting non-random sampling schemes, e.g., by distributing questionnaires to the riders at railway stations or in buses or trains. This is called choice-based sampling because sampling is done based on the choices made by the sample individuals (i.e., to ride a bus or a train). The resulting choice-based sample contains obvious bias, which can be corrected by applying appropriately defined weights for parameter estimation.

A situation often encountered in transportation planning is the need to estimate demand for non-existent transportation services, e.g., a proposed subway line. In this case, data showing the use of the subway are not available simply because it does not yet exist. In such a case, a survey may be conducted that solicits the respondent to answer hypothetical questions that assume the presence of the subway line. For example, the respondent may be asked "Suppose you must visit the city hall, and there are the following two ways of traveling: (1) walk 8 minutes to the subway station, ride the subway for 15 minutes, then walk 5 minutes to the city hall, with a subway fare of $1.20, or (2) drive your car for 20 minutes, and park at a parking lot next to the city hall, and pay $3.50 for parking. Which one would you prefer?" Such a survey is called a stated-preference survey because the data it produces reflect "stated," as opposed to "revealed," preferences. The resulting data are most often used to develop mathematical models to explain the choices, which are in turn used to estimate demand for non-existent transportation services.

Information Technologies in Travel Surveys

Advances in information and computer technologies (ICT) have made it possible to trace the trajectory of a person or a vehicle in time and space with levels of accuracy that were unthinkable in the past. Much of the information associated with trips that has been collected through household travel surveys can now be obtained automatically using new technologies. In one experiment, a GPS device was installed on a passenger vehicle, and the position information was transmitted from the GPS unit to the experiment headquarters every second. In addition, supplementary information (driver, accompanying travelers, and the purpose of travel) was entered into a hand-held computer for each trip made using the vehicle. The results of the survey were compared with data obtained from conventional household travel surveys. Likewise, study results to determine the location of the holder of a cellular phone based on the intensities of its radio waves received at three or more communications stations have been accumulated.

Studies have shown that GPS devices, cellular phones, hand-held computers, and other ICT devices facilitate acquisition of accurate information on time and location. Attributes of trips such as beginning and ending time, origin and destination locations, and routes can then be inferred with high levels of accuracy that traditional household travel surveys cannot possibly attain. Combined with supplementary interviews of the travelers, most, if not all, information that has been provided by conventional surveys can be obtained with the new types of surveys that deploy ICT devises. Although there are technical problems that need be resolved (e.g., disruption of radio waves by high-rise buildings), it is likely that new forms of travel surveys that do not rely on respondents' retrospective reporting of trips will be prevalent in the near future.

See Also the Following Articles

Election Polls • Surveys • Urban Economics • Urban Studies

Further Reading

Cambridge Systematics, Inc. (1996). *Scan of Recent Travels Surveys*. DOT-T-37-08, Final Report prepared for the U.S. Department of Transportation and U.S. Environmental Projection Agency, Technology Sharing Program. U.S. Department of Transportation, Washington, D.C.

Louviere, J. J., Hensher, D. A., and Swait, J. D. (2000). *Stated Choice Methods: Analysis and Application.* Cambridge University Press, Cambridge, UK.

Ortuzar, J., de D., and Willumsen, L. G. (1994). *Modelling Transport.* 2nd Ed. Wiley, Chichester, UK.

Pas, E. I., and Harvey, A. S. (1997). Time use research and travel demand analysis and modeling. In *Understanding Travel Behaviour in an Era of Change* (P. Stopher and M. Lee-Gosselin, eds.), pp. 315–338. Pergamon, Oxford, UK.

Pendyala, R. M., and Pas, E. I. (2000). Multi-day and multi-period data for travel demand analysis and modeling. In *Transport Surveys: Raising the Standard, the Proceedings of an International Conference on Transport Survey Quality and Innovation, May 24–30, 1997, Grainau, Germany*. Transportation Research Circular, No. E-C008, August, Transportation Research Board, National Research Council, Washington, D.C.

Weiner, E. (1997). *Urban Transportation Planning in the United States: An Historical Overview*, 5th Ed., DOT-T-97–24. Technology Sharing Program, U.S. Department of Transportation, Washington, D.C.

Treatment Effects

Charles S. Reichardt
University of Denver, Denver, Colorado, USA

Glossary

before–after quasi-experiment A comparison of outcomes before and after a treatment is introduced.
between-participant design A comparison of participants who receive different treatments.
correlational design A between-participant quasi-experiment in which the treatment is a continuous rather than discrete variable.
interrupted time series quasi-experiment A comparison of a time series of observations before a treatment is introduced with the continuation of that time series of observations after the treatment is introduced.
nonequivalent group quasi-experiment A comparison of participants who receive different treatments in which the participants are not assigned to the treatments at random.
quasi-experiment A comparison in which different treatments are not assigned at random.
randomized experiment A comparison in which different treatments are assigned at random.
regression-discontinuity quasi-experiment A comparison in which participants are assigned to treatments using a cutoff score on a quantitative assignment variable.
threat to internal validity An alternative to the treatment as an explanation for an observed outcome difference.
treatment effect The difference between what happens after a treatment is administered and what would have happened if the treatment had not been administered, but everything else had been the same.
within-participant design A comparison in which each participant receives all the different treatments.

A treatment effect is the difference between what would have happened if a treatment had been implemented and what would have happened if the treatment had not been implemented, but everything else had been the same. Such a comparison is called the ideal comparison. Unfortunately, the ideal comparison cannot be obtained in practice. In any comparison that can be obtained in practice, every thing else cannot have been the same. Whatever else is not the same is called a threat to validity. A threat to validity can cause a difference in the observed outcomes and therefore either masquerade as a treatment effect when none is present or bias the estimate of a treatment effect when it is present. One of the critical tasks in estimating treatment effects is to take account of threats to validity. This article describes different types of comparisons, the primary threats to validity for each comparison, and the categories of threats to validity that must be addressed when estimating treatment effects.

Types of Comparisons

A treatment is an intervention such as a medical treatment, job training program, or remedial reading course. Comparisons that are drawn to estimate treatment effects involve two or more treatment conditions. To simplify the discussion, the simplest case is considered first in which there are only two treatment conditions. A treatment is implemented in one of the two conditions (the treatment condition) and either no treatment or an alternative treatment is implemented in the other condition (the comparison condition). For example, a novel medical treatment could be compared either to the absence of medical treatment or to a standard medical treatment. Also for convenience, the participants in a study are referred to as individuals and assumed to be humans, although the participants could be other animals and either individuals or groups of individuals.

It is customary to distinguish between two broad categories of comparisons (or research designs): randomized experiments and quasi-experiments. In randomized experiments, the treatment conditions are assigned at

random. Quasi-experimental comparisons are implemented without the benefit of random assignment.

It is also customary to distinguish between comparisons that are drawn "between" participants and those drawn "within" participants. In between-participant designs, some participants receive the treatment condition, whereas the other participants receive the comparison condition. The treatment effect is estimated by comparing the performances of these two groups of participants on an outcome variable. In within-participant designs, each participant receives both the treatment and the comparison conditions. The treatment effect is estimated by comparing each participant's performance in the treatment condition to that same participant's performance in the comparison condition. In both types of comparisons, the difference between the treatment conditions is called the independent variable. The outcome variable is called the dependent variable.

Between-participant comparisons can be either randomized experiments or quasi-experiments. The same is true for within-participant comparisons.

Between-Participant Randomized Experiments

In a between-participant randomized experiment, individuals are randomly assigned to treatment conditions. After the different treatments have been administered to the groups, a posttreatment, outcome measure is assessed. Effects of the treatment are estimated based on differences between the groups on the outcome measures. For example, in a series of classic studies of conformity by Solomon Asch in the 1950s, individuals were randomly assigned either to a treatment condition in which peer pressure was exerted or to a comparison condition without peer pressure. Peer pressure caused the treatment group to make far more errors on a subsequent perceptual task than the comparison group.

One of the primary threats to validity in a randomized between-participant experiment is "random selection differences." Because the treatment effect is estimated by comparing the performances of different individuals in the two treatment conditions, observed differences in outcomes could be due to differences in the composition of the two groups (i.e., random selection differences) as well as to the effects of the treatment. Statistical inference is the classic means of distinguishing between these two sources of effects. A statistical significance test reveals whether the observed outcome difference between the groups is larger than would be expected due to random selection differences (or to other random effects) alone. Alternatively, a confidence interval estimates the size of the treatment effect within a range of scores that takes account of uncertainty due to the effects of random differences.

Threats to validity due to nonrandom differences can also arise in between-group randomized experiments. Two of the most common such threats are local history and differential attrition. Local history arises when external events differentially affect the two treatment groups. Differential attrition means that different types of individuals in the two treatment groups fail to complete either the prescribed treatment protocols or the outcome measurements. Randomized experiments conducted under laboratory conditions can often minimize local history through careful implementation of controls such as experimental isolation from outside forces, and differential attrition can often be minimized by using short time intervals between the treatment and outcome measurement. Randomized experiments in the field should try to implement the same controls but often cannot avoid these problems even then.

Randomized experiments do not require that pretreatment measures be collected on the participants. However, pretreatment measures can help diagnose and take account of differential attrition. Pretreatment measures can also be used to increase the precision of treatment effect estimates and to assess how the effect of the treatment varies across individuals.

Within-Participant Randomized Experiments

In a within-participant randomized experiment, each participant receives all of the different treatment conditions. The treatment effect is assessed by comparing the performance of each individual under each of the different treatment conditions. For example, a series of studies by Benton Underwood and associates in the 1960s assessed the effects of massed versus distributed practice on learning by having participants study a list of words. Target words were repeated three times in the list and were randomly assigned to presentation in either massed fashion (i.e., the repetitions of a target word appeared one right after the other in the list) or distributed fashion (i.e., filler words appeared between repetitions of a target word). Following the presentation of the list, participants were asked to recall the words. The differences in the numbers of massed and distributed words that were recalled by each participant were averaged across the participants. The mean difference between the number of massed and distributed words that were recalled reflects the effect of massed versus distributed practice.

A threat to validity in within-participant studies arises when different materials (e.g., different words) are used in the different treatment conditions (e.g., the massed versus

distributed practice conditions), which could cause differences in outcomes (e.g., differences in the number of massed and distributed words that are recalled). In within-participant randomized experiments, the materials are assigned to the different treatment conditions at random. Therefore, a threat to validity due to using different materials in different treatment conditions can be addressed by the classic statistical procedures of confidence intervals and statistical significance tests.

Another threat to validity, called practice effects, can arise because of the ordering in which participants are exposed to the different treatment conditions (e.g., the ordering of massed and distributed words within a list). Participants can become more skilled as they perform a task, or they can become tired or bored as a task continues. To the extent that the materials for one treatment condition come earlier in time than the materials for the other condition, practice effects can bias the estimate of the treatment effect.

Practice effects can be controlled by counterbalancing the order of presentation. If the different treatments are repeated only a few times for each individual, the ordering of the treatments can be varied randomly across individuals. If the different treatments are repeated a large number of times for each individual, the order of the repetitions can be randomized for each individual. Another alternative is to order the treatment (A) and comparison (B) conditions in repeating ABBA sequences over time.

A final threat to validity is called transfer or carryover effects, which arise when the effect of a treatment condition depends on the condition that preceded it. For example, if the treatments being compared are different medications, the effect of the medication administered second could vary depending on the medication administered first. The effects of the first medication might not have worn off before the second medication is taken, so a drug interaction occurs. Also, the first medication might cure the illness, so the second medication can have no ameliorative effect.

When transfer effects are plausible, a between-participant design is usually preferred over a within-participant design. Otherwise, within-participant designs tend to be preferable because (i) participants serve as their own comparison condition, which increases statistical precision, and (ii) within-participant designs tend to be less expensive to implement.

Between-Participant Quasi-Experiments

The most common between-participant quasi-experiments are nonequivalent group designs, correlational designs, and regression-discontinuity designs.

Nonequivalent Group Designs

In a between-participant randomized experiment, participants are assigned to treatment conditions at random. A nonequivalent group design is identical to a between-participant randomized experiment except that individuals are not assigned to the treatment conditions at random. For example, individuals might be assigned to treatment conditions by self-selection or based on convenience for administrators.

With or without random assignment, different individuals will be in the different treatment groups, and these group differences (i.e., selection differences) may produce a difference in the outcome measures that would be mistaken for a treatment effect. The advantage of random assignment to treatments is not that it removes selection differences but that it makes selection differences random, which means that they can easily be taken into account using the classic methods of statistical inference. Without random assignment, selection differences are unlikely to be random, and it is much more difficult to be convinced that they have been taken into account properly.

Taking account of the biasing effects of nonrandom selection differences requires a pretreatment measure. In most cases, the best pretreatment measure is one that is operationally identical to the posttreatment measure. A number of statistical methods have been proposed for analyzing data from nonequivalent group designs so as to take account of the effects of nonrandom selection differences. The analysis strategies all use a pretreatment measure to model the effects of the nonrandom selection differences, but in most situations it is not clear which, if any, of the statistical methods are likely to produce credible estimates of treatment effects.

Implementing additional design features is the preferred approach for separating the effects of nonrandom selection differences from the effects of the treatment. Useful design elaborations include multiple comparison groups (including cohort comparison groups), dependent measures that are affected by selection differences but not by the treatment, and pretreatment measures collected at more than one point in time prior to the treatment. However, no matter the approach, the nonequivalent group design rarely produces as credible an estimate of the treatment effect as randomized experiments.

Correlational Designs

So far, it has been assumed that the independent variable is discrete. That is, it has been assumed that a discrete treatment is compared to a discrete comparison condition. Alternatively, the amount of a treatment can vary continuously rather than discretely. For example, the effects

of smoking could be assessed by comparing smokers to nonsmokers (so the treatment is a discrete variable) or by comparing individuals who smoke different amounts (so the treatment is a continuous variable). A design in which the level of the treatment varies continuously and nonrandomly is called a correlational design. Correlational designs are the counterpart to nonequivalent group designs, when the treatment is a continuous rather than a discrete variable.

As in nonequivalent group designs, a primary threat to validity in correlational designs is nonrandom selection differences. Different types of individuals tend to get different amounts of the treatment and these selection differences can bias the estimate of the treatment effect. The most common way to take account of the biasing effects of selection differences in correlational designs is to measure the sources of selection differences and control for their effects statistically. As in nonequivalent group designs, however, it is difficult to be convinced that adequate statistical controls have been applied.

In addition to selection differences, other threats to the validity of between-participant randomized experiments are also threats to the validity of correlational and nonequivalent group designs. These threats can be addressed in similar ways in the different types of between-participant designs. The primary advantage of correlational and nonequivalent group designs is that they often can be implemented when randomized designs cannot, especially in field settings

Regression-Discontinuity Designs

In a regression-discontinuity design, participants are assigned to discrete treatment conditions using a quantitative assignment variable (QAV). The participants are measured on the QAV before the treatments are introduced and assigned to treatment conditions according to a cutoff score on the QAV. All participants with QAV scores above the cutoff score are assigned to one of the treatment conditions, whereas all participants with QAV scores below the cutoff score are assigned to the other treatment condition. The treatments are then introduced and an outcome measurement is obtained on each participant.

Figure 1 is a scatterplot of data from a hypothetical regression-discontinuity design. The outcome measure is plotted on the vertical axis and the quantitative assignment variable is plotted on the horizontal axis. The vertical line indicates the cutoff score on the QAV. Individuals with QAV scores above the cutoff are placed in the experimental group, and their scores are represented by X's in Fig. 1. Individuals with QAV scores below the cutoff are placed in the comparison group, and their scores are represented by O's in Fig. 1.

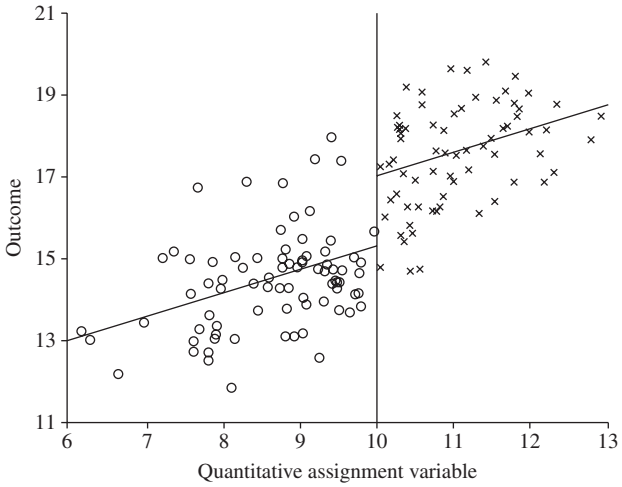


Figure 1 Data from a hypothetical example of a regression-discontinuity design.

In essence, the data from a regression-discontinuity design are analyzed in the following fashion. The scores on the outcome measure are regressed onto the QAV scores separately in each treatment condition and the two regression lines are compared. In Fig. 1, the regression lines are the two sloping lines fit through the data scatters in each group. A treatment effect is evidenced to the extent the two regression lines are not the same. If the treatment effect is a constant for all participants, a vertical displacement in the regression lines would occur at the cutoff point on the QAV. An upward displacement of the treatment group's regression line compared to the comparison group's regression line means that the treatment increases the scores on the outcome variable. Conversely, a downward displacement means that the treatment reduces the outcome scores. If the effect of the treatment interacts with the QAV score, the regression lines are not only displaced upward or downward compared to each other but also differ in slope. Figure 1 reveals a positive effect of the treatment because the regression line in the treatment groups is shifted upward compared to the regression line in the comparison group. However, there is no treatment effect interaction because the regression lines are parallel.

To obtain an unbiased estimate of the treatment effect, the regression lines in the two treatment groups must be fit correctly. For example, if the true regression surface is a straight line, a straight-line regression is the correct model to fit. However, if the true regression surface is curvilinear, a straight-line regression model is likely to produce a biased estimate of the treatment effect: A curvilinear regression model would be required. A curvilinear regression model can be fit in a variety of ways. The most common approach is to add polynomial terms to the regression equation. However, neither this

approach nor any other is likely to produce as credible estimates of effects as are produced in randomized experiments. In addition, the regression-discontinuity design requires far more participants (at least 2.7 times more in common cases) to have the same statistical power as a between-participant randomized experiment.

One of the primary advantages of the regression-discontinuity design is that it can sometimes be implemented in situations in which a randomized experiment cannot be used. In this regard, the regression-discontinuity design is likely to be particularly attractive to program administrators and staff because, unlike in between-group randomized experiments, the QAV can be used to assign participants to treatment conditions based on measures of need or merit. When it can be implemented, a randomized experiment is likely to be superior to a regression-discontinuity design. However, a regression-discontinuity design is likely to produce far more credible results than nonequivalent group or correlational designs. The primary drawback is that the regression-discontinuity design can be implemented in fewer settings than nonequivalent group or correlational designs.

Within-Participant Quasi-Experiments

Within-participant quasi-experiments most often take one of two forms: before–after designs and interrupted time series designs.

Before–After Designs

As the name suggests, in a before–after design, participants are observed on the same variable both before and after a treatment is implemented. The difference between the two measures is used to estimate the treatment effect. A before–after design is often particularly easy to implement, but it is also usually susceptible to severe bias from threats to validity.

Six threats to validity are often plausible in before–after designs. First, a maturation threat to validity occurs when changes from before to after the treatment are introduced because the participants grow older, more experienced, more fatigued, and so on. Second, a history effect occurs when an external event, which occurs between the times of the before and after measurements, causes a difference in the outcomes. Third, a testing effect arises when the mere measurement of the before observation causes a change in the after observation. Fourth, an effect due to instrumentation arises when the variable being measured before the treatment is introduced differs from the variable being measured after the treatment intervention, and this change alters the before–after difference that is observed. Fifth, attrition arises when some of the participants measured before the treatment is introduced are not assessed after the treatment is introduced. Sixth, a regression artifact can arise, for example, if the "before" measurement is obtained when the participants seek treatment because they are experiencing unusually severe problems and these problems would diminish, even without treatment, by the time of the "after" measurement.

A before–after design can produce credible results if few threats to validity are likely to be operating, such as in a learning experiment in which the content matter is unlikely to be learned in any way other than from the treatment intervention. However, such circumstances are relatively rare. The credibility of a before–after design is often improved by adding a before–after comparison using participants who do not receive the treatment. However, adding such a comparison simply converts the before–after design into a nonequivalent group design, which, as previously noted, still tends to be inferentially weak. The primary advantage of the before–after design is that it is easy to implement. However, in most circumstances, alternative designs are preferable.

Interrupted Time Series Designs

An interrupted time series design is an extension of a before–after design in which additional observations are added at points in time both before and after the treatment is implemented. That is, observations on the same variable are collected at several points in time before the treatment is introduced, the treatment is then introduced, and observations on the same variable are collected at several additional time points. Figure 2 is a plot of data from a hypothetical interrupted time series design. The outcome measure is plotted on the vertical axis and time is plotted on the horizontal axis. The vertical line in the middle indicates the time at which the treatment was introduced.

In essence, the data from an interrupted time series design are analyzed in the following fashion. A regression line is fit to the time series data that precede the treatment, and a separate regression line is fit to the time series data that follow the treatment. In Fig. 2, the regression lines are the two sloping lines fit through the two time series of data. These two regression lines are compared, and shifts in one regression line compared to the other are used to estimate the treatment effect. If the treatment effect is constant over time, an upward or downward displacement in the two regression lines should occur. An upward displacement in the "after" regression line compared to the "before" regression line means the treatment increases the scores on the outcome variable. Conversely, a downward displacement means the treatment reduces

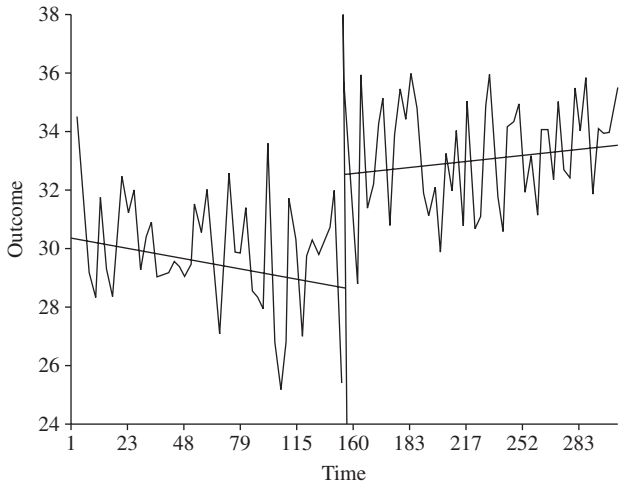


Figure 2 Data from a hypothetical example of an interrupted time series design.

the scores on the outcome variable. If the effect of the treatment varies over time, the regression lines are not only displaced upward or downward compared to each other but also shifted in slope. Figure 2 evidences a positive effect of the treatment because the "after" regression line is displaced upward compared to the "before" regression line. In addition, the effect of the treatment varies over time because the regression lines have different slopes.

Compared to the simple before–after design, the interrupted time series design tends to diminish the plausibility of the threats to validity of maturation, testing, (non-treatment-related) attrition, and regression artifacts. However, the interrupted time series design can still be susceptible to the threats to validity of history and instrumentation. The plausibility of these threats can often be reduced by adding a time series of observations collected on participants who are influenced by the same effects of history and instrumentation but who do not receive the treatment. The data in the comparison time series are analyzed in the same fashion as the data in the treatment time series. A treatment effect is evidenced to the extent the estimate of the treatment effect in the treatment condition time series differs from the estimate of the "treatment" effect in the comparison condition time series. Treatment-related attrition can also be a threat to validity and can usually be assessed by collecting additional data on the degree to which participants did not complete the study.

In some circumstances, such as in behavioral assessments in clinical psychology, interrupted time series designs are implemented with a single participant. In addition, the design is often easy to implement using archival data. An advantage of the interrupted time series design is that it can reveal how the effect of a treatment varies over time. Another advantage is that interrupted time series designs often produce relatively credible estimates of treatment effect. The obvious disadvantage is that data must be obtained over more time periods than in other designs.

Factorial Designs and Treatment Interactions

So far, this article has considered designs with only a single (either discrete or continuous) independent variable. Any of the designs described previously could be implemented (although sometimes only with substantial complications) so as to assess the effects of two different independent variables. For example, a researcher could assess the effects of (i) the independent variable of ingesting caffeine versus not ingesting caffeine and (ii) the independent variable of sleep deprivation versus no sleep deprivation. In what are called factorial designs, participants would be assigned to the four treatment combinations that are generated by these two independent variables. That is, participants would be assigned to the four conditions of (i) caffeine and sleep deprivation, (ii) caffeine and no sleep deprivation, (iii) no caffeine and sleep deprivation, and (iv) no caffeine and no sleep deprivation.

In such a factorial design, three different treatment effects can be estimated. First, the researcher can estimate the effect of caffeine versus no caffeine. Second, the researcher can estimate the effect of sleep deprivation versus no sleep deprivation. These two effects are called main effects. Third, the researcher can estimate the effect of the interaction between caffeine and sleep deprivation. If an interaction effect is present, it means that the effect of caffeine versus no caffeine is different under the condition of sleep deprivation than under the condition of no sleep deprivation.

Four Categories of Threats to Validity

Threats to validity are often partitioned into categories. This section describes the categories of threats to validity developed in a highly influential series of works by Donald Campbell, Julian Stanley, Thomas Cook, and William Shadish that have been published from 1957 to the present.

Statistical Conclusion Validity

Statistical conclusion validity concerns the proper use of statistical procedures in analyzing data. Typical threats to statistical conclusion validity are violating the

assumptions of statistical procedures such as the independence of observations, accepting the null hypothesis, having low statistical power, and increasing the chance of a type I error by data fishing.

Internal Validity

Threats to internal validity are alternatives to the treatment that could explain the observed outcome difference. Most of the threats to validity described in this article (e.g., selection differences, history, and maturation) are instances of threats to internal validity.

Construct Validity

Construct validity concerns the identification of the causes, effects, settings, and participants that are present in a study. For example, a medication might have an effect not because its putative active ingredients are absorbed into the bloodstream but because of its placebo effects. In this case, the cause would be misidentified if its effects were attributed to the absorption of the medication's active ingredients rather than to a placebo effect. Assessing the causal chain by which a treatment has its effects (i.e., determining the variables that mediate an effect) can reduce such misattributions. Alternatively, a cause can be misspecified when the treatment is implemented with less strength or integrity than intended. Manipulation checks are often used, especially in laboratory studies, to assess whether treatments are implemented as planned.

External Validity

External validity concerns the extent to which the results from a study can be generalized to other treatments, outcome variables, settings, and populations of participants. Useful generalization can sometimes be accomplished by identifying the causal mediators of an effect because similar causal mediators often produce similar effects. Generalizations can also be accomplished by identifying the moderators of an effect. When the size of a treatment effect varies with a characteristic of the participants, for example, that characteristic is said to moderate the effect. Moderators can be identified based on variation in the size of treatment effects within a study. Moderators can also be identified via meta-analysis, which is the quantitative synthesis of results from multiple studies. After a moderator has been identified, generalizations can often be made via extrapolation or interpolation.

Maximizing Validity

In many cases, changes made to strengthen one type of validity serve to weaken another type. For example, randomized experiments typically have superior internal validity compared to quasi-experiments, but randomized experiments often can be implemented only using volunteer participants, which tends to make generalizing to nonvolunteers more difficult. In general, limited resources mean that researchers should focus their attention on those types of validity that are judged to be most important in the given circumstances.

Conclusions

Estimating the size of treatment effects is often thought to be more august than "mere" measurement, perhaps because the noble goal of testing theories is done far more often by estimating treatment effects than by simply measuring, for example, public opinions. However, in fact, assessing a treatment effect is nothing more than measurement. The distal goal of a randomized experiment, for example, may be to test a theory, but measuring a treatment effect is always the proximal goal.

What distinguishes the measurement of treatment effects from other forms of measurement is the nature of the attribute that is being measured. A treatment effect is defined as the difference between what would have happened if a treatment had been implemented and what would have happened if the treatment had not been implemented but everything else had been the same. If one side of this comparison is obtained, the other cannot be, if everything else were the same. This impossibility raises unique problems for the measurement of effects.

Many types of designs can be used in place of the ideal but impossible comparison that defines a treatment effect. Each is subject to threats to internal validity. Which design is best depends on the circumstances. Some designs tend to be easier to implement. Others tend to be stronger in terms of internal validity. Yet others tend to offset relative inferiority in internal validity with greater strength in other forms of validity. Nonetheless, internal validity is of particular importance in estimating effects because it is concerned with how closely the comparison used in practice matches the ideal comparison that defines a treatment effect, which is the distinguishing feature of the measurement of effect sizes.

Internal validity is usually stronger in randomized experiment than in quasi-experiments. In quasi-experiments, threats to internal validity can be addressed with statistical procedures but are often better addressed by adding design features. The fundamental challenge for the researcher is determining which threats to internal validity are most plausible and which design features are likely to control them most effectively.

See Also the Following Articles

Correlations • Experiments, Overview • Explore, Explain, Design • Quasi-Experiment • Randomization • Research Designs • Sample Design • Survey Design

Further Reading

Aronson, E., Brewer, M., and Carlsmith, J. M. (1985). Experimentation in social psychology. In *The Handbook of Social Psychology* (G. Lindzey and E. Aronson, eds.), 3rd Ed., Vol. 1, pp. 441–486. Random House, New York.

Boruch, R. F. (1997). *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Sage, Thousand Oaks, CA.

Campbell, D. T., and Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research.* Rand McNally, Chicago.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Rand McNally, Chicago.

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs.* Jossey-Bass, San Francisco.

Judd, C. M., and Kenny, D. A. (1981). *Estimating the Effects of Social Interventions.* Cambridge University Press, New York.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences,* 3rd Ed. Brooks/Cole, Pacific Grove, CA.

Maxwell, S. E., and Delaney, H. D. (1990). *Designing Experiments and Analyzing Data: A Model Comparison Perspective.* Wadsworth, Belmont, CA.

Mohr, L. B. (1995). *Impact Analysis for Program Evaluation,* 2nd Ed. Sage, Thousand Oaks, CA.

Overman, E. S. (ed.) (1988). *Methodology and Epistemology for Social Science: Selected Papers of Donald T. Campbell.* University of Chicago Press, Chicago.

Reichardt, C. S., and Mark, M. M. (1998). Quasi-experimentation. In *Handbook of Applied Social Research Methods* (L. Bickman and D. J. Rog, eds.). Sage, Thousand Oaks, CA.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin, Boston.

Trochim, W. M. K. (ed.) (1986). *Advances in Quasi-Experimental Design and Analysis*, New Direction for Program Evaluation No. 31. Jossey-Bass, San Francisco.

Type I and Type II Error

Alesha E. Doan
California Polytechnic State University, San Luis Obispo, California, USA

Glossary

alternative hypothesis An expectation or prediction that is tested; also referred to as the research hypothesis.
coefficient An unknown but fixed parameter estimated in a regression equation. The coefficient refers to the intercept coefficient and slope coefficient.
dependent variable A variable that is influenced or caused by another variable.
***F* test** An overall test of significance that can be used in a multiple regression equation. The test determines whether all of the slope coefficients simultaneously equal zero. The test is based on an analysis of the total sum of squares, which encompasses the explained sum of squares and the residual sum of squares.
independent variable A variable that influences or causes another variable.
null hypothesis A statement that is the opposite of the alternative hypothesis, stating that there is not a relationship between the independent and dependent variables.
***p* value** The exact level of significance; or, the exact probability of committing a type I error.
***t* statistic** A test of significance that uses the sample results to determine the accuracy of the null hypothesis (e.g., whether the null hypothesis should be accepted or rejected).
type I error Accepting the alternative hypothesis when the null hypothesis is true.
type II error Accepting the null hypothesis when the alternative hypothesis is true.

Committing a type I error or a type II error refers to the probability of erroneously rejecting the null hypothesis or, conversely, accepting the null hypothesis in error. When testing hypotheses, researchers are faced with a dilemma—attempting to minimize the risk of committing a type I error increases the probability of committing a type II error. The opposite relationship holds as well: minimizing the risk of type II error leads to an increased probability of committing a type I error. Both errors present unique problems for the investigator and can result in serious consequences for statistical inferences.

Introduction

Explanation of Type I Error and Type II Error

The goal of conducting research is often to test relationships between two or more variables (dependent variable, Y_i, and explanatory variable(s), X_i) within a sample and draw inferences to a larger population. Investigators set up competing research hypotheses and use statistical analysis to determine if there is empirical support in favor of one hypothesis over the other. Two hypotheses are derived, the null hypothesis, which is often notated as H_0, and the alternative hypothesis, notated as H_a. In empirical research, the null hypothesis typically tests the relationship H_0: $\beta_2 = 0$, which indicates that the slope coefficient is zero, or stated differently, there is no relationship between Y_i and X_i. The alternative hypothesis can be stated in a variety of forms, usually depending on theoretical expectations or prior empirical research. However, for simplicity, the following alternative hypothesis will be used H_a: $\beta_2 \neq 0$. The alternative hypothesis states that the slope coefficient is not equal to zero.

Unfortunately, as with all statistical analysis, a degree of uncertainty exists. An investigator is always confronted with the possibility of accepting the null or alternative hypothesis in error. If a researcher finds empirical support for the alternative hypothesis, thus rejecting the null hypothesis when in fact the null hypothesis is correct, the researcher has committed a type I error.

Alternatively, if a researcher accepts the null hypothesis when the alternative hypothesis is actually correct, then he or she has committed a type II error. Historically, researchers have predominantly focused on type I error; however, both types of errors can be costly.

Significance of Committing a Type I versus a Type II Error

Depending on the purpose of the research, the cost of committing a type I or type II error can be great, because it leads to drawing incorrect inferences and conclusions that can have potentially severe consequences for society. Although the hypothesis has been simplified, the following example highlights the significance of committing a type II error. Early medical trials assessing the benefits of using hormone replacement therapy (HRT) in postmenopausal women concluded that the use of HRT did not pose significant harm to women (H_0: HRT is not detrimental to a woman's health). Based on these studies, physicians routinely prescribed HRT for postmenopausal women. Unfortunately, subsequent research determined that HRT posed significant risks for women, such as an increased likelihood of developing breast cancer, coronary heart disease, and stroke. Clearly, a type II error had been committed in the early studies. Researchers had erroneously accepted the null hypothesis when in fact the alternative hypothesis (H_a: HRT is detrimental to a woman's health) was correct. In this case, committing a type II error resulted in thousands of postmenopausal women being exposed to serious health problems resulting from the use of HRT. The cost of making a type I or type II error is significant and can result in serious consequences for both researchers and society.

Article Overview

The remainder of this article explores how type I and type II errors arise in research. First, the symbols that are used throughout the article are explained, including a summary table. The next section examines the initial decision-making process regarding the formation of hypothesis. More specifically, attention is given to the reasons for setting up a null and alternative hypothesis. The issues pertaining to determining a level of statistical significance in research are explored (e.g., what is the relative advantage of using alpha (α), a probability value (p value), or confidence intervals), and the relationship between significance levels and type I and type II errors is discussed. The following sections review common misperceptions of type I and type II errors as well as some strategies for minimizing both types of errors. The article concludes with a discussion of the trade-offs between type I and type II errors.

Table I Symbols Used

α	The Greek letter alpha, used to represent a type I error.
β	The Greek letter beta, used to represent a type II error.
β_2	The slope coefficient in a regression analysis.
H_0	The null hypothesis.
H_a	The alternative hypothesis.
X_i	The independent variable in an analysis.
Y_i	The dependent variable used in an analysis.
N	The sample size of a study.
F test	An overall test of significance; used to determine if multiple slope coefficients simultaneously equal zero.
p value	The level of significance.
t statistic	A test of significance.
TSS	The total sum of squares; the total variation of the actual Y values about their sample mean.
ESS	The explained sum of squares; the variation of Y that is explained by the regression line.
RSS	The residual sum of squares; the unexplained variation of the Y values in the regression line.

Definition of Symbols

The symbols and notation used throughout this article mirror the conventional, standard Greek symbols and notation widely used in statistical textbooks and articles. Table I lists and defines all the symbols used in this article.

The Formation of Hypotheses

Generating Hypotheses

In empirical research, an investigator generates an expectation or set of expectations to test via statistical analysis. The expectations emerge from an exploration of existing theoretical or empirical work on the given topic, or from a combination of the two. The expectation is stated as a hypothesis, which is simply a statement to be tested. The generation and testing of hypotheses is a process that aids a researcher in making rational decisions about the veracity of the effects being investigated.

Importantly, the hypotheses must be established prior to the actual empirical testing, otherwise the work runs the risk of being *post hoc* and circular in reasoning. If a researcher engages in "data mining" (e.g., generating his or her expectations to match the results of the empirical analysis), then the researcher is guilty of foregoing the scientific method, thus largely invalidating the p values from that specific empirical test. *Post hoc* studies are not considered valid and should be avoided.

The Alternative or Research Hypothesis

The alternative (also referred to as the research) hypothesis is the expectation a researcher is interested in investigating. Typically, the alternative hypothesis will naturally emerge from the particular area under investigation. An alternative hypothesis may, for example, test whether the slope coefficient (β) is not zero (H_a: $\beta_2 \neq 0$), or it may test whether a relationship exists between the dependent variable (Y_i) and the independent variable (X_i). In this case, the alternative hypothesis is a two-sided hypothesis, indicating that the direction of the relationship is not known. Typically a two-sided hypothesis occurs when a researcher does not have sound *a priori* or theoretical expectations about the nature of the relationship. Conversely, in the presence of strong theory or *a priori* research, the alternative hypothesis can be specified according to a direction, in which case it takes on the following forms: H_a: $\beta_2 > 0$ or H_a: $\beta_2 < 0$.

The Null Hypothesis

The null hypothesis is set up to reflect the opposite statement from the alternative hypothesis and can take on multiple forms. For example, the null hypothesis may state that a relationship does not exist between the dependent (Y_i) and independent (X_i) variable. Or, the slope coefficient is expected to be zero (H_0: $\beta_2 = 0$). The null hypothesis is actually the hypothesis that is tested, not the alternative hypothesis. If the empirical testing does not produce any evidence for the alternative hypothesis, then the researcher fails to reject the null hypothesis. Similarly, if evidence is found in support of the alternative hypothesis, then the researcher fails to accept the null hypothesis. Importantly, because most empirical research is based on a sample rather than the true population, the alternative hypothesis (and likewise, the null hypothesis) is not "proved." A possibility always exists that more empirical tests, based on another sample, will yield different results, or another null hypothesis exists that is just as compatible with the empirical evidence. Therefore, a researcher is simply stating that based on the current sample evidence, there is reason to either accept or fail to accept the null hypothesis.

Determining Significance Levels

Choosing a Level of Significance

After generating the research and null hypothesis, a level of significance (α) must be established in order to determine the point at which a researcher will either accept or reject the null hypothesis. Essentially, hypothesis testing hinges on the level of significance chosen by the researcher. The level of significance is simply the probability of committing an error and the investigator—prior to the empirical analysis—sets it. In other words, the level of significance is the probability of committing a type I error. Of course, the more stringent a researcher is in setting the significance level (e.g., 0.05, 0.01, 0.001), the less likely the conclusions will be incorrect. Although the level of significance can be set by the investigator, in social science research, α is conventionally set at 1%, 5%, and occasionally at 10%. However, the α level is flexible and usually depends on several factors, such as the type and purpose of the research under investigation, the sample size, and the costs associated with drawing incorrect conclusions.

Relationship between α and Type I and Type II Errors

The α level is critically linked to the probability of committing a type I or type II error. Unfortunately, trying to reduce type I error (typically represented by α) leads to an increased probability of committing a type II error (typically represented by β). Type I and type II errors are inversely related and cannot be simultaneously minimized. Recall that type I error occurs when a researcher falsely rejects the null hypothesis. In other words, he or she found a false positive relationship. Conversely, a type II error exists when the null hypothesis is falsely accepted. The researcher's evidence suggests that there is not a significant effect, but in the population the effect is significant. In both cases, the researcher's sample has failed, either erroneously indicating an effect or no effect, respectively, in the population. Table II contains a summary table defining type I and type II errors.

By setting the α level, a researcher is actually setting the parameters for committing a mistake. In other words, α determines how much risk an investigator is willing to accept in his or her research. For example, by setting α at 5%, a researcher is accepting the chance of committing a mistake one in 20 times. Stated differently, a researcher will reject the null hypothesis when it is actually correct one in 20 times. Historically, statisticians have advocated for reducing the probability of type I error over type II error by setting α at a low level.

As a researcher attempts to mitigate against type I error, he or she increases the risk of committing a type II error. However, whereas the α level is determined in advance by a researcher, the β error rate is much more difficult to determine. In order to set β, a researcher

Table II Decision Analysis

Test decision	*H_0 is true*	*H_0 is false*
Reject H_0	Type I error	No error committed
Accept H_0	No error committed	Type II error

would have to know the distribution of the alternative hypothesis, which is rarely known. Consequently, the concept of power can be used to reduce the probability of committing a type II error.

Power refers to the probability of not committing a type II error. Stated differently, power is the likelihood that a test will reject the null hypothesis when it is truly false; it is determined by taking $1 - \beta$. Power can be strengthened by increasing the sample size of the study. In fact, consideration to type II error should be given at the initial stage of research design, when the investigator assesses how large of a sample is needed to detect any potentially significant effects.

Alternatives to α: *P* Value and Confidence Intervals

Instead of setting the α level, which is often arbitrary or done out of convention, a researcher can use a test statistic (e.g., the t statistic) to find the p value. The p value is the probability value; it provides the exact probability of committing a type I error (the p value is also referred to as the observed or exact level of significance). More specifically, the p value is defined as the lowest significance level at which the null hypothesis can be rejected. Using the test statistic, a researcher can locate the exact probability of obtaining that test statistic by looking on the appropriate statistical table. As the value of the test statistic increases, the p value decreases, allowing a researcher to reject the null hypothesis with greater assurance.

Another option in lieu of relying on α is to use a confidence interval approach to hypothesis testing. Confidence intervals can be constructed around point estimates using the standard error of the estimate. Confidence intervals indicate the probability that the true population coefficient is contained in the range of estimated values from the empirical analysis. The width of a confidence interval is proportional to the standard error of the estimator. For example, the larger the standard error of the estimate, the larger the confidence interval, and therefore the less certain the researcher can be that the true value of the unknown parameter has been accurately estimated.

The null hypothesis is frequently set up as an empirical straw man because the objective of empirical research is to find support for the alternative hypothesis (hence the conventional wisdom that null findings are not newsworthy findings). The null hypothesis may reflect a fairly absurd scenario that is actually used to dramatize the significance of empirical findings. Consequently, some econometricians argue for the use of confidence intervals, which focus attention on the magnitude of the coefficients (findings) rather than on the rejection of the null hypothesis. According to De Long and Lang (1992) "if all or almost all null hypotheses are false, there is little point in concentrating on whether or not an estimate is indistinguishable from its predicted value under the null" (p. 1257).

Both of these options present alternatives to simply choosing a level of significance. The p value yields an exact probability of committing a type I error, which provides the researcher with enough information to decide whether or not to reject the null hypothesis based on the given p value. Using confidence intervals differs in approach by concentrating on the magnitude of the findings rather than the probability of committing a type I error. Every approach to hypothesis testing—using α, p values, or confidence intervals—contains some amount of trade-offs. Ultimately, a researcher must decide which approach, or combination thereof, suits his or her research style.

Minimizing Type I and Type II Errors

Common Misperceptions

Conventional statistical wisdom often suggests that in order to reduce the likelihood of committing a type I or type II error, a researcher should increase the sample size of the empirical study. Greater confidence is typically attributed to the findings from studies that have a large sample size. This assertion, however, is only partially correct and can lead to inaccurate interpretations of research results.

Increasing the sample size will in fact reduce the probability of committing a type II error. In a smaller sample, detecting treatment effects is much more difficult compared to a larger sample. For example, consider two identical studies with varying sample sizes. The first study has a sample size of 25 ($N = 25$), whereas the second study has a sample size of 250 ($N = 250$). Using the same level of significance, the treatment effect in the first study would have to be three times larger than the treatment effect in the second study in order to be detected. This disparity results from the formula used to derive statistical significance, which contains the sample size in the denominator of the formula. Consequently, increasing the sample size will reduce the probability of committing a type II error, but not a type I error. Rather, reducing the α level is the only way to reduce the probability of committing a type I error.

Reducing the Probability of Committing a Type I Error

The probability of committing a type I error can dramatically increase when testing for multiple hypotheses. Indeed, when relying on the t statistic for multiple comparisons, the α level will become inflated,

leading to a greater propensity for a type I error. Essentially, the problem occurs because as the sample size increases, the value of the range increases at a faster rate than the value of the standard deviation. The t statistic, while accounting for the standard deviation in the formula, does not consider the larger increase in range in relation to the standard deviation. Consequently, when used in multiple comparisons, the t test will inflate the α level, thus leading to a greater risk of committing a type I error.

Although a perfect solution does not exist, there are several measures a researcher can take to mitigate the risk of committing a type I error. Two major approaches include using ordered p values or employing a comparison of normally distributed means. The easiest ordered p value approach is based on the first-order Bonferroni inequality. Simply stated, the Bonferroni technique takes into account the number of hypotheses being tested and adjusts the p value accordingly. For example, if three hypotheses are being tested, the p value is adjusted downward by dividing the p value by three (e.g., 0.05/3). The new p value becomes 0.02, based on the rules of probability. The Bonferroni adjustment should only be applied if the hypotheses are assumed to be independent; if they are interrelated, the Bonferroni adjustment is too severe.

Similarly, several comparisons of normally distributed means techniques exist. Three of the more common techniques are the Scheffe, Tukey, and Dunnett methods. Each of these techniques differs slightly in its approach according to the specific means contrasts under investigation. Despite their variations, they share several elements. First, they are particularly designed to deal with comparison of means; second, they assume normally distributed observations; and third, they are premised on the joint distribution of all observations. In the end, these approaches enable a researcher to decrease the likelihood of committing a type I error.

Reducing the Probability of Committing a Type II Error

As discussed previously, the most straightforward—although not always feasible—way to decrease the probability of committing a type II error is simply to increase the sample size. If increasing the sample size is not practical, then Fisher's ANOVA procedure can be employed. ANOVA is the analysis of variance, which is a study of the total sum of squares (TSS) in a regression analysis. The TSS is a combination of the explained sum of squares (ESS) and the residual sum of squares (RSS). ANOVA can be used to calculate an F ratio (assuming the disturbances are normally distributed). The F ratio follows the F distribution and provides a test of the null hypothesis (H_0: $\beta_2 = 0$). Fisher's test minimizes the probability of a type II error; however, the trade-off is that it increases the chance of making a type I error.

Conclusion: The Trade-off between the Two Errors

Type I and type II errors present unique problems to a researcher. Unfortunately, there is not a cure-all solution for preventing either error; moreover, reducing the probability of one of the errors increases the probability of committing the other type of error. Although a researcher can take several measures to lower type I error, or alternatively, a type II error, empirical research always contains an element of uncertainty, which means that neither type of error can be completely avoided.

Type I error has historically been the primary concern for researchers. In the presence of a type I error, statistical significance becomes attributed to findings when in reality no effect exists. Researchers are generally adverse to committing this type of error; consequently, they tend to take a conservative approach, preferring to err on the side of committing a type II error. The major drawback to exclusively emphasizing type I error over type II error is simply overlooking interesting findings. Typically, once statistical relationships are discovered, more studies follow that confirm, build upon, or challenge the original findings. In other words, scientific research is cumulative; therefore, false positives are revealed in subsequent studies. Unfortunately, in the presence of a type II error, the line of inquiry is often discarded, because in most fields of research, a premium is placed on statistically significant results. If a type II error has been committed and that particular line of inquiry is not pursued further, the scientific community may miss valuable information.

Ultimately, the scientist must decide which type of error is more problematic to his or her research. Essentially, the investigator is confronted with the question of what type of error is more costly. The answer to this question depends on the purpose of the research as well as the potential implications in the presence of a false positive (type I error) or false negative (type II error) findings.

See Also the Following Articles

Confidence Intervals • Data Mining • Hypothesis Tests and Proofs • Sample Size

Further Reading

Dunlap, W. P., Greer, T., and Beatty, G. O. (1996). A Monte-Carlo study of Type I error rates and power for Tukey's Pocket Test. *J. Gen. Psychol.* **123,** 333–339.

De Long, B. J., and Lang, K. (1992). Are all economic hypotheses false? *J. Polit. Econ.* **100,** 1257–1272.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assn.* **50,** 1096–1121.

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance cases. *J. Am. Stat. Assn.* **75,** 796–800.

Dunnett, C. W., and Tamhane, A. C. (1992). A step-up multiple test procedure. *J. Am. Stat. Assn.* **87,** 162–170.

Fisher, R. A. (1935). *The Design of Experiments.* Oliver & Boyd, Edinburgh.

Fisher, R. A. (1952). Sequential experimentation. *Biometrics* **8,** 183–187.

Fisher, R. A. (1955). Statistical methods and scientific induction. *J. Roy. Stat. Soc.* **17,** 69–78.

Ganzach, Yoav. (1998). Nonlinearity, multicollinearity and the probability of type II error in detecting interaction. *J. Manage.* **24,** 615–623.

Gill, Jeff. (1999). The insignificance of null hypothesis significance testing. *Polit. Res. Q.* **52,** 647–674.

Gujarati, D. N. (1995). *Basic Econometrics,* 3rd Ed. McGraw-Hill, Inc., New York.

Kim, K., and DeMets, D. L. (1997). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74,** 149–154.

Sato, T. (1996). Type I and Type II error in multiple comparisons. *J. Psychol.* **130,** 293–303.

Scheffe, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40,** 87–104.

Scheffe, H. (1959). *The Analysis of Variance.* Wiley, New York.

Scheffe, H. (1970). Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios. *Ann. Math. Stat.* **41,** 1–19.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Ann. Rev. Psychol.* **46,** 561–585.

Tukey, J. W. (1994). *The Collected Works of John W. Tukey, Volume VIII: Multiple Comparisons* (H. I. Braun, ed.) Chapman & Hall, New York.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics* **5,** 99–114.

Wilkerson, M., and Olson, M. R. (1997). Misconceptions about sample size, statistical significance, and treatment effect. *J. Psychol.* **131,** 627–632.

Typology Construction, Methods and Issues

Kenneth D. Bailey
University of California, Los Angeles, California, USA

Glossary

classification The grouping of entities by similarity.
constructed type A qualitative type used to compare empirical cases.
ideal type A qualitative type with accentuated dimensions.
monothetic type A type in which all members are identical on all characteristics.
polythetic type A type in which all members are similar, but not identical.
reduction The process of reducing the number of types in a typology.
substruction The process of identifying and expanding the underlying dimensions of a type in order to form a full typology.
taxonomy An empirical typology used in biology and related fields.
three-level model An expanded measurement model recognizing the conceptual, empirical, and indicator or document levels.
typology A multidimensional classification, generally conceptual rather than empirical.

A typology is an array or complete set of types. It is a multidimensional classification. Typology construction is part of the more general process of constructing and utilizing classification schemes of various sorts. The use of types is endemic in everyday speech. It is common to speak of types of birds, types of trees, and so forth. Traditional typologies were generally qualitatively derived. They often remained conceptual, without empirical referents. More recently, typologies have been quantitatively derived through the computerized statistical analysis of empirical specimens. This approach has incorporated a variety of techniques, variously called cluster analysis, numerical taxonomy, or pattern recognition.

Classification

If classification in general, and typology construction in particular, has such a great value for social theory and measurement, it seems logical that scholars would devote special effort to its development and explication. Sadly, this is generally not the case. Typology construction is often taken for granted, if not neglected. Most of the literature on typology construction in social science is quite old. Much like the foundation of a building, the typological foundation of social measurement is often out of view, and thus easily out of mind. This is unfortunate, because social theory and measurement are often no stronger than their typological underpinning. If the latter is weak, so is the former. Yet ironically, something as important as typology can often become seemingly invisible, not only to the lay public, but even to many scholars as well. Terms such as typology, substruction, reduction, or classification elicit little name recognition. The only concept that many people seem to recognize is that of "sorting," a term that ironically is generally absent from the formal lexicon of typology construction. Many people have a drawer in their kitchen for sorting silverware into spoons, forks, and knives, and this seems to be about the extent of their awareness of classification principles and procedures. Social scientists often recognize the concepts of "ideal types" and "polar types," but this may represent the extent of their awareness of typological principles and procedures as well.

Basic Definitions

Classification

Classification is the generic process for grouping entities by similarity. The goal is to make each cell of the classification as similar as possible (to minimize within

group variance). This may involve maximizing between-group variance, by maximizing the distance between each cell. The individual cell or category within the larger classification is called a class. The term "classification" is used to refer both to the process and the result of the process, as in "The classification process produced an excellent classification." An adequate classification must be simultaneously mutually exclusive and exhaustive. In other words, the classification must provide a place (but only one place) for every individual in the sample.

Typology

A typology is a multidimensional classification. The term "typology" is often reserved for classification schemes that are verbal, or said to be conceptual, theoretical, or "heuristic," as opposed to empirical. A single cell of a full typology is called a type. Types can include both heuristic and empirical types. Heuristic types are conceptually derived, and thus may lack empirical examples. Empirical types are constructed entirely through empirical data analysis, and thus may lack precise conceptualization.

Taxonomy

A taxonomy is an empirical classification, generally associated with the field of biology. A single cell of the taxonomy is known as a taxon.

Numerical Taxonomy

The process of generating an empirical classification schema quantitatively is called numerical taxonomy. One generally begins with a set of empirical cases believed to have certain similarities. Then various algorithms are selected to mathematically group the empirical data according to similarity. Numerical taxonomy is virtually always a computerized analytic procedure. Although not limited to biology, numerical taxonomy was largely developed within biology, and is often associated with it.

Cluster Analysis

This procedure is similar to numerical taxonomy, and in fact the two terms are often used interchangeably. Cluster analysis also refers to a variety of techniques for quantitatively grouping data by similarity, generally (but not universally) on the computer. The term cluster analysis is not as closely associated with biology as is numerical taxonomy. Cluster analysis techniques are used by researchers in a wide variety of disciplines, including sociology, anthropology, psychology, and education.

Divisive Methods

These are methods which begin by treating the entire original sample as one cluster and then successively dividing it until the desired set of clusters is reached. Divisive methods are generally (but not always) computerized.

Agglomerative Methods

Agglomerative methods are the converse of divisive methods. They begin by treating each of the N cases in the original sample as N different clusters. Then these N clusters are successively agglomerated into a smaller number of classes. Although noncomputerized agglomerative methods do exist, most agglomerative methods fall under the rubric of either numerical taxonomy or cluster analysis. They generally proceed by utilizing some quantitative, computerized algorithm to iteratively generate the desired set of clusters M, where $1 < M < N$. Both agglomerative and divisive methods use algorithms or quantitative rules that dictate how clusters are formed. Though these are generally utilized in the form of "canned" computer programs, there may be decisions that researchers have to make in using them, concerning such things as the level of similarity required for adding an additional case to a preexisting cluster or even the desired number of clusters.

Identification

Conceptual verbal typologies and empirically derived concrete taxonomies represent different dimensions. Typologies may be purely conceptual, while taxonomies may be purely empirical. Each of these alternatives may prove sufficient for some research purposes, depending upon the researcher's particular goals. A verbal typology may suffice for heuristic purposes, or for explicating certain theoretical notions. Conversely, a taxonomy may be sufficient if one desires primarily an empirical description of cases and does not desire further theoretical analysis. However, in many instances the researcher may not be satisfied with either a purely verbal typology or with a computer-generated empirical mathematical taxonomy. If a researcher has constructed a purely verbal or conceptual typology, he or she may wish to find empirical cases to fit one or more of the verbal type cells. This process of finding empirical cases to fill in the cells of verbal typologies is called identification.

Qualitative Typologies

The typological tradition has a long history in social science, especially in sociology. Here, types were often viewed as multidimensional concepts, or constructs.

Often these types were utilized individually or in pairs, without the benefit of a full typology. Such a verbal type was generally seen as a set of correlated variables or dimensions that were somehow "connected" to each other. For example, "student athlete" is a multidimensional type concept comprising five separate but intercorrelated dimensions: (1) university affiliation; (2) sex; (3) age; (4) nationality; and (5) amateur standing.

The Ideal Type

Perhaps the most famous type concept is Weber's ideal type. This frustrating concept has proven to be both valuable and confusing, as scholars have argued over exactly what it is, and how to utilize it. As defined by Weber, the type concept is exaggerated on one or more of its basic dimensions, generally to the point that it cannot be found empirically (at least in its most exaggerated form). Weber states explicitly that the ideal type is a utopia. However, he also specifies that a given ideal type should be used as a criterion point to measure the degree to which a specific empirical case diverges from it. For example, Weber says that a researcher can begin with the ideal type of "city economy," and then study the degree to which the economy of an actual city departs from this ideal type.

Weber's strategy is actually ingenious, which explains all the acclaim that it has received. However, it is also unfortunately vulnerable to misinterpretation, which explains the criticism it has received, generally from critics who simply failed to understand its underlying logic and its proper use. Critics have seized upon Weber's comments that the ideal type is a utopia that "cannot be found empirically anywhere in reality." Critics have misconstrued this to mean that the ideal type is "imaginary" or hypothetical and thus lacking a fixed position in typological space. This is not true. To say that an ideal type is a utopia does not mean that it is imaginary or lacks a fixed location in typological space. It simply means that it occupies such an extreme position in the typology that examples of it will rarely (or ever) be found empirically. Critics charge that since the ideal type is imaginary, it can be placed anywhere the researcher desires, thus rendering it useless as a comparative tool. This is not a fair representation of the ideal type. The ideal type cannot be moved or revised at will, but instead occupies a firmly fixed position or location in typological space. In reality, types that are similar to the ideal type can be routinely found empirically, but without the perfection of the ideal type.

A contemporary example is the system currently in use for grading collectible United States coins. MS-70 is the highest, or perfect grade. It is the equivalent of the ideal type. It is a utopia as Weber stated, but it is certainly not imaginary, as critics have charged. It is very real and clearly defined in a multidimensional state, but it is simply difficult to attain as it represents perfection. It is, however, very valuable for grading coins of lesser grade. The MS-70 demonstrates what a perfect specimen would look like, and thus documents the degree of imperfection of the coin being graded. Coins grading around MS-60 or MC-62 are abundant, but the ideal type coin of MS-70 is rarely if ever found.

What Weber was doing by comparing all empirical cases with a single ideal type was effectively reducing the entire typology (which might have contained hundreds of individual types), down to a single type for comparative purposes. In Weber's precomputer era, a large typology was entirely too unwieldy to use efficiently. For example, if a researcher devised a 10-dimensional typology, the minimum size of this typology (if all dimensions were merely dichotomies) would be 1024 cells. If some or all of the dimensions contained more than two categories, the typology could be much larger, encompassing thousands of cells. Problems with using such an unwieldy typology for studying actual empirical cases are not limited to the number of cells, but stem from the number of dimensions as well. If a researcher wants to use a ten-dimensional typology, he or she is greatly constrained by the fact that only two dimensions at a time can be represented on a two-dimensional sheet of paper.

Fortunately, at this point Weber's ideal type emerges as a valuable compromise. Weber offered a single cell for the typology, generally one that was the most visible (often the most extreme) on each of the dimensions comprising the typology. The researcher was now spared the complexity of attempting to work with thousands of types. He or she could simply identify each empirical case, measure the degree that it departed from the ideal type on each dimension, and specify the theoretical typological cell corresponding to this case.

Constructed Types

An alternative to the ideal type is the constructed type, as popularized by McKinney. The constructed type is also a single type (like the ideal type, but without the mystery and confusion of the ideal type). As defined by McKinney, the constructed type is a purposive combination of a set of criteria with empirical referents. Although these criteria may be accentuated, they are generally not as extreme as an ideal type. While not an average, the constructed type is designed to be more central than the ideal type. This means that it is nearer in value to the empirical cases it is being compared with, thus facilitating measurement. The constructed type is designed to serve as a basis for the comparison of empirical cases.

Polar Types

A typologist who is reluctant to depend on a single ideal type may choose to add additional types, but without

going to the full typology. One popular strategy is the use of polar types. These types represent both extremes of the correlated dimensions comprising the type concept. That is, one of these would generally be the ideal type, and the other would be its polar opposite. One advantage of this strategy is that it mirrors the common practice of reducing empirical reality into simple dichotomies, rather than continuous dimensions. Common polar type pairs are gemeinschaft/gesellschaft, local/cosmopolitan, and rural/urban. This strategy could be further extended by choosing four types, such as the four corners of the typology, or by utilizing one or both of the diagonals of the typology.

Ethnographic Types

In addition to ideal, constructed, and polar types, another form of qualitative type is the ethnographic type. "Heuristic types" such as ideal and constructed types are generally deductively derived without empirical data analysis. In contrast, ethnographic types, while qualitatively produced, tend to be inductively derived through ethnographic analysis of empirical data (so called "grounded theory"). The standard procedure is for ethnographers to conduct field research on a particular group and to derive empirical type concepts that serve as descriptive labels for the varieties of social phenomena observed. A complicating factor, though, is that there are often two separate sets of qualitatively derived empirical types. One set is generated by the outside observer (ethnographer) while the other set is derived by the inside group members themselves. The existence of two parallel sets of types from the same phenomenon is an indication (at least to some degree) of an existing "insider–outsider" distinction.

Insider types tend to be more specific and more favorable than outsider labels. For example, among the labels that transients use to describe themselves are "bindle stiff" and "fruit tramp." These type concepts are generally unknown to the lay public, who tend to use alternative and more stigmatized labels such as "bum" or "wino." Note that like classical heuristic types, ethnographic types represent a purposive selection of types (the ones evident in the analysis) rather than a full typology.

Reduction

The use of ideal and constructed types for the identification and comparison of empirical cases can be seen as a rudimentary form of reduction. Lazarsfeld introduced the dual procedures of reduction and substruction of typologies. Reduction is the process of reducing the complexity of an unmanageable typology. Lazarsfeld introduced three forms of typological reduction—functional, pragmatic, and arbitrary numerical. All three forms of reduction assume that a full typology (for example, of 1024 cells) exists, but that the researcher desires to reduce the size and complexity of the typology.

Functional reduction is similar to the process we described for the ideal type, but does not entail prior specification of a criterion type such as an ideal or constructed type. With functional reduction, one first specifies the full typology, then seeks to identify empirical cases for the respective cells. Cells for which empirical examples can be found are retained, while the remainder (null cells) are removed from the typology. The second form of reduction is pragmatic reduction. This consists of collapsing together a number of contiguous cells in the typology. While having the advantage of reducing the total number of cells, it has the disadvantage of increasing the within-cell heterogeneity of the new aggregated type.

Lazarsfeld's third form of reduction is arbitrary numerical reduction. This is essentially an unequal-weighting scheme whereby two or more distinct and unequal types are rendered equal, and thus one or more becomes redundant and can be removed from the typology. In elaborating this strategy, Lazarsfeld presented each dimension of the type as a dichotomy, so that binary coding could be used. Thus, in a type of three dimensions, each of the three could be coded as either present (1) or absent (0) in a particular study. This yields eight possible code patterns from (0,0,0) to (1,1,1). Lazarsfeld's basic strategy entails setting some of these patterns equal through unequal weighting. For example, paraphrasing Lazarsfeld, imagine that we are constructing an index of residential housing attractiveness based on three dimensions—central heat, a fireplace, and a swimming pool. A home possessing all three of these would be the most attractive, and would be coded (1,1,1), while a home lacking all of these would be least attractive, and would be coded (0,0,0). Arbitrary numerical reduction, as the name implies, is both arbitrary and numerical. It entails specifying (somewhat arbitrarily) unequal weightings. For example, we could say that central heat is more important than the other two factors, so that a home with central heat, but lacking the other two, is equally attractive to a home lacking central heat, but possessing both a fireplace and a swimming pool. In coded form, $(1,0,0) = (0,1,1)$. This means that the original eight three-dimensional types have now been reduced to seven.

A huge typology that is theoretically rich is nonetheless of little use if it remains too complex to be used efficiently. At first glance it might seem that this was chiefly a problem for typologists such as Weber working in a time before computers. Now, typologists can let the computer deal with the complexity of a large typology, without the need for the drastic reduction of types. In reality, though, this is a gross oversimplification. Although the computer certainly stores more types than any device available in

Weber's time, computers have not effectively dealt with the monumental problems of interpreting the complex data. Thus, while computers may store a typology of multiple dimensions, they are not yet able to present them all simultaneously in a format that researchers can easily comprehend and interpret.

A researcher might construct a 10-dimensional typology, yet when he or she is ready to print it, the reality remains that the paper that the results are printed on remains intractably two-dimensional. The 10 dimensions cannot be presented simultaneously, but are generally printed two at a time. That is, dimensions one and two are printed, then dimensions one and three, and so forth. The worry is that the holistic nature of the full typology is almost totally lost. With no way to present the congruence of all dimensions simultaneously, the typologist can only view a piecemeal two by two presentation. This loses most of the holistic richness and complexity of the typology.

Thus, ironically, the need for reduction remains almost as strong in the computer era as it was in the precomputer era of Weber's time. Now, more dimensions can be stored, but they still cannot be simultaneously viewed and interpreted. In reality, rather than printing the *N*-dimensional typology two dimensions at a time, it may still be better to reduce the full typology to a few key types, each which can be viewed holistically, with its full complexity intact.

Substruction

Substruction is the opposite of reduction, and was also explicated by Lazarsfeld. Many times the literature does not present a full multidimensional typology, and so there may be no pressing need for reduction. However, the opposite problem may arise. That is, a writer may present a complex type such as "gemeinschaft" or "cosmopolitan," without adequately specifying all of its underlying dimensions. It may be evident to the reader that the type is multidimensional rather than unidimensional, but he or she may not know exactly what the latent dimensions are. The process of substruction entails identifying the underlying dimensions so that other relevant types, or even the full typology, may be constructed. This process of extending the complete property space and the resulting complete typology from one or a few multidimensional types is called substruction. Barton provided an example of substruction. He began with four types of social norms—folkways, mores, laws, and customs. He performed a substruction that identified three underlying dimensions of norms—"how originated," "how enforced," and "strength of group feeling." These intercorrelated dimensions were subsequently extended and combined to form a full property space.

Quantitative Typologies

Quantitative typologies differ in significant ways from their qualitative counterparts. They generally have different names, being referred to as taxa or clusters rather than as types. Quantitative types are almost exclusively derived through the analysis of empirical data and tend to be inductive. While it would be possible to construct a quantitative "heuristic" type by applying an algorithm to conceptual dimensions rather than to empirical data, this approach is rare or nonexistent. The chief goal of quantitative typologists, whether they term their efforts quantitative typology, numerical taxonomy cluster analysis, or some other label, is to take some empirical data set and group all the cases into categories that maximize the internal similarity of each group.

Q- versus R-Analysis

A major distinction in quantitative taxonomy is whether one is seeking to group objects (Q-analysis) or variables (R-analysis). The usual procedure is to begin the quantitative analysis with a basic data matrix in which the rows list data for objects, and the columns list data for variables. The basic data matrix is shown in the work of Bailey. The internal scores of the data matrix are the same whether one inspects rows or columns, and they represent the matrix of scores on each of *M* variables for the sample of *N* objects (persons, animals, plants, and so forth). In sociology and most of the other social sciences, it is customary to conduct R-analysis by correlating pairs of columns in the data matrix. This yields a set of R-correlation coefficient among variables. For example, variables 1 and 3 may exhibit a correlation of 0.67, while variables 2 and 6 have a correlation coefficient of 0.42, and so forth. In biology and related fields it is customary to use Q-correlations for numerical taxonomy and cluster analysis. The procedure is essentially the same except that correlation coefficients are computed for pairs of rows rather than pairs of columns. This yields Q-correlations between objects.

Most quantitative algorithms used in numerical taxonomy and cluster analysis are quite robust, enabling them to accommodate a variety of data. These include Q- and R-similarity coefficients for all levels of data measurement (ratio, interval, ordinal and nominal data and binary-coded data, as well as distance coefficients. However, due to degrees of freedom issues, Q- and R-analyses generally require somewhat different kinds of data sets. Specifically R-analysis is best with a large sample of objects (*N*) and a smaller set of variables (*M*). The converse is true for Q-analysis. It requires a small sample of objects (*N*) and a larger set of variables (*M*).

Thus, the fact that R-analysis predominates in sociology and Q-analysis predominates in biology is not merely

the result of whimsy or tradition. There are sound methodological reasons for this. In sociology and in related fields such as political science and policy studies, there is an increasing emphasis on the use of large national samples of individuals. If one is using a sample of 10,000 cases, it is virtually impossible to measure a larger number of variables (more than 10,000) for each of the individuals. Thus, the researcher has more objects than variables, making the analysis most appropriate for R-correlations. The converse is often true in biology or medicine, and sometimes in psychology and other fields as well. Here, it may be difficult to find more than a few rare specimens, but if binary coding is used, it may be possible to code data for a large number of characteristics (perhaps hundreds) each case. Data of this sort is conducive to the use of Q-analysis. Thus, the majority of quantitative analyses cluster objects via Q-analysis, but a few cluster variables via R-analysis.

Quantitative empirical methods generally cannot form full typologies. However, they can attempt to group the sample of specimens as well as possible on the basis of similarity. Ideally, all clusters or taxa would be tight, meaning that the individuals within each cluster were highly similar if not identical. That is, internal cluster variance or distance would be minimized. Further there will be maximum distance between clusters, so that there are no overlapping clusters, and no arbitrary decisions need to be made regarding when one cluster ends and another cluster begins.

Quantitative typologists need to make at least four main decisions. These are: the type of analysis (Q or R), whether the analysis is agglomerative or divisive (as discussed previously), what kind of coefficient to use (similarity or distance), and if an agglomerative method is chosen, the type of nucleus formation need to be decided. The basic decision is whether one chooses the most similar pair as the nucleus for one cluster, or the most dissimilar pair as the nuclei for two different clusters, or some variation of these basic strategies.

A Typology of Types

Figure 1 shows a typology of types. Row 1 shows the qualitative typologies, previously discussed. These can be divided into qualitative heuristic (Cell 1) and qualitative empirical types (Cell 2). The heuristic types are generally conceptual or theoretical in nature and are derived deductively prior to empirical investigation. These include the ideal type, the constructed type, and polar types. These may or may not have clear empirical referents. Types that are said to be purely theoretical, and to lack empirical referents, still may be valuable for heuristic purposes. However, if they are used empirically, they are vulnerable to charges of improper reification.

	Heuristic	Empirical
Qualitative	Ideal Type 1	Ethnographic Type 2
Quantitative	Rare or Nonexistent 3	Clusters or Taxa 4

Figure 1 A typology of types. From *Encyclopedia of Soc* 2E 4V, by, 5, Macmillan Library Reference, © 2000, Macmillan Library Reference. Reprinted by permission of The Gale Group.

The types of Cell 2 are inductively derived through ethnographic research. Outsider types may be imposed by external researchers or by the lay public, while insider types are generated and used by the insider participants themselves. The qualitative types of both Cell 1 and Cell 2 generally represent selected types, rather than the full typology. Thus, they can be viewed as examples of functional reduction, as described above.

Quantitatively derived types are found almost exclusively in Cell 4 (quantitative–empirical). Examples of Cell 3 (quantitative–heuristic) are rare if not nonexistent in the typological literature. About the only way they could be generated would be to take purely conceptual, nonempirical categories and process them through quantitative algorithms. The rationale for such analysis, and the value of the subsequent types, is unclear. The prospects for such activity in Cell 3 are low. In contrast, Cell 4 (quantitative empirical) is the site of a great deal of activity. A wide variety of cluster and numerical taxonomy techniques, both agglomerative and divisive, are available for constructing clusters and taxa in Cell 4. Such clusters and taxa generally do not represent a full typology, but rather can be seen as examples of pragmatic reduction. They are equivalent to the set of types that can be found by collapsing multiple contiguous cells to make a single new cell that is more parsimonious, but also more heterogeneous, than the prior types. The clusters and taxa of Cell 4 represent the maximum amount of intratype similarity that can be quantitatively achieved from the given sample of empirical cases.

Monothetic Types

The great advantage of constructing types that are purely theoretical is that they can be made entirely monothetic. Monothetic types are fully homogeneous. There is no internal variation within a monothetic type. In statistical terms, this would be represented by an example where all cases in the particular type possesses the mean value on all dimensions. Thus, the internal variance is zero.

Imagine a typology formed from 10 binary-coded characteristics, when each of the 10 is coded 1 if present and 0 if absent. The full typology will consist of 1024 cells or types. Each of these is monothetic, as a case cannot

occupy a given cell unless it possesses all of the characteristics represented by that cell. In order for a typology to be monothetic, the possession of a unique set of features must be both necessary and sufficient for identifying cases as belonging to a particular cell in the typology. Each specific characteristic is necessary, and the set alone is sufficient. Before a case can be identified with a particular type, it must possess all of the characteristics of that type, but no others. Thus, all of the cases in a given monothetic type are identical in all ways (at least in all the ways that were specified and measured).

Polythetic Types

Polythetic types are more eclectic and heterogeneous than monothetic types. While monothetic types display no within-type variance, polythetic types may display such internal variation, because not all cases within a given type are completely identical. Thus, compared to the pure internal homogeneity of monothetic types, polythetic types can be characterized as pragmatic or practical approximations to pure monotheticism. In a polythetic type, the cases are selected to display the greatest possible degree of similarity. That is, the internal variance of a polythetic type is not zero (as in a monothetic type), but is minimized. In a polythetic type, no single characteristic is either necessary or sufficient. For example, in the 10-dimensional typology comprising 1024 monothetic types, the cases in each type will all be identical on all 10 characteristics. In contrast, in a polythetic type based on the same dimensions, it might turn out that none of the pairs of cases identified for the type share all 10 characteristics, however, they might each have 9 of the 10 characteristics, but not the other one. For example, Case 1 might possess characteristics 1 through 9, while Case 2 possesses characteristics 2–10. These eight characteristics (2 through 9) are held in common by each case. However, in addition, Case 1 possesses characteristic 1 (which Case 2 lacks), while Case 2 possesses characteristic 10 (which Case 1 lacks). In a polythetic type, each specimen possesses a large number (but not necessarily all) of the characteristics, and each property is possessed by a large number of specimens (but necessarily all). A special case of the polythetic types is called fully polythetic. A type is termed full polythetic if no single property is held in common by every individual in the type.

A purely verbal quantitative type, constructed without analysis of empirical data, can be constructed to be monothetic if the full typology is constructed. The 10-dimensional typology of 1024 cells is monothetic as long as all 1024 cells are retained and the typology is not reduced. The monothetic typology is mutually exclusive and exhaustive. If a sample of empirical specimens is coded for the 10 characteristics, each specimen can be identified with some cell in the typology, although there may be many null cells (if the ample size is less than 1024, and there may be no cell with more than one specimen).

The three types of reduction discussed previously have different ramifications for monotheticism. It is assumed that the full multidimensional typology is monothetic. Functional reduction can maintain monethicism by only retaining the cells for which empirical cases are identified. Functional reduction will also ensure that the types remain mutually exclusive. Further, the reduced typology can still be considered exhaustive for the given sample. However, if a new sample is chosen for the identification of empirical specimens, the researcher may find that some type needed to identify a specimen has been deleted in the course of functional reduction. In this case, the typology remains monothetic, but it is inadequate because it is no longer exhaustive and is of no use for identifying the specimen in question. The two other types of reduction, pragmatic and arbitrary numerical, reduce the number of types by converting monothetic types into polythetic types. This is done by putting specimens in a single cell that previously would have occupied separate cells in the full monothetic typology. These forms of reduction enable the typology to remain both mutually exclusive and exhaustive, while relinquishing its monotheticism.

Returning to Fig. 1, it is clear that the qualitative types of Row 1 are monothetic. Both the heuristic types of Cell 1 and the ethnographic types of Cell 2 retain their monotheticism, even though in most cases they do not represent a full typology. In some cases, a full typology may have been constructed and then reduced, but in most cases the typology was only partially constructed, with other types remaining latent having never been constructed. The reduction is equivalent to functional reduction, as null cells are simply not included in the typology. The cells that remain are not collapsed or compromised in any way, so their full monothetic status is maintained.

In Row 2 of Fig. 1, examples of Cell 3 are virtually nonexistent. If any did exist, they would probably be polythetic. Virtually all of the examples of Cell 4 (and there are many) are polythetic. It would be possible to construct a monothetic type through cluster analysis or numerical taxonomy, but this would be a rare even using a special data set. The polythetic clusters of Cell 4 (and of Cell 3 if any existed) are the equivalent of reduced typologies produced through pragmatic or arbitrary numerical reduction. That is, they represent a reduced number of types from a full typology, with the remaining types being polythetic, and being more heterogeneous than monothetic types would have been.

To summarize Fig. 1 by rows, qualitative types (Row 1, Cells 1 and 2) are generally monothetic, while quantitative types (Row 2, Cells 3 and 4) tend to be polythetic. To summarize Fig. 1 by columns, heuristic types (Column 1,

Cells 1 and 3) tend to be deductively derived, while empirical types (Column 2, Cells 2 and 4) are inductively derived. The qualitative and quantitative typological traditions appear to be worlds apart. The qualitative traditions originated largely in sociology. They tend to be verbal typologies that may or may not have clear empirical referents. They are generally deductively derived. Types tend to be monothetic. Full typologies can be reduced through pragmatic, functional, and arbitrary numerical reduction. Full typologies can be constructed from one or a few types through substruction.

Quantitative typologies are generally empirically derived through either agglomerative or divisive methods. They originated in fields such as biology and psychology and go by names such as numerical taxonomy and cluster analysis. The types formed empirically through qualitative analysis tend to be called clusters or taxa and are generally polythetic. The processes of substruction and reduction are generally not recognized in quantitative approaches to typology construction. However, the resulting set of polythetic types is similar to a typology reduced through arbitrary numerical reduction or pragmatic reduction (but not functional reduction, in which the types of the reduced typology remain monothetic).

The Three-Level Model

An adequate understanding of typology construction and use entails understanding the relationships between qualitative and quantitative typologies. In traditional measurement terminology, the heuristic verbal type can be labeled a "mental construct" or "latent variable." The measurement process entails demonstrating a relationship between this conceptual level (qualitative type) and the empirical level (quantitative type). The correlation between the theoretical and empirical levels has been called an epistemological correlation. For example, a verbal typologist might construct the type concept of "career criminal." The quantitative typologist might group a sample of criminals via cluster analysis and judge one cluster to represent career criminals. Thus, the traditional approach envisions only two levels, the conceptual or theoretical level and the concrete or empirical level, and the epistemic correlation which measures how well the empirical cluster represents the underlying theoretical construct (heuristic type).

In reality, there are three levels, not two. These are the conceptual or theoretical level, the empirical level, and the documentary or index level. One can envision the 10-dimensional typology of 1024 monothetic cells, with each of the 10 dimensions coded in binary fashion. Thus, full typology is mutually exclusive and exhaustive. It is also a relatively easy task to search for empirical examples of each monothetic type. One might find 5000 empirical specimens and determine that examples of all 1024 cells can be identified from these. This process of first constructing multidimensional monothetic conceptual types and then identifying empirical examples of each is clearly a form of measurement. The type concept (e.g., student athlete) is constructed, an empirical specimen is measured on all relevant dimensions, and the specimen identified as fitting that type.

The problem comes when one seeks to represent the typology and its empirical examples in a documentary dimension that renders the typology amenable to dissemination and interpretation. An example of this documentary or index level would be any medium, such as a printed page (the most common), a computer screen, or a computer disk. As noted earlier, the 10-dimensional type cannot be properly represented on the two-dimensional page. This causes major problems. A typologist can envision the 1024 cell multidimensional full typology, and envision the process of identifying empirical examples. This can all be envisioned without any representation on paper, film, computer, etc. But portraying this on a printed page so that the information can be stored, interpreted, analyzed, and transmitted to others is currently impossible.

However, a small full typology such as the fourfold typology of Fig. 1 can be represented on paper (as long as it remains in one or two dimensions), and so it can be used to illustrate the three levels. The classical method of typology construction and use pioneered by Weber involved three successive stages. The first stage was for the researcher to conceive a mental image of the ideal type (e.g., "All American Boy" in his or her mind). This is Level A, the conceptual or mental level. The second stage was to write the designation of the type on paper. This is Level B, the documentary or index level. The third stage was to look for an example of a living person who fit the type concept. This is Level C, the empirical level. Qualitative typologists start with Level A and next proceed to Level B, but they often do not get to Level C and do not find empirical examples of the type. Quantitative typologists generally proceed in reverse. They begin with Level C (the empirical level) and construct polythetic clusters of objects that are printed on paper (Level B). They often do not proceed to Level A (conceptual).

Without using the full three-level model, it is difficult to clearly understand the typological process. Unfortunately, the full three-level model is rarely recognized in the measurement literature. The tendency is to present only two "measurement" levels. These dichotomies have many names, such as the latent variable and the concrete variable, concept and index, concept and empirical level, and theoretical and concrete levels. All of these labels fail to recognize the three-level model. More tragically, they all represent some particular conflation of Levels A, B, and C into only two levels. This is generally done unwittingly, without recognition (or even knowledge) of the three levels.

Thus, one researcher may present a mental concept (Level A) and an empirical example (Level C) as a "two-level measurement model," while another researcher presents Levels B and C as purportedly the same model, or another researcher presents A and B erroneously as the same model. Until these are recognized as three distinct levels (instead of simply alternative presentations of the same two levels) in the measurement literature, understanding of the complex typological process will remain flawed.

Advantages and Disadvantages of Typologies

Advantages

Typologies have a number of distinct advantages for research and measurement, and in fact are a central feature of those endeavors.

1. Description. The typology is the premier descriptive tool. It is the cornerstone of any discipline, as it provides the core set of descriptive, multidimensional types or taxa.
2. Exhaustiveness. No other research technique provides the comprehensiveness of a full typology. The typology represents the definitive reference source for a discipline.
3. Multidimensional complexity. No other concept or presentation can match the complexity and conceptual range of a mutidimensional type. Unidimensional descriptions are simply no alternative for the mutidimensional type.
4. Clarity. A rigorous explication of a multidimensional type exhibits a degree of clarity and absence of ambiguity that is badly needed in research.
5. Comparison. From the single ideal type to the full typology, the typology is the premier tool for the rigorous multdimensional comparison and analysis of both conceptual and empirical types. The comparative procedure is very parsimonious, as it allows one to only identify the types for which empirical cases exist. Other potential types can remain latent and unused as long as they are not needed, but still are available if needed.
6. Differences. Typologies, especially full typologies and polar types, are useful for illustrating differences among two or more empirical cases.
7. Identification of empirical cases. Typologies are the ultimate tool for identifying empirical examples of particular type concepts.
8. Illustration of Possibilities. A full typology allows one to illustrate possible types, even if they cannot be found empirically.
9. Reduction of complexity. Type concepts such as the ideal and polar types, along with the reduction processes of arbitrary numerical reduction, pragmatic reduction, and functional reduction, are excellent means of reducing complexity to manageable levels.
10. Theoretical explanation. Devices such as heuristic types, including ideal, constructed, and polar types, as well as the process of substruction, are excellent tools for facilitating theoretical illustration and explication.

Disadvantages

1. Unmanageability. Typologies can on occasion be so unwieldy and large that they are difficult to utilize efficiently.
2. Difficulties in interpretation. Even typologies that are not overly complex, such as the ideal type, can sometimes be difficult to interpret. The ideal type became immersed in controversy regarding its proper usage and interpretation because some critics found it difficult to understand or to use effectively.
3. Little explanatory or predictive power. Typologies have often been faulted as being overly description, with limited explanatory or predictive value. Some critics are wont to eschew typologies in favor of techniques such as multiple regression that are considered to have more explanatory power.
4. Noncausal. A related criticism is that typologies say little if anything about cause, mainly presenting a set of correlated objects or attributes.
5. Outmoded. Heuristic typology construction, most notably in the form of the ideal type, is associated with 19th and early 20th century social science. As such, it is considered to be obsolete by many, particularly in the age of computerization. Modern methods of cluster analysis and numerical taxonomy certainly cannot be regarded as obsolete, but many social scientists are not familiar with them and do not understand their relation to classical typological methods.
6. Primitive. Despite their underlying complexity, some classical methods (such as the ideal type), may be viewed by some critics as primitive and simplistic.
7. Lack of integration. Critics that are unfamiliar with the three-level model may lack a framework to integrate qualitative and quantitative approaches to typology construction. Such critics may view the field of typology construction as a whole as eclectic and not properly integrated.

Typology Construction as Measurement

The transparently descriptive nature of the typological endeavor masks its essential nature as a measurement tool. There are at least five chief factors that hinder recognition of the typology as a measurement tool. One is the claim by heuristic typologists that classical types such as the ideal type often do not have empirical

referents. A second is the descriptive nature of the typology. A third is the apparent limitation of clusters and taxa to the empirical level. A fourth is the obfuscation of the three-level model through its frequent conflation into a dichotomy. A fifth is use of Q-typology (objects) rather than R-analysis (variables).

A complete typological cycle involves three stages, and it can begin with either Level A or Level C (but not with Level B). The classical qualitative approach begins with the genesis of a conceptual type (Level A). The second stage is the mapping of the full typology onto the document level (Level B). This can be accomplished, for example, by writing the full typology on paper (even if this must be done two dimensions at a time). The third stage is the identification of empirical cases (Level C) for each respective type. Modern quantitative methods such as cluster analysis and numerical taxonomy repeat the process in the opposite direction. They begin with empirical analysis (Level C). The second stage is to represent the cluster on paper (Level B). The third stage is to envision a mental concept of the cluster (Level A), often by mentally generating a name or label for the cluster (such as "career criminal"). Whether one begins with Level A (qualitative approach) or Level C (quantitative approach), the process unfortunately remains incomplete in actual analyses. Generally, Level B (the second stage) is reached, either from Level A or from Level C. However, verbal typologists who begin with Level A generally reach Level B and write their types on paper. However, they often do not reach Level C (they do not identify empirical cases for the types). Similarly, quantitative researchers who begin with Level C also reach Level B, but may not reach Level A (may not form mental images of, or names for, the empirical clusters or taxa).

No matter whether the researcher begins with Level A or Level C, when the whole process is complete, empirical types are connected with their conceptual-type counterparts, in a quintessential measurement operation. Critics often fail to recognize that this is measurement simply because they are accustomed to measurement conducted in terms of R-analysis rather than Q-analysis. For example, contemporary measurement procedures generally operated solely in terms of R-analysis. The researcher might conceive of a multivariate concept such as "alienation" and then construct a multiple-item alienation scale to measure this characteristic empirically.

Typologists would have essentially the same goals, but would approach the problem in a different (but parallel) manner. The classical typologist might conceive of the multidimensional heuristic type of "alienated intellectual," and then would next seek to identical empirical examples of the real-life alienated intellectual person. The "alienated intellectual" must be recognized as a Q-type, as the basic unit is the object (person) rather than the variable. Further, the traditional typologist might not attempt the quantitative measurement of all the variables making up the alienation concept. However, in order to identify empirical cases of the type, all of the variables must be measured in some manner. Similarly, the quantitative cluster analyst or numerical taxonomist, while making more precise quantitative measurements than the classical typologist, and proceeding from the other direction (by beginning with the empirical level rather than the conceptual level), is also engaged in a measurement process. Once the polythetic empirical type is constructed, one can then seek a label for it (for example, the "alienated intellectual"), thus completing the measurement process in reverse.

See Also the Following Articles

Clustering • Lazarsfeld, Paul • Weber, Max

Further Reading

Bailey, K. (1972). Polythetic reduction of monothetic property space. In *Sociological Methodology 1972* (H. Costner, ed.), pp. 82–111. Jossey–Bass, San Francisco.

Bailey, K. (1973). Monothetic and polythetic typologies and their relationship to conceptualization, measurement, and scaling. *Am. Sociol. Rev.* **38,** 18–33.

Bailey, K. (1974). Cluster analysis. In *Sociological Methodology 1975* (D. Heise, ed.), pp. 59–128. Jossey–Bass, San Francisco.

Bailey, K. (1984). A three-level measurement model. *Quality Quantity* **18,** 225–245.

Bailey, K. (1986). Philosophical foundations of social measurement: A note on the three-level model. *Quality Quantity* **20,** 327–337.

Bailey, K. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques.* Sage, Thousand Oaks, CA.

Bailey, K. (2000). Typologies. In *The Encyclopedia of Sociology* (E. Borgatta and R. Montgomery, eds.), 2nd Ed., Vol. 5, pp. 3180–3189. MacMillan, New York.

Barton, A. (1955). The concept of property space. In *The Language of Social Research* (P. Lazarsfeld and M. Rosenberg, eds.), pp. 40–53. Free Press, New York.

Lazarsfeld, P. (1937). Some remarks on the typological procedures in social research. *Z. Sozialforsch.* **6,** 119–139.

McKinney, J. (1966). *Constructive Typology and Social Theory.* Appleton, New York.

Sneath, P., and Sokal, R. (1973). *Numerical Taxonomy.* Freeman, San Francisco.

Spradley, J. (1970). *You Owe Yourself a Drink: An Ethnography of Urban Nomads.* Little, Brown, Boston.

Tiryakian, E. (1968). Typologies. In *International Encyclopedia of the Social Sciences* (D. Sills, ed.), pp. 177–186. MacMillan/Free Press, New York.

Weber, M. (1947). *Theory of Social and Economic Organization,* (A. Henderson and T. Parsons transl.; T. Parsons, ed.). Free Press, New York.

Winch, R. (1947). Heuristic and empirical typologies: A job for factor analysis. *Am. Sociol. Rev.* **12,** 68–75.

Units of Analysis

Victor C. de Munck
State University of New York at New Paltz, New Paltz, New York, USA

Glossary

attributionist fallacy Confusing a behavior for a trait or attribute of a person or group.
ecological fallacy The error of interpreting variations in environmental settings as variations among individuals.
individual differences fallacy The error of interpreting individual characteristics as group characteristics.
large-scale collectivity Groups that have fuzzy boundaries and that cannot be observed as wholes but whose members share a common social identity. These collectivities are imagined communities.
small-scale collectivity Clearly bounded and observable social units consisting of a minimum of 2 and seldom more than 1000 members.
supracommunal level Two or more levels of institutionalized authority are considered a supracommunal collectivity.
unit A person, element, or some discrete indivisible property that can be treated as an entity and can therefore be measured.

Scant attention has been paid in the social sciences to the problem of defining units of analysis. Instead, the methodological lens has been aimed at describing, measuring, and analyzing variables. This article describes the problems that occur when the researcher neglects to clearly define the units of analysis and how to avoid them. Units of and for analysis are always entities, whereas variables refer to the attributes, events, or processes that are part of or impact on entities. For any research question it is necessary to clearly define the unit as well as the variables of the study. Ignoring this distinction leads to either the ecological or the individual fallacy. Both of these problems and their resolution are addressed. This is followed by an examination of defining units for cross-cultural research. Cross-cultural research requires that the units of comparison are independent of one another; otherwise, similarities found may be a result of diffusion rather than of independent origin. It is demonstrated that this problem can only be resolved on a case-by-case rather than a systematic basis. A checklist for determining the appropriate units of analysis for any social research project is provided.

Introduction

This article provides a detailed discussion of the meaning, uses, and limitations of the concept "unit" in the social sciences. There is some disagreement about how to define unit and what kinds of things can and cannot be classified as units. The first two sections of this article provide a definition and typology of social units. The following sections consider the three most obdurate and enduring problems associated with the unit concept: (i) the ecological fallacy—that is, conflating or confusing situational constraints with personality; (ii) the particulate-systemic or attributionist fallacy—that is, confusing a behavior for a trait or attribute of a person or group; and (iii) the problem of determining the independence of units. This latter problem has been dealt with most comprehensively in cross-cultural studies in which it is known as "Galton's problem." The final section offers a checklist for researchers in helping them decide what cultural units to choose for their research and the limitations and advantages for analysis of these different kinds of units.

Defining Unit

A unit is a person, element, or some discrete indivisible property that can be treated as an entity and can therefore be measured. Variables are never units. A unit contains or

expresses the variable that is under study. A unit may be an individual or a collectivity. What distinguishes a unit from a variable is that a unit is a discrete thing in itself. Time and space, as in minutes or inches, are units of measurement and not of analysis.

A unit of analysis is always treated as if it were an entity. If one is interested in the height of sixth-grade boys and girls, then each individual boy and girl is a unit of analysis. In the statement "he is six-feet-five," "he" is the unit of analysis and "six-feet-five" is an attribute of that unit, whereas feet and inches are the units of measurement. The individual is a discrete and indivisible thing that has the property of being "six-feet-five." Discussing a property as if it is an entity is called the attributionist fallacy, which is discussed later. In short, a unit is always an entity from which measurements are taken.

In a typical social science two-mode data matrix, rows identify cases and columns identify variables. In two-mode matrices, cases are usually equivalent to units and the column labels identify the variables in the study. A cell is the intersect of a unit and a variable. The value in any cell is a measure of the amount of that variable associated with that unit (Fig. 1). Social scientists usually analyze the cell value as a variable rather than as a social or cultural unit value.

Researchers use a two-mode matrix to test hypotheses about the relations between variables. The unit itself is seldom considered in the analysis because "variable" and "unit" are typically perceived to be in a figure–ground relationship with the units as the ground and the variables as the figure(s). The analytic lens is aimed at the variables rather than the units (i.e., the individuals). The units are figuratively drawn and quartered with only their salient parts incorporated into the analysis. For example, one can compare suicide rates across time, space (e.g., a rural–urban dimension), cultures, or religions without ever considering the units (i.e., the people committing suicide) as individuals. The individuals who ultimately comprise the units of analysis are extricated from the study except as they provide anecdotal material.

An example of a research strategy that retains the individuals as units of analysis and measurement is the 1995 study, "Environmental Values in American Culture" by Kempton *et al*. The researchers compared the environmental values of a sample of sawmill workers, dry cleaning managers, the general public, Sierra Club members, and members of Earth First! The study focused on the response profiles of all the individuals and compared these profiles both within and between groups. The researchers found surprisingly high agreement among the members of these groups. In this study, the individual was the unit of analysis and the unit of measurement, and the statistics and charts depicted the relationship between individuals rather than between variables. This study relied on consensus analysis (a relatively new method introduced in 1986 by Romney *et al*.) to measure the aggregate level of agreement across all variables for each individual, thus figuring the unit of analysis and grounding the individual variables. However, such studies are rare and most foreground the variables rather than the units. The following discussion of the different kinds of units considers the former rather than the latter situation. A problem with using the individual *in toto*, as both analysis and the unit of measurement, is that the profile variables may range from nominal to ratio and vary considerably in how they are operationalized and whether they are independent of one another. For example, measures of social class often aggregate different types of variables, such as income, religion, prestige, and education, as if they can all be measured by the same scale. This crude amalgamation of different types of variables to construct and then profile social classes both mystifies class and "washes out the actual grid of causal processes that distribute people across several dimensions of the social landscape" (Collins, 1990, p. 48). On the other hand, such mixing can lead to greater generalizability if it takes into account the limitations and difficulties of making summarizing inferences from an array of disparate variables.

Units

Variables

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Abel					
Beth					
Carl					
Don					

Figure 1 A two-mode matrix with the cases as the units of analysis.

Kinds of Units

The canonical taxonomy for types of variables was first presented by Stevens in 1946. Stevens described nominal, ordinal, interval, and ratio variables. Although there is debate over the utility and appropriateness of this classification scheme, it is found in all introductory statistics texts and most texts on research methods. A typology for units of analysis, although heretofore lacking, is presented in Fig. 2. This typology is, of course, open, and more types of units can be added onto it. Its purpose is to provide an initial taxonomy that is subject to further development. A discussion of the typology follows.

The individual is the most common unit of analysis. Studies on personality traits, agent-based analysis, decision making, and life histories are examples of individuals as the units of analysis. When data on individuals are aggregated and the analysis is based on the aggregate

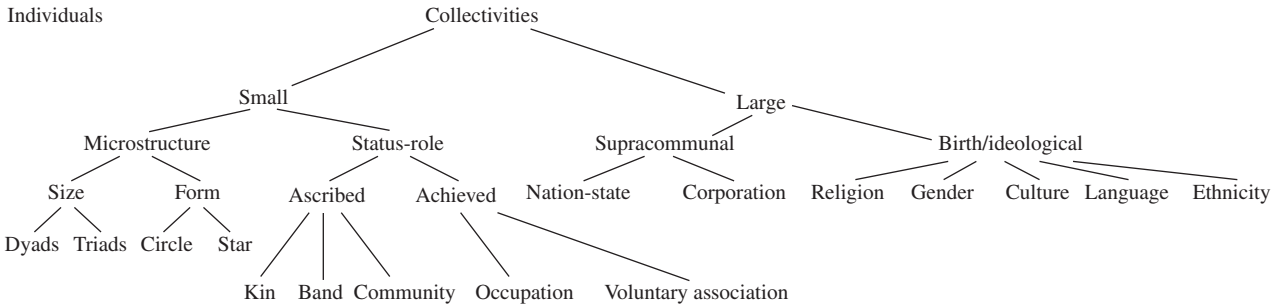


Figure 2 Units of analysis.

profiles, the collective is the unit of analysis. At a minimum, a collective consists of two or more individuals who recognize their common identity. The term *collectivity* is favored over *group* because the criterion "interaction" is often a defining feature of "groupness." Individuals can recognize their common identity on the basis of religion, gender, ethnicity, common language, culture, or nationality without interacting. Durkheim, Toenneis, Weber, and Benedict are among the many social theorists who have thought to distinguish the social structural and psychological differences between small and large collectivities. Relations in small collectivities are multiplex (i.e., many connections) and "face-to-face" with interactions characterized by intimacy or *gemeinschaft* (i.e., informality). Bonds in large collectivities are usually singleplex, relationships are imagined rather than actual, and interactions when they occur are characterized by formality or *gesellschaft*.

Small collectivities are often defined in terms of their structural components or their social function. Structural analyses of small groups use either the size of the group as the unit of analysis or the form of the group. Both Heider and Simmel were interested in the inherent sociological properties of dyads; both Simmel and Granovetter were interested in triadic relations. The study of dyads and triads is a vital cottage industry in the subdisciplines of social networks (in sociology) and personal relationships (in psychology). In these studies, it is the number of individuals that comprise a group that is the unit of analysis. For example, Simmel noted that one property of the dyad is that it is the only social group in which the individual has the option of terminating the group. Granovetter points out that a property of triadic relationships is that their stability rests on the mutual affinity of each member for the other. Group size, not its shape or its members, comprises the unit for analysis. Collectivities such as dyads and triads are studied as if they were single entities with the relationship of properties to collectivity analogous to that of traits and the physiological properties that pertain to an individual.

In 1950, Alex Bavelas initiated a series of experimental studies on the shape of small groups. He and his colleagues wanted to investigate the communicative efficiency of different social forms. By efficiency, he meant a combination of accuracy and speed in solving problems. Bavelas and associates discovered that "star-shaped" structures were the most efficient at solving problems in a laboratory setting, whereas "circle-shaped" structures were the least efficient. In a posttest survey, he discovered that those subjects who were placed in the circle structures were much happier participating in the experiment than those who were placed in the star structures (with the exception of the individual at the center of the star). The reason for these responses is based on emergent behavioral and emotional properties of the respective groups. The center person in a star-shaped structure received all the information from the people at the end of each "ray" or spoke. The subjects at the end of each ray were not in communication with each other and were completely dependent on the central person to make a decision for the group. Each would pass their bit of information to the central person, who then had complete information to solve the problem that was posed to the group. In circle-shaped structures, the information is distributed equally and the group seeks to make its decision via consensus, which is time-consuming, may be faulty, but necessitates the input of each participant in the decision-making process. The Bavelas research project was very successful in isolating group-level characteristics and analyzing their effects on communication. One can imagine many other types of social forms, as indeed Bavelas and many others have done, but most other forms of small groups lie somewhere between star and circle structures. In the growing field of network analysis (largely spurred by Bavelas's findings), much research has been conducted on cliques (where each member of a group is connected to everyone else), the directionality and the strength or valence of ties, and informal as well as formal networks. In these studies, the unit of analysis is not the individual or group but the pattern of relations.

Another division under "small collectivities" refers to those groups that are organized by status–role relations. Statuses are distributed through a bounded group creating a rich network of reciprocal privileges and obligations.

Roles are understood to be the rules for the appropriate engagement of a status. Thus, a status set is a system of reciprocal rights and duties. Ascribed statuses are acquired by birth or are assigned by convention. Two examples of an ascribed status acquired by convention are age–grades and kin statuses. Achieved statuses are those that are acquired by choice and through a social decree. In contrast to a social convention, a social decree refers to a claim for a status that is then legitimated by a referent group. A few statuses, such as adopted kin status, may be considered both achieved and ascribed, but the majority clearly belong to one or the other category. Research that relies on status–role as the unit of analysis typically examines (i) the qualifications of the position, (ii) the distribution of rights and duties of the position, (iii) the activities spawned by those rights and duties, and (iv) the settings associated with the status–roles.

Large collectivities are divided into supracommunal and birth/ideological types. Any collectivity that contains two or more levels of institutionalized authority is considered a supracommunal collectivity. For example, in a chiefdom there are local chiefs who have authority over the commoners in their locale and paramount chiefs who have authority over the local chiefs. A nation-state has, at a minimum, three supracommunal levels: national, regional, and local seats of authority. The nation-state is the unit of analysis in world systems theory, in which the transnational flow and exchange of natural resources, as well as humans, between core and peripheral nation-states is the subject of study. Supracommunal levels were used as the units of analysis by Elmond Service in 1962, who placed cultures into a cultural evolutionary scheme based on sociopolitical levels of integration. Bands and tribes were the simplest societies, with no or one supracommunal level; chiefdoms had two supracommunal levels; and states consisted of a minimum of three such levels. Corporations also have hierarchically nested seats of authority. Corporations are units of analysis in studies on corporate organization, the production and flow of goods and capital, and in analyses of the stock market.

Collective identities predicated on birth and/or ideology include culture, sex, language, and ethnicity. Sex is a unit of analysis in evolutionary psychology, in which it is assumed that sexual differences promote different mating strategies and work activities. Expendability theory, for example, suggests that in foraging societies the reproductive value of women was far greater than that of men and thus women were protected and kept near camps while men took on the high-risk activities such as big-game hunting. Sex is a unit of analysis when social differences are seen to be a result of inherent psychobiological differences between males and females. When these differences are evaluated in terms of gender, culture becomes the unit of analysis. Cross-cultural studies have shown that patriarchy is greater in Islamic countries and communities than in Hindu or Christian countries or communities. In this kind of study, religion is the unit of analysis, culture is held constant, and the properties of the religion are the variables. Religion and gender are cultural constructs and cannot be units of analysis in ethnographic studies. That is, culture is always the unit of analysis when the subject of study is a subsystem of a culture. Religion, gender, ethnicity, and other macrocollectivities are units of analysis only when they are being studied cross-culturally. In these instances, these units are treated as megacultures.

Ruth Benedict's *Patterns of Culture*, first published in 1934, laid the groundwork for the study of national cultures and the theoretical predilection of researchers to reify culture. She explicitly argued that cultures were not merely a collection of traits but were "like individuals," forming a more or less integrated pattern of emotions, thoughts, and actions. In reifying culture, Benedict promoted the use of culture and other social constructs as units rather than objects of analysis. Bellah *et al*. repopularized the notion of culture as an entity in their 1986 book, *Habits of the Heart*, and its 1992 successor, *The Good Society*.

The typology of social units of analysis presented previously will undoubtedly be modified over time. However its two primary nodes—the individual and the collectivity—must be the core divisions of the typology. Although it is true that collectivities are composed of individuals, it is not true that the properties of collectivities are identical with those of individuals. The following sections take particular issue with problems that result from confounding collectivities with individuals and individuals with collectivities.

The Ecological Fallacy

In 1950, Robinson coined the term ecological fallacy to refer to the error of interpreting variations in environmental settings as variations among individuals. One tactic for solving Robinson's ecological fallacy is to construct surveys in which questions clearly state whether they are asking personal opinions of the subject or general assessments of an environment setting. A Likert scale example of an ecological (i.e., environmental) question is to ask respondents to agree or disagree with the comment, "Sometimes class is very disorganized." A comparable example in which the individual is the unit of analysis is to ask respondents to agree or disagree with the comment, "Sometimes I am not prepared when I come to class." The ecological question provides a generalized assessment of the environment without targeting the source of disorganization. In the ecological example, it is unclear as to what unit of analysis the

subjects are responding to—the setting, the teacher, the other students, themselves, or all of these.

Richards and colleagues compared the use of individualist and ecological units to analyze classroom environments. They used the Classroom Environment Scales developed by Moos and Trickett, which consist of true–false questions about the classroom environment. Richards *et al*. (1991) noted that the questions were "modeled on and resemble the type of questions used in objective personality tests" (p. 425). Consequently, measures of dispersion (such as standard deviation) were much higher among individuals in settings than across settings and reliability measures (alpha) were also higher across than within settings. Richards *et al*. also suggested that assessments of setting measures were mediated by personality differences between the individuals and that this confounded the results within any one setting. Thus, survey questions should be crafted so that they distinguish and elicit assessments of the environmental setting rather than serve as "disguised measures of individual differences."

Richards and colleagues use the terms ecology and settings interchangeably. However, it should be remembered that, strictly speaking, the setting is not the unit of analysis but the group that inhabits the setting. The actual classroom does not fill out a questionnaire, students do. The Richards *et al*. study is important because it unequivocally confirms that by themselves, and without a theoretical justification, individuals as the unit of analysis are invalid and unreliable units by which to measure setting-level characteristics. It should be noted that by "setting," Richards *et al*. are referring to the small-scale groups that inhabit the setting and thus setting is a group-level unit. If the goal of the study is to understand the characteristics and dynamics of settings (in this case, the classroom), then the proper sample for the study is settings and not individuals, and the goal of the researcher is to examine variation between settings and not between individuals.

In 1997, Gary King proposed a statistical solution to the ecological inference problem. Leo Goodman had previously proposed an ecological regression model to estimate individual differences from census data. King added to Goodman's model by using random coefficients to further minimize the aggregation bias. His solution has met with partial success in finding estimators of subpopulations within a larger population. However, although statistical sampling is a powerful tool, statistics is not good at low-level inferences—that is, reducing the whole to its components, a kind of reverse statistics.

The ecological fallacy is the error of attributing the characteristics of a population to an individual. Statistical inference is intended to generalize from a sample population to the whole population. The goal of statistics is to generalize from the particular to the whole and not from the whole to the particular. As such, statistics cannot offer a solution to the ecological fallacy. Data on individuals or on subpopulations within a larger population can best be obtained by ensuring that the unit of analysis is the individual or the subpopulation and not the larger population. As Richards and colleagues note, this problem can be avoided by designing survey instruments that elicit individual characteristics and attitudes. It is only from individualistic data that the researcher can track individual and subpopulation characteristics when necessary.

In a study in which local and individual hospitalization rates were derived from community-level estimates of various indicators of socioeconomic status (SES), Hofer noted that SES community profiles may not be representative of those individuals in the community who are actually going to the hospital. For example, it is known that the proportion of elderly who have medical coverage is far greater than it is for young adults, and that some of these elderly patients will use the hospital many times. To obtain accurate estimates of the subpopulations using and not using the hospital, it is necessary to obtain data on samples of individuals, not social aggregates. The best aggregate estimator of subpopulation or individual differences is to either ensure that the individual characteristics to be analyzed are representative of the aggregate or to use complete analytical models that target only that set of SES data pertinent to a target population. In their study on hospitalization rates, Billings *et al*. found it necessary to include age and income interactions in assessing SES variables in small area studies.

Although ecological (groups) units comprise individuals, their characteristics are not equivalent to those of the individuals in the group; therefore, one has to apply a different theory to studies that use collectivities as units of analysis than to studies that use the individual as the unit of analysis. When collectivities are the units of analysis, the proper subject of inquiry should be the overall characteristics and emergent properties of populations. Group-level characteristics may be very different from those of the individual members of the group. Ethnographic and psychological studies are frequently guilty of the opposite of the ecological fallacy: the fallacy of mapping individual characteristics onto a group. This problem, called the "individual differences fallacy" by Richards in 1990, is discussed next.

The Individual Differences Fallacy

The individual differences fallacy occurs when the individual is used as the unit of analysis in order to investigate and describe the characteristics and behaviors of a collectivity. This error will be examined by discussing

the individual differences fallacy in studies of small- and large-scale collectivities. The appropriate stratagems for resolving this fallacy are substantively different for small- and large-scale collectivities. Small-scale collectivities are usually discrete molecular units; large-scale collectivities have indeterminate or "fuzzy" boundaries and are always "imagined" rather than actual. For small-scale collectivities, it is often possible to observe, survey, and collect qualitative and quantitative data from all the members of the collectivity; this is never possible for large-scale collectivities. Hence, the individual differences fallacy is of greater magnitude when dealing with large-scale collectivities, such as gender or culture, than with small-scale collectivities, such as classrooms and juries.

One example of this problem can be illustrated by a study of jury behavior by Kerwin and Shaffer. Verdicts are group-level decisions, so studies of juries should be based on a theory of group-level characteristics and dynamics and use the group as the unit of analysis. Kerwin and Shaffer committed the individual differences fallacy by assuming that central tendency measures of jury member characteristics as elicited through survey instruments mirrored the characteristics of the jury as a whole. They committed what Galtung (1967) called the "fallacy of the wrong level" (p. 45), taking for granted that there is a direct correspondence between aggregated individual and group characteristics.

In their study of a mock trial, juries were categorized as "dogmatic" or "nondogmatic" on the basis of their mean dogmatism score obtained through a survey. Statistically, a jury could fall into the nondogmatic column if five people graded average on the dogma index and one graded unusually low on this index. Kerwin and Shaffer assume that individual levels of dogmatism cause the jury, as an entity, to be dogmatic or nondogmatic.

The authors of the study also used analysis of variance inappropriately because it can only be used to analyze nominal-level data under specific conditions that were not met. In studies of small groups, researchers should rely on groups as the unit of analysis unless they are interested in the effect of the group on individual behaviors. When using groups as the unit of analysis, the researcher should consider whether group characteristics are independent of, derived from, or adequately represented by aggregate statistics of the individual members of the group. There are three stratagems a researcher can employ to develop a theory of group-level behaviors: (i) The group's behaviors and characteristics are completely independent of the individuals who comprise the group, (ii) group characteristics and behaviors reflect the overall statistical properties of those individuals, and (iii) the behaviors and characteristics of the individuals are by-products of the behaviors and characteristics of the group. It is probable that each of these approaches is valid for different types of groups. It is recommended that for small-scale studies concerned with group characteristics, the researcher should pretest each of these three stratagems to determine which best suits his or her needs.

The individual differences fallacy is magnified when the unit of analysis is a large collectivity such as culture. Culture is the central organizing concept for the major subfields of anthropology: sociocultural, linguistic, physical, and archeology. The core defining attributes of culture are that it is shared, learned, and holistic. No other social/behavioral science discipline has made holism such a central tenet of study as has anthropology. Psychologists may study memory, personality, perception, and so forth, but few are likely to claim that they are studying or describing the entire human psyche in the same way that an ethnographer presumes he or she is giving a description of a whole culture, more or less, even if this claim is unstated.

Unlike the individual or a jury, culture is, at best, a slippery unit of analysis because it lacks what Campbell in 1958 referred to as "entitativety"—a collection of material and/or mental things that "interact strongly, have a common fate, and resist dispersion." Although anthropologists customarily treat culture as a holistic unit, culture seems to be composed of a more or less random collection of things, some of which are only loosely connected. For example, things such as mousetraps, computers, and fast-food restaurants seem to have little in common with each other, much less with the various values, beliefs, and behavioral repertoires that also constitute culture.

Culture is not a unit of analysis like a jury is a unit of analysis. It is also a more ambiguous unit of analysis than religion, ethnicity, or gender—units that are possible to identify and define. Culture is a heterogeneous, not homogeneous, unit, and not only is it composed of different kinds of things but also it is understood at different levels of abstraction. At the macrolevel, it may be understood as a seamless weaving together of values, beliefs, and behaviors, much as Ruth Benedict in her typology of cultural patterns as personality writ large. At the mesolevel, culture may be construed as a set of interdependent but distinct functional systems. This is the most common understanding of culture and is reflected in ethnographies in which the material is divided into religious, kinship, political, economic, and other areas. At the microlevel, cultures are often depicted as token events, social interactions, and actors.

These different levels of abstraction are seldom, if ever, underpinned by, or justified in terms of, a theory of units. Anthropologists move with intellectual innocence from one unit of analysis and level of explanation to another. Given that culture is shared, the anthropologist is partially justified in presenting ritual, customary events, and profiles of individuals as token types, representative of the culture as a whole. Partial assurance of the validity

and reliability of ethnographic representations of a culture, in the absence of any attempt to address the cultural units dilemmas that inevitably arise, comes from the length of the ethnographer's stay in the field and a comparison with other ethnographies written about the same cultural area. Nonetheless, Barrett is correct in using the label "no name anthropology" to describe most of the work done in anthropology.

Cultural theorists have dealt with the problem of cultural units by considering symbols or cognitive structures to be the basic units of people. Clifford Geertz has been most influential in developing the notion of symbols as the units of culture. Humans are symbol-generating animals and culture consists of symbolic interactions and interpretations. Cultural symbols are public in that members of a culture know them and know that all other members of the culture also know and use these symbols in semantically appropriate ways. Geertz is the founding father of the interpretivist school of cultural theory, and he has influenced constructionist, poststructural, subaltern and hermeneutic theorists. What these schools have in common is the notion that culture is a text consisting of symbols that are publicly accessible and knowable, but these symbols are open to multiple interpretations depending on the position of the "reader."

The Geertzian position solves the individual fallacy problem since it does not view culture in terms of animate entities but in terms of symbolic entities, or "vehicles" as Geertz called them. However, this problem has been replaced by Whitehead's "fallacy of misplaced concreteness," whereby mental phenomena such as symbols are treated as if they were entities. This may work if we assume that these entities are in the minds of individuals and exist as a neural network, but interpretivists as constructionists are emphatic that symbols are not to be studied in the mind but in public arenas. It is also not true that the symbol can be the unit of analysis because symbols differ in meaning and levels of abstraction, all symbols are different, and we cannot study the categorical label "symbol" because, like all labels, it designates a category of things and does not, in itself, possess the property of "thingness." The Geertzian position ultimately exacerbates the fallacy problem because it is impossible to identify a unit of analysis from this theoretical perspective.

Metaphorically perceiving culture as text leads logically to the idea of culture being constituted of loosely or independent modules or systems of knowledge. This idea, that there are no grand cultures but only cultural modules, pares down the possibility of creating a cultural unit that is of manageable size. In 1956, Goodenough defined culture as knowing how to behave in a normative or appropriate manner in any given situation. This proposal is central to the ethnoscience and ethnomethodological schools of cultural theory. From this perspective, the researcher took a particular task, behavior, concept, or setting in a culture and found out what one needed to know in order to understand the concept or act in a setting in the same way as a native. Culture here is viewed as a series of recipe-like books that describe the lawlike regularities that govern behavior and knowledge in a particular setting. The position is similar to that of the interpretivists except that the symbol systems are located in the mind of the individuals, and the set of symbols to study are specific to a particular cultural target. The unit of analysis is not the mind or cognitive processes, as some ethnoscientists insist, but the individual from whom information about cultural models or schemas is elicited.

This theoretical position resolves the ecological–individual differences fallacy by using individuals as the unit of analysis and by investigating regularities in their thoughts, emotions, and behavioral repertoires. The study of culture is invested in the study of individuals to discover normative knowledge and behavioral clusters that are specific to a particular sociocultural domain, such as illness, attending funerals, and ideas of success. The limitation of this approach is that there is no theory of collective behaviors and characteristics except as these are normatively construed by individuals. Such a theory cannot provide an understanding of the characteristics of social units as social units because collective units are not the units of analysis.

The Independence of Units

In order to conduct statistical analysis, the units of analysis must be independent of one another. Random sampling and care that survey questions are not mere replicas of one another are the primary means to ensure independence between units and variables. However, culture is shared, and the members of a culture share common cultural experiences. Statistical analysis can be employed only when we distinguish between individuals or groups within a culture. Thus, we can compare across gender, ethnicity, age groups, regions, and the like, assuming that these differences are significant enough to ensure the independence of units. In cross-cultural studies, we can assume that cultures are independent units. However, what if they are not? What if many of the cultures in a sample have had extensive historical contact with one another and some have not? The cultures that have had extensive contact may unduly bias our results so that we are guilty of a type I error; that is, any statistical test will show that our results are significant when in fact they are not. In cross-cultural research this is known as "Galton's problem."

In 1889, Francis Galton attended Sir Edward Tylor's presentation of what appears to be the very first cross-cultural study relying on statistical methods. Galton questioned the results by suggesting that many of Tylor's

cases were replicas of one another because many of the cultures included in the sample had had extended historical contact with one another. Galton suggested that Tylor's conclusions were suspect because his sample was biased.

Since then, there have been many attempts, none completely successful, to solve Galton's problem. The failure to solve this problem is due to two separate issues. First, theoretically, anthropologists have assumed that cultures are holistic linguistic communities in which members share common traits, values, technologies, and complementary (if not similar) behavioral repertoires. Second, methodologically, anthropologists have assumed that Galton's problem is a statistical, not theoretical, problem; hence, solutions have focused on devising a proper sampling technique. However, it has been shown that there are no ultimate primitive units of culture and that the units researchers use depend on the questions asked. A common cross-cultural sampling criterion is to select only those cultures for study in which the natives speak different languages. It is argued here that if cultures diverge on language, then they are independent cultural units subject to statistical analyses. Unfortunately, the language criterion for independence does not always hold. For example, if the researcher hypothesizes that male circumcision is related to cultural prohibitions against drinking alcohol and has a sample containing predominantly Islamic cultures whose members speak different languages, the results will still be biased by these cultures sharing a common Islamic heritage. Since Christianity and Islam are global religions, it does not matter if a sample consists of cultures that are from disparate regions and where the indigenous people speak different languages: their common religious background is enough to bias the sample.

In an actual cross-cultural study, Andrey Korotayev and I hypothesized that the number of supracommunal levels (i.e., levels of political integration) is directly correlated with the presence of a class structure. The correlations for this hypothesis were found to be significant with a sample of only Islamic societies ($\text{Rho} = 0.55$, $p = 0.01$) and with a worldwide sample ($\text{Rho} = 0.62$, $p < 0.001$). In this case, it simply does not matter whether a large and diverse worldwide sample or a sample of just Islamic societies is used; any number of cultures from the Islamic world can be included in the worldwide sample without distorting the final results.

We decided to perform the same test using only ethnicity (rather than language or religion) as the criteria for delineating cultural units. In addition, we chose a sample of only equatorial Bantu cultural groups, selecting a sample of cultures that were, according to all sampling criteria, dependent units. The correlation between the two variables selected in this closely knit cultural region is even higher and more statistically significant than the one for the worldwide sample ($\text{Rho} = 0.72$, $p < 0.00000000000000001$). This cannot be explained by a Galton effect. The hypothesized correlation between political integration and class structure—the greater the political integration, the more stratification we find in that society—holds for all our three samples, with the range of variation more pronounced in the ethnically similar sample than in the worldwide sample. In this case, the Galton effect was utterly irrelevant to the analysis. The concern voiced first by Galton and later by George Murdock and Douglas White with regard to selecting independent cultures for comparative research appears not to be relevant for some cross-cultural comparisons.

These results demonstrate that it is impossible to know a priori what criteria to use for selecting units for cross-cultural analysis. The criteria will vary on a case-by-case (or rather hypothesis-by-hypothesis) basis. If culture is seen as modular, rather than holistic, and if those modules consist of a knowledge–emotional–behavioral repertoire cluster, then one has to determine what criteria are responsible for organizing the cluster. In certain circumstances, cultural clusters are not necessarily geographically contiguous or even proximate but can be organized on the basis of global transcultural and transcontinental systems such as Christianity or Islam. In other circumstances, linguistic and cultural area boundaries serve as useful criteria for identifying cultural units. At still other times, it is unnecessary to be concerned about linguistic or cultural areas because differences are predominantly affected by environmental or other extracultural factors.

Checklist for Determining Units of Analysis

Units of analysis are always entities. The least problematic entity is the individual. The individual is the primary unit of analysis for studies in which the focus is on individual agency or the characteristics and behaviors of individuals. The limitations of using the individual as the unit of analysis are that group characteristics and behaviors can only be measured indirectly and studies are prone to the individual differences fallacy in which groups are attributed to possess the characteristics of either token-type individuals or those characteristics derived from central tendency measures of a sample population. In order to describe and analyze group-level characteristics from the individual as the unit of analysis, the researcher must posit an *a priori* theoretical justification for making such a leap. Collectivities can be the units of analysis. For small-scale groups the use of such units is relatively unproblematic and has great advantages if one is going to analyze group or ecological characteristics and dynamics. However, one cannot reduce group-level units to

particular individuals. Statistical methods were devised to work from the part to the whole rather than from the whole to the part. Thus, inferences from the individual to the group can be plausible and inferentially appropriate if theoretically specified, but inferences from the group to the individual are seldom, if ever, plausible or appropriate. Symbols and concepts should be avoided as units of analysis because they lack entativity. Large-scale collectivities, such as gender, religion, and culture, lack generalized entativety but may have specific entativety; this is to be discovered on a case by case or, as mentioned previously, hypothesis-by-hypothesis basis. A unit of analysis must be clearly defined. It cannot be used as a variable; rather, variables are extracted from the unit of analysis. Most important, there should always be a theory of analysis that justifies the choice of the units for analysis.

Acknowledgments

This article has benefited from the comments of two anonymous reviewers and the perspicuous editorial eye of the author's wife, Trini Garro.

See Also the Following Articles

Ecological Fallacy • Socio-Economic Considerations

Further Reading

Anderson, B. (1983). *Imagined Communities.* Verso, London.

Bailey, F. G. (1988). *Humbuggery and Manipulation: The Art of Leadership.* Cornell University Press, Ithaca, NY.

Bavelas, A. (1950). Communication patterns in task oriented groups. *J. Acoust. Soc. Am.* **22,** 271–282.

Bellah, R., Madsen, R., Sullivan, W., Swidler, A., and Tipton, S. (1985). *Habits of the Heart.* University of California Press, Berkeley.

Bellah, R., Madsen, R., Sullivan, W., Swidler, A., and Tipton, S. (1992). *The Good Society.* Vintage, New York.

Billings, J., Zeitel, L., Lukomnik, J., Carey, T. S., Blank, A. E., and Newman, L. (1993). Impact of socioeconomic status on hospital use in New York City. *Health Affairs* **12**(1), 162–173.

Collins, R. (1990). Stratification, emotinal energy, and the transient emotins. In *Research Agendas in the Sociology of Emotions* (T. D. Kemper, ed.), pp. 27–57. State University of New York Press, Albany.

Gatlung, J. (1967). *Theory and Methods of Social Research.* Columbia University Press, New York.

Goodenough, W. H. (1956). Componential analysis and the study of meaning. *Language* **32,** 195–216.

Goodman, L. (1977). *Measures of Association for Cross Classifications.* Springer-Verlag, New York.

Granovetter, M. (1973). The strength of weak ties. *Am. J. Sociol.* **78,** 1360–1380.

Heider, F. (1977). On behavior and attribution. In *Perspectives in Social Networks* (P. W. Holl and S. Leinhardt, eds.), pp. 51–62. Academic Press, New York.

Hofer, T. P. (1998, June). Use of community versus individual socioeconomic data in predicting variation in hospital use. *Health Services Res.,* 1–11.

King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press, Princeton, NJ.

Richards, J. C., Abbot, R., and Tull, D. S. (1991). What can be done about interviewer bias. *Research in Marketing* **3,** 443–462.

Romney, K. A., Weller, S. C., and Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *Am. Anthropologist* **88,** 313–338.

Service, E. (1962). *Primitive Social Organization: An Evolutionary Perspective.* Random House, New York.

Simmel, G. (1950). *The Sociology of Georg Simmel* (K. Wolff, ed. and trans.), Free Press, Glencoe, IL (and Trans.).

Weber, M. (1964). *The Theory of Social and Economic Organization.* Free Press, New York. [Original work published 1947].

Williams, J. E., and Best, D. L. (1990a). *Measuring Sex Stereotypes: A Thirty Nation Study.* Sage, Newbury Park, CA.

Williams, J. E., and Best, D. L. (1990b). *Sex and Psyche: Gender and Self Viewed Cross-Culturally.* Sage, Newbury Park, CA.

Unobtrusive Methods

Raymond M. Lee
Royal Holloway University of London, Egham, Surrey, United Kingdom

Glossary

accretion measure A type of unobtrusive measure based on the deposit of material in some setting.
artifact A feature of a research design that affects the validity the results produced.
episodic records Archival records, the discontinuous form of which does not allow trends to be identified.
erosion measure A type of unobtrusive measure produced by wear on some material.
reactivity The potential for research participants to change their behavior due to the presence of an investigator.
running records Type of unobtrusive measure involving ongoing, continuous documentary sources.
simple observation A method of observation in which the observer acts in a largely passive and nonintrusive way.
traces A type of unobtrusive measure in which physical remnants are used to provide information about social behavior.
unobtrusive measure A measurement collected without directly eliciting information from a respondent or informant.

Unobtrusive measures are measurements of social phenomenon that derive from methods of data collection *not* involving the direct elicitation of information from research participants.

Introduction

In 1966, Eugene J. Webb, Donald T. Campbell, Richard D. Schwarz, and Lee Sechrest published a witty and somewhat irreverent book entitled *Unobtrusive Measures*, which made a case for using sources of data that did not rely on direct interactional involvement between researcher and research participant at the point of data collection. The strength of unobtrusive measures was argued by Webb *et al*. to lie in their nonreactivity; that is, they avoided sources of invalidity produced when researchers directly elicit information from respondents or informants. (Unobtrusive measures are sometimes also referred to as nonreactive measures, a term which served as the title of the revised version of Webb *et al*.'s book published in 1981.)

Webb *et al*. argued that data collection instruments of many kinds suffer from methodological weaknesses associated with reactivity. Thus, respondents in interview- and questionnaire-based studies often try to manage impressions of themselves in order to create a positive image in the eyes of an interviewer. There are well-documented tendencies, for example, for respondents to overreport socially desirable behaviors and underreport socially undesirable behaviors, to claim to have opinions about fictitious topics, or to choose responses to questionnaire items based on their perceptions of the social characteristics of an interviewer.

Respondents, moreover, must be accessible and cooperative if survey-based methods are to be effective. Yet, there appears to be evidence from a number of countries suggesting a decline in levels of survey response. Concern about the artifactual character of data might also extend to experimental methods if research subjects act in ways they presume will ensure a successful outcome, or if those willing to participate in experiments differ markedly in their social and attitudinal characteristics from those who do not.

Taking concepts, metaphors, and examples from disciplines as diverse as geology, archaeology, and historiography, Webb *et al*.'s writing on unobtrusive measures is strongly informed by "multiple operationism." The assumption here is that research findings are potentially subject to the hypothesis that they are artifacts of the method used to collect the data. Sources of

invalidity, however, are not the same across methods. In particular, the problems of reactivity that afflict direct elicitation methods are absent when data are collected unobtrusively. One implication of this is that unobtrusive methods are not simply alternatives to direct elicitation methods, but complementary to them. For any given theory, testing is only possible at those points, "outcroppings" as Webb *et al*. describe them, where theoretical predictions and available instrumentation meet. The outcome of any one test is necessarily equivocal. The more remote or independent such checks, however, the more confirmatory their agreement. There is clearly here a justification for the use of multiple sources of data; configuring different methods, each of which is fallible in a different way, gives greater purchase on the problem to hand than a reliance on a single method.

Sources of Unobtrusive Measures

Webb *et al*. propose three types of data as sources of unobtrusive measures: physical traces (the evidence that people in traversing their physical environment leave behind them), nonparticipant observation, and documentary sources. It seems that this classification was developed primarily for expository purposes rather than being intended to convey conceptual clarity. Some alternative typologies have been proposed. Emmison and Smith emphasizing the visual character of unobtrusive data distinguish between: (a) two-dimensional visual sources such as images, signs, and representations, (b) three-dimensional sources, like settings, objects, and traces, and (c) lived and living forms of visual data, i.e., the built environment, human bodies, and interactional forms. Lee has proposed that Webb *et al*.'s passive typology of data sources be recast into a more active typology of data acquisition methods to allow a greater understanding of how particular measures come to be generated. In this schema, traces become "found data," observation methods yield "captured data," and documents can be thought of as forms of "retrieved data." Whatever their provenance, Webb *et al*. place a heavy premium on measures which are novel, playful, creative, or serendipitous.

Traces

Traces are physical remnants produced by erosion of the environment or accretion to it. A classic example of an erosion measure is the wear on floor tiles as an index of the traffic passing over them. Webb *et al*. give the instance of an exhibit showing live, hatching chicks at the Chicago Museum of Science and Industry. So popular was the exhibit that the vinyl floor tiles around it needed to be replaced approximately every six weeks. Floor tiles in other areas of the museum lasted for years without needing to be replaced. Another such example is the use of smudges, finger marks, turned-down pages and the like in library books as an index of their popularity. Graffiti provide an example of an accretion measure. The materials and techniques used to make graffiti, the various forms graffiti take, and the content of messages are all suitable topics of study. Garbage collected either at the curbside or excavated from landfills is also a fruitful source of unobtrusive data. For example, the number of condom wrappers found in garbage might be taken as a measure of the effectiveness of public health messages about protection against HIV infection. Comparing garbage counts with self-report provides one way of validating survey findings. Alcohol consumption, it seems, is often underreported on surveys when compared with the numbers of discarded drinks containers found in garbage.

Traces are ubiquitous, available at low cost and easily quantifiable. There are few ethical problems associated with their use. Gathering trace data causes little or no inconvenience to research participants who are anonymous (indeed in many cases their identity is completely unknown). Traces are often cumulative, permitting the collection of longitudinal data. In contrast, traces usually produce conservative estimates of behavior; some activities leave no traces or obliterate those that already exist. Trace data can take time to accumulate. It is often difficult to obtain the necessary population data that would allow a rate for some measure to be calculated. In addition, detecting the presence of response sets and patterns of selectivity in data based on trace measures can be rather difficult. For example, what finds its way into garbage is affected by recycling practices. A measure like differential floor wear, for example, depends on the physical properties of the floor covering itself; carpet wears out more quickly than tile, while materials are eroded or deposited in ways that are not necessarily independent of other erosions or deposits.

Observation

What people do and say during the daily passage of their lives, how they move through time and space, and the social patterns associated with posture, position, demeanor, and display, are all amenable to simple observation. Emblematic objects such as tattoos, body piercings, and clothing styles lend themselves to observation, as do interactional gestures and the social organization and use of physical space. Simple observation, put rather baldly, is field observation. It differs from observation in experimental contexts in that the observer has relatively little control over the setting, and differs from the postcoding of film or video records because observation and recording occur contemporaneously with the behavior being studied. In its systematicity it can also be

distinguished from participant observation methods used in sociology and anthropology. Social processes involving very large or very small spans of time or space might need to be manipulated through, for example, the use of high speed filming or time-lapse photography if our sensory apparatus is to be fully able to apprehend them. Recognizable "observational genres" have grown up with researchers designing observation studies around activities such as driving behavior, help-seeking, the return of lost objects, and the provision of goods and services.

Observational methods are often used where interview methods are inappropriate, such as in studies where research subjects, young children for example, have limited verbal ability, where potential informants lack a social vocabulary for answering questions about some kinds of behavior, or where participants are deeply engrossed in activities that would be disrupted by the intrusion of an interviewer. Observation is appropriate in some cases because analytic interest is focused, not on individuals, but on the relationships or interactions between them. Settings such as bars or factories where ambient noise levels are high do not always lend themselves to interviewing but might be suitable for observational study. Observation might be the only way to capture activities that are fleeting, or where respondents are likely to react to questioning in a strongly defensive way. Weick suggests that many everyday activities can be modified in ways that yield opportunities for observation, and that naturally occurring "provocations" such as accidental disruptions to a setting often yield valuable information.

Observational studies minimize problems of reactivity because people who do not know they are being studied do not change their behavior. Studying people without their permission, however, potentially negates the principle of informed consent. In fact, since the social expectations that govern behavior in public places assume that it will be observable and subject to scrutiny by others, unobtrusive observation carried out in public settings is frequently regarded as being less problematic than, for example, covert participant observation, or the deception of subjects in social science experiments.

Observational Sampling

Observation inevitably involves sampling. Decisions need to be made about what is to be studied, where and when an observation is to take place, and what the observer should notice and record during the observation. It can, however, be difficult to identify suitable sampling frames for observational studies or to decide how many periods of observation are needed and how long each should be. Although their work draws mainly on field studies of animal behavior, Martin and Bateson have identified the major sampling procedures for observational studies: (a) *ad libitum* sampling, (b) focal sampling, (c) scan sampling, and (d) behavior sampling. In *Ad libitum* sampling, systematic procedures are not followed. The observer notes what is visible and potentially relevant. There is a tendency, therefore, to focus on behaviors that are visible and discernible, and the potential exists for missing transitory or subtle behaviors. Focal sampling involves observing for a specified time one sample unit, such as an individual or relational pairing, and recording during that time all instances of a number of different categories of behavior. Because the focal unit can leave the setting or be out of sight of the observer, focal sampling can be difficult under field conditions. Since behavior out of sight might differ from behavior in sight, focal sampling potentially produces a bias toward recording public behavior. Scan sampling involves scanning a group of subjects at regular intervals. At a particular moment the behavior of each individual in the setting is recorded. Conspicuous individuals or behaviors are more likely to be noticed and therefore overrepresented; it is also often only possible to record relatively few categories of behavior. With behavior sampling some group or setting is observed in its entirety. Each time a particular behavior occurs, its occurrence is recorded along with a note of which sample element was involved.

Recording Methods

Martin and Bateson identify two methods for recording behavioral data: continuous (or "all occurrences") recording, and time sampling. An intensive and thorough procedure, continuous recording aims to produce a precise and faithful record of how often and for how long particular behaviors occur, with accurate recording of start and stop times. The method allows the frequency and duration of behaviors to be measured precisely, and it does not involve the loss of information inevitably associated with sampling. Continuous recording, however, is a burdensome activity, usually making it possible for an observer to attend only to a relatively few categories of behavior.

Time sampling requires that observations be recorded periodically, rather than continuously, with a random selection of time points for observation. The intermittent character of the observation means that the burden of work on the observer is reduced. In consequence, time sampling is arguably more reliable than continuous sampling because it allows more categories to be measured and more of the subjects present in the setting to be studied. In sampling behavior there is a need to balance the accuracy of measurement against its reliability and the ease with which measures can be obtained. The former implies short sample intervals, the latter long ones. There is no automatic way of determining how long or how short a sample interval should be. Choosing an interval will often be a matter of trial and error and/or judgement, or dependant on a pilot study.

There are two types of time sampling: instantaneous sampling and one-zero sampling. In instantaneous (or point) sampling the period of observation is divided up into short sample intervals. At the instant each sample point is reached, the observer records whether the behavior of interest is occurring or not. Rare events, those of relatively short duration or inconspicuous activities are not well captured by instantaneous sampling. In one–zero sampling, the observer records at each sample point whether the behavior of interest occurred or did not occur during the preceding sample interval. The method can produce biased results since behavior is recorded no matter how often it appears or for how long it occurs, and because events clustered at particular times tend to be undercounted relative to those spaced evenly across the whole period of observation. One–zero sampling might be appropriate in studies of intermittent behavior which are difficult to capture with either continuous recording or instantaneous sampling.

Reliability

Observation often provides "content-limited data," as Webb *et al*. describe it, since the information available to the researcher is contained within what is visible, the import of which might not be obvious without access to other more potentially reactive forms of data. Levels of reactivity are affected by the extent to which people are caught up in what they are doing, and the extent to which they have become accustomed to the presence of the observer. To increase the reliability of an observation it might be appropriate to use multiple observers and multiple methods of data recording. Observers can also be asked to assess the degree to which they thought subjects were acting in a reactive manner during the observation. Reliability can be assessed by looking at levels of intraobserver reliability, i.e., the extent to which individual observers are consistent in their practice, and at levels of interobserver agreement, the extent to which different observers produce similar results when they observe the same behavior on the same occasions. In some contexts it might also be appropriate to assess the combined effects of observer, setting, situation and observational categories by estimating the ratio of within-individual or setting variation to between-individual or setting variation.

Documentary Sources

Webb *et al*. make a rather arbitrary distinction between "running records," that is, records or documents produced (and often published) on a regular basis, and "episodic and private records," which are archival materials having a discontinuous form.

Running Records

Actuarial records registering the volume of births, marriages, and deaths are perhaps the most obvious example of running records, while the mass media provide a vast wealth of measures in the form of news stories, obituaries, wedding announcements, personal advertisements, advice columns, cartoons, editorials, advertisements, and the like. Running records, which are increasingly available in digital form, are valued for their ubiquity, their low cost, and their convenience. They can serve as a source of validation data for self-report data. Although they produce data restricted in content, running records typically cover lengthy time periods and generate considerable volumes of material. Because they extend over long periods, running records allow trends to be established, permit the exploration of temporal patterns, and provide opportunities for quasi-experimentation. In addition, associations, continuities, and discontinuities between different sets of records can sometimes yield information of a kind difficult to obtain by other means, for example, producing estimates of deviant behavior routinely hidden from view.

Running records are socially situated products. The quality of records varies depending on the nature of the processes involved in recording them. Apparent trends found can reflect external factors such as changes in record-keeping practices, or result from selective recording. Moreover, researchers in using running records might need to make careful judgements about issues such as transformation, aggregation, and time-series analysis, in other words, about how measures are to be expressed, how they might be combined, and how changes over time are to be handled.

There are few ethical problems involved in the use of running records, except where record linkage makes it possible to deduce the identity of individuals in the data. Even here the likelihood of deductive disclosure can be minimized by using strategies such as releasing results only on random subsamples of the data, including small amounts of random error, or more technical strategies such as broadbanding (the avoidance of finely detailed report categories) or microaggregation (where average responses for small clustered aggregates of research subjects are reported rather than individual scores). Because these strategies cloak the identity of individuals by introducing indeterminacy into data, their use involves some inevitable degradation of data quality.

Episodic Records

The use of documentary sources by researchers outside disciplines like history has grown appreciably in recent years. The interest in personal documents—letters, diaries, journals, memorabilia, family photographs, and the like—reflects a growing interest in discursive and textual

practice, a desire to incorporate elements of personal experience into the research process, and a commitment to give voice to those, such as women and members of minority groups, that social science has traditionally excluded, silenced, or marginalized. The mass media are full of visual images. Changes in the composition of such images provide an indication of shifts in the social valuation of particular groups. How images are assembled often gives an insight into how activities, events, and roles are socially constructed. Visual materials provide a cheap, effective, and readily accessible vehicle for making cross-national comparisons, and in less recognized forms, such as product packaging, attention to the visual can produce insights into the social use of objects. The judicial and legislative processes are a frequently unanticipated source of rich material on white collar crime, political malfeasance, and interrogatory discourse. Court transcripts, testimony before legislative committees and commissions, and the results of requests made under freedom of information laws can all yield data useful to the social researcher.

The relationship between documents available to a researcher and the wider universe of potentially relevant documents that exist or have existed is a problematic one. How far a particular set of documents can be taken to be representative of a wider universe of documents is affected by patterns of differential survival and variations in the accessability of documents. Documents are also fragile. They can be deliberately or accidentally destroyed, or be lost. Neither the recovery of a document's intended meaning, nor of the meaning received by its recipient(s) or audience(s) is unproblematic. Documentary analysis requires an appreciation of genre and stylistics, and an understanding of the context in which a given document was produced. The potential for reactivity present when data are elicited face-to-face is absent in documents. Pressures toward positive self-presentation may, however, still be present, for example, in material produced with an eye to eventual publication.

A range of strategies have been developed for the analysis of textual and graphical data, including quantitative content analysis, grounded theory procedures, and semiotic analysis. A wide variety of software tools is now available to help in the analysis of textual, and to a lesser extent, graphical data.

Unobtrusive Measures and the Internet

The Internet lends itself rather readily to data collection that does not involve the direct intervention of an investigator. While not yet universal, the Internet has recast the constraints of space, time, and cost typically juggled by researchers, providing opportunities not previously available to researchers in remote, small-scale or resource-poor environments.

The possibilities for the retrieval of secondary data and archival material have expanded massively with the advent of the Internet, enhancing the availability of running records and at least some kinds of documentary sources. Large volumes of machine-readable data on computer use, networking, communication processes, and message content are all available to the researcher at low cost, and with relatively little effort. The patterns revealed in these data are often not necessarily discernible by the participants involved. For qualitative researchers, the Internet allows for first-hand naturalistic investigation into the character of computer-mediated communication itself. The Internet opens up opportunities previously unavailable to study behavior prospectively. The transmission of rumours, arguments, the development of relationships, the unmasking of previously unsuspected selves, and the remedial strategies used to reinstate the good character of those who trangress against normative expectations, all of which previously could only be captured in retrospect if at all, can now be followed in their electronic manifestations from their inception to their conclusion.

Research in cyberspace falls within the scope of existing guidelines on ethical research practice in respect of informed consent, privacy and confidentiality and the need to protect research participants from harm. It is less clear, however, that existing guidelines are adequate. For example, obtaining informed consent in cyberspace where patterns of participation and involvement shift rapidly can be much more problematic than in face-to-face contexts.

Generating Unobtrusive Measures

A difficulty with unobtrusive measures is that there is little explicit guidance on how to generate unelicited data relevant to a particular research problem. Some writers have adopted an *orientational* approach, which stresses the importance of a creative and playful stance toward possible and actual sources of data. This approach can be contrasted with a *taxonomic* strategy, the basic aim of which is to identify particular properties of measures. From these properties, it might be possible to develop a generative taxonomy which would allow unobtrusive measures fitted to specific research purposes to be generated on demand. Neither approach is entirely satisfactory. An alternative view is that the generation of

unobtrusive measures involves the use of a variety of implicit heuristic strategies for finding data sources relevant to a particular research problem. Heuristics derived from a close reading of Webb *et al.*'s work include asking: "What features of some setting or situation can be made perceptually, normatively or culturally problematic and how?"; "How are the physical properties of objects inadvertently implicated in their social use?"; "What performative opportunities do objects offer?"; and "At what points and in what ways in society is information logged about social behaviour?" To date, the development and elaboration of such heuristics remains limited.

Unobtrusive Measures: An Assessment

The case against self-report methods can be exaggerated, but the presence of the researcher potentially shapes the responses of research participants in socially patterned ways. Research based on self-report is also vulnerable to the social factors affecting both the availability of research participants and their willingness to respond to researchers' questions. The problems of reactivity that potentially affect data collected by direct elicitation are absent when data are collected unobtrusively.

Unobtrusive measures are not used as widely as they might be. Some researchers reject an opportunistic attitude toward data feeling it better to produce, design, or create data for a specific purpose. The inferential weakness inherent in particular measures, at least when used on their own, is also taken to be a liability. A playful or creative attitude to data might not be easy to generate at will. Existing approaches to the generation of unobtrusive measures have generally not been satisfactory.

Unobtrusive measures commend themselves as ways of producing data complementary to direct elicitation methods, but with different weaknesses and strengths. Configuring different methods, each of which is fallible in a different way, might potentially give more reliable outcome than the measures produce by a single method. Unobtrusive measures can also be used where direct elicitation is difficult or dangerous. Simplicity and accessibility are advantages of unobtrusive measures. They rarely require great technical or technological sophistication, and they can provide a pathway for the verbally inaccessible. Unobtrusive methods are valuable in themselves because they encourage playful and creative approaches to data, undermining the tendency to use particular research methods because they are familiar or routine rather than appropriate to the problem in hand.

See Also the Following Articles

Data Collection, Primary vs. Secondary • Internet Measurement • Neutrality in Data Collection • Observational Studies

Further Reading

Dabbs, J. M., Jr. (1982). Making things visible. *Varieties of Qualitative Research* (J. Van Maanen, J. M. Dabbs, Jr., and R. B. Faulkner, eds.). Sage, Beverly Hills, CA.

Emmison, M., and Smith, P. (2000). *Researching the Visual.* Sage, London.

Fielding, N. G., and Lee, R. M. (1998). *Computer Analysis and Qualitative Research.* Sage, London.

Lee, R. M. (2000). *Unobtrusive Methods in Social Research.* Open University Press, Berkshire, UK.

Martin, P., and Bateson, P. (1993). *Measuring Behaviour: An Introductory Guide.* Cambridge University Press, Cambridge, UK.

Rathje, W., and Murphy, C. (1992). *Rubbish! The Archaeology of Garbage.* Harper Collins, New York.

Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences.* Rand McNally, Chicago.

Webb, E. J., Campbell, D. T., Schwartz, R. D., Sechrest, L., and Belew Grove, J. (1981). *Nonreactive Measures in the Social Sciences.* Houghton Mifflin, Dallas.

Weick, K. E. (1968). Systematic observational methods. *Handbook of Social Psychology* (G. Lindzey and E. Aronson, eds.). Addison-Wesley, Reading, MA.

Urban Economics

Daniel P. McMillen
University of Illinois, Chicago, Illinois, USA

Glossary

agglomeration economies Cost advantages from locating near other firms.
CBD A city's central business district.
Gini coefficient, Herfindahl index Measures of industrial concentration.
gradient The rate of decline in a variable with respect to distance.
hedonic price function A function relating prices to housing characteristics.
internal economies of scale Cost advantages owing to large firm size.
localization economies Cost advantages from being near other firms in the same industry.
location quotient A descriptive statistic measuring industrial concentration.
monocentric city model A theoretical model of urban spatial structure in which the central business district is the primary employment site.
polycentric city An urban area with large suburban employment centers.
subcenter A large employment center outside of the central city.
urbanization economies Cost advantages from being in an urban area with diverse industries.

Urban Economics is the application of economic modeling to problems affecting urban areas. A distinguishing feature of the field is the analysis of spatial relationships: urban economists traditionally seek to explain where economic activity takes place and why the spatial distribution of activity changes over time. Another major role of urban economists is the analysis of urban social problems using standard economic tools.

Agglomeration Economies

Why Do Cities Exist?

The starting point for any discussion of urban economics is the seemingly obvious question, why do cities exist? In the past, cities often were a means of defense—walled cities were easier to defend than scattered farms. Cities also served as religious centers. But throughout history, cities have served as trading centers. The question addressed in urban economics is what is it about a city that attracts economic activity?

The first answer to this question is that cities arise at critical points in the transportation system to facilitate trade between regions. Due to economies of scale in transportation, shipping is typically cheaper in large batches. Still today, most cities are located around harbors, rivers, and railroad and highway crossings. These sites offer advantages to large firms that ship their products to many different locations. Large firms exist because of internal economies of scale—cost advantages enjoyed by firms producing large quantities of a product. Large firms employ many workers, and these workers tend to live near the firm. The concentration of homes in turn attracts stores, restaurants, and other services that cater to the residents. The original transport advantage becomes the catalyst for diverse economic activity that begets more activity. Steel in Pittsburgh, meat packing in Chicago, and beer brewing in St. Louis are examples of industries with internal

economies of scale that were attracted to cities offering significant transportation cost advantages.

Cities also attract economic activity because of agglomeration economies, a term used to denote cost advantages enjoyed by a firm simply because other firms are located in the vicinity. Agglomeration economies are traditionally classified as localization or urbanization economies. A localization economy is a cost advantage that accrues to all firms operating in an industry as that industry expands within an urban area. For example, Silicon Valley in California's Santa Clara County has many small firms that together comprise a very large industry. Among other reasons, these firms find it worthwhile to locate in Silicon Valley because it is easy to hire the right workers in an area with a very large concentration of high-tech firms. In contrast, urbanization economies are cost advantages that are enjoyed by firms even when there are few other firms in the urban area in the same industry. A large urban area provides enough demand that even highly specialized companies can find local firms to provide a service, rather than having to provide the service internally with their own employees at a higher cost.

Measuring Agglomeration Using Location Quotients

Urban economists have used two primary measures of agglomeration economies. The first, the location quotient, is a simple descriptive statistic measuring the extent to which firms in a given industry are concentrated within individual cities. It is useful for identifying industries in which localization economies may be present. The second approach involves direct estimates of production functions to identify firms in an industry that are subject to internal economies of scale, localization economies, or urbanization economies. The production function approach is more direct and more insightful, but it is far more data intensive than the location quotient approach.

Location quotients are typically constructed using data on employment by industry. For a given urban area, the location quotient for industry i is simply:

$$LQ_i = \frac{\text{Percentage of the Urban Area's Employment in Industry } i}{\text{National Percentage of Employment in Industry } i} \quad (1)$$

For example, if 15% of a city's employment is in an industry, compared with 10% for the entire country, the location quotient is 1.5. High location quotients—those in excess of 1.2 or so—imply that a city specializes in an industry, which in turn implies that localization economies may be present.

Table I presents location quotients for selected cities for 2002. New York has significant specializations in information, financial activities, and educational and health services. The information sector also stands out in Los Angeles–Long Beach, Washington, DC, and San Jose. Although professional and business services are important in all five cities, this sector is particularly important in Washington, DC, and San Jose. Interestingly, the government sector is important but not overwhelming so in Washington, DC. Manufacturing—mostly in the computer industry—is a critical sector in San Jose. The preponderance of location quotients near 1.0 suggests that Chicago has a relatively diverse economy. Table I suggests that localization economies may be most likely in the information and professional services sectors.

Measuring Agglomeration Using Production Functions

The production function approach to measuring agglomeration economies provides direct estimates of the extent to which firms' production within urban areas is subject to

Table I Location Quotients for Selected Large PMSAs

	New York	*Los Angeles–Long Beach*	*Chicago*	*Washington, DC*	*San Jose*
Construction and mining	0.64	0.61	0.85	1.07	0.84
Manufacturing	0.36	1.13	1.02	0.23	1.91
Wholesale trade	0.97	1.25	1.33	0.56	0.92
Retail trade	0.70	0.86	0.90	0.81	0.79
Transportation and utilities	0.89	1.14	1.23	0.65	0.46
Information	1.81	1.97	0.99	1.59	1.43
Financial activities	1.93	0.95	1.26	0.90	0.64
Professional and business services	1.21	1.17	1.29	1.68	1.54
Educational and health services	1.46	0.90	0.94	0.85	0.83
Leisure and hospitality	0.78	0.95	0.88	0.90	0.83
Other services	1.01	0.88	1.04	1.41	0.71
Government	0.97	0.91	0.76	1.34	0.65

internal scale economies, localization economies, and urbanization economies. The typical production function is

$$y_{ij} = g(S_j) f(K_{ij}, L_{ij}, Z_{ij}), \tag{2}$$

where y_{ij} represents output for the ith firm in the jth city, and K, L, and Z represent capital, labor, and other inputs. The term $g(S_j)$ is a shift factor representing internal economies of scale or agglomeration economies, with $g \geq 0$. Scale, S_j, is measured by either metropolitan employment or population to represent urbanization economies, or industry employment to represent localization economies. In general, empirical studies find evidence that both urbanization and localization economies are important determinants of output in a variety of industries, countries, and times. A salient example is Henderson (2003).

Measuring the Spatial Distribution of Economic Activity

Urban theory predicts that agglomeration economies give firms an incentive to locate in large metropolitan areas even though the cost of land and labor is higher in cities. How spatially concentrated is the actual distribution of employment? Do large cities attract new firms? Are firms attracted to cities that already have a concentration of jobs in the same industry? Several methods have been proposed to study the spatial distribution of employment in urban areas. Of these, the most common are the Herfindahl and Gini indexes, and a new method proposed by Ellison and Glaeser (1997). These measures require the researcher to first specify a unit of analysis, such as counties or zip codes. Let z_i denote region i's share of national employment in a given industry, and let n be the number of regions under consideration. Then the Herfindahl index is given by

$$H = \sum_{i=1}^{n} z_i^2. \tag{3}$$

The index ranges from $1/n$ to 1. Complete concentration—all employment in an industry in one region—implies that $H = 1$ because $z_i = 0$ for all but the one region in which $z_i = 1$. Completely diversified employment implies that $z_i = 1/n$ for each region, which implies that $H = \sum_{i=1}^{n} (1/n)^2 = (1/n)$. That the Herfindahl index is simple to compute is its advantage. Its disadvantage is that it compares the actual spatial distribution of employment to an unrealistic counterfactual of complete homogeneity. Whereas the lowest possible value of H, $1/n$, implies that each region has exactly the same employment share, it is unreasonable to expect a small region such as Cheyenne, WY, to have a similar share of an industry as New York or Denver.

Due to this disadvantage of the Herfindahl index, a more commonly used measure of employment concentration is the Gini coefficient. As a simple example, suppose that we want to measure geographic concentration across five regions in the manufacturing sector. The first region is the largest, accounting for 60% of all employment. The remaining four regions each have 10% of total employment. Sorted from lowest to highest, the cumulative shares are 10, 20, 30, 40, and 100%. Manufacturing employment is more dispersed than the total: each region accounts for 20% of manufacturing employment. Figure 1 is a graph of the cumulative manufacturing employment shares against the cumulative shares of total employment. If manufacturing employment were just as dispersed as total employment, then the plot of manufacturing shares would follow the 45° line—the dotted line in Fig. 1. The area between the two lines is the Gini coefficient. The area is zero if manufacturing and total employment are equally dispersed. The Gini coefficient equals 0.5 if all employment is in one region since the plot of the cumulative share of manufacturing then follows the x axis, while the area under the 45° line is simply 0.5 by construction.

Ellison and Glaeser propose an index that formally incorporates random firm locations—a "dartboard approach"—as the alternative. They begin by defining a gross geographic index for a given industry within the manufacturing sector, $G = \sum_i (s_i - x_i)^2$, where s_i is the share of the industry's employment in region i, and x_i is manufacturing's share of total employment in the region. Thus, G measures the extent to which the industry's employment shares differ from the overall distribution of manufacturing employment. Ellison and Glaeser show that the expected value of G is $(1 - \sum_i x_i^2)H$ when the distribution of firms is completely random, where H is the Herfindahl index for the industry. Using this result, they define the following index:

$$\gamma = \frac{\left(G - \left(1 - \sum_i x_i^2\right)H\right)}{\left(1 - \sum_i x_i^2\right)H} \tag{4}$$

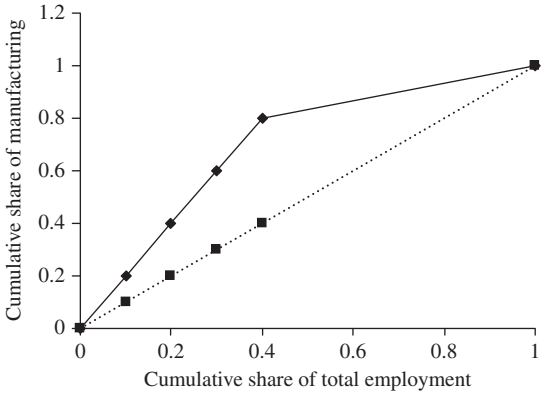


Figure 1 Constructing Gini coefficients for the manufacturing sector.

High values of G indicate that an industry is more spatially concentrated than expected given the overall spatial distribution of manufacturing employment.

If firms' location decisions are driven by internal scale economies alone, then there is no reason to expect that the distribution of the industry's employment should be different from the overall distribution of manufacturing jobs, and γ is close to zero. Localization economies imply clustering, and $\gamma > 0$. Ellison and Glaeser find that the U.S. automobile, carpet, and computer industries are highly concentrated spatially, whereas soft drink bottling, manufactured ice, and concrete products are not.

The Monocentric City Model

The monocentric city model, which is most closely associated with the work of Muth and Mills, is the core theoretical model of urban land use. In the Muth–Mills model, consumers receive utility from housing and other goods. Each household has a worker who commutes each day to the central business district (CBD). Each round trip to the CBD costs $\$t$ per mile. Since consumers have no direct preferences for one location over another, they would all try to live in the CBD in order to minimize their commuting costs unless house prices adjust to keep them indifferent between locations. In equilibrium, the price of housing must fall with distance from the CBD

$$\frac{\partial P_h(d)}{\partial d} = \frac{-t}{H(d)} \tag{5}$$

where $P_h(d)$ is the price and $H(d)$ is the quantity of housing at a site d miles from the CBD. This equation describes a simple relationship between house prices and distance from the CBD, as illustrated in Fig. 2.

The Mills–Muth model also includes producers who combine land and capital to produce housing. Since producers will pay more for land that is at sites with high house prices, land values will be high near the CBD. Where land prices are high, lot sizes are low, buildings are tall, and population density is high. Thus, Fig. 2 can be used to represent the price of housing, land values, building heights, and population density. The mirror image of Fig. 2—a smooth upward sloping function—represents lot sizes. All of these predictions are easily tested using simple regression procedures.

Measuring Urban Spatial Structure

The monocentric city model provides the basis for the method urban economists most often use to measure urban spatial structure. Figure 1 implies that variables such as the unit price of housing should decline smoothly and uniformly with distance from the CBD. The most commonly used functional form is the simple exponential function $y_i = e^{\alpha - \beta x_i}$ or $\ln y_i = \alpha - \beta x_i$, where x_i is the distance from the CBD for observation i, and y_i is any of the variables list in the previous section—population density, building heights, land values, etc. For example, suppose that y is population density. Then α represents the natural logarithm of population density in the CBD, and β measures the rate of decline in land values with respect to distance: land values fall by $100\beta\%$ with each additional mile from the CBD.

The coefficient β is referred to as the "gradient." Higher values of β imply a greater rate of decline in y. The gradient is the most frequently used measure of centralization. For example, Fig. 3 shows McMillen's estimates of the land value gradient for Chicago from 1836 to 1990. The gradient falls from 0.61 in 1836 to 0.14 in 1990. These numbers imply that land values fell by 61% with each additional mile from the CBD in 1836, compared with 14% in 1990. Clearly, Chicago was much more centralized near the time of its incorporation as a city than in 1990. Interestingly, the gradient has risen somewhat since 1960 as the city center has revitalized.

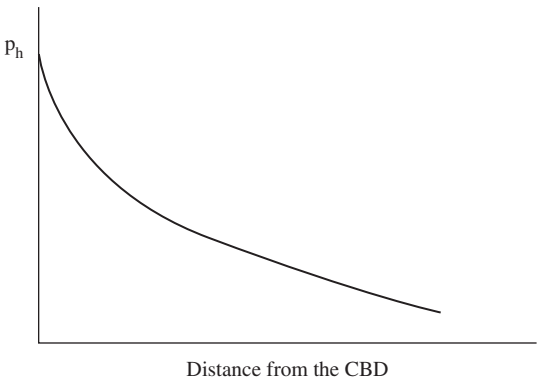


Figure 2 Predicted functional form for the price of housing.

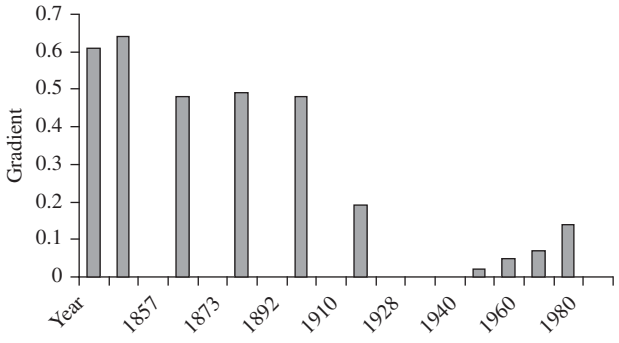


Figure 3 Land value gradients in Chicago.

Distance gradients have been used most commonly to study population density. Clark is the classic study. Both Muth and Mills made them an integral component of their research on urban spatial form. Coulson examines house prices using the approach. Gradients also have proved useful as a means of measuring rates of suburbanization and "sprawl."

Polycentric Cities

Recent extensions of the monocentric city model take account of the ongoing trend toward decentralized urban structure. As employment has moved to the suburbs, cities are more accurately portrayed as "polycentric"—a term coined to characterize urban areas with several employment centers outside of the traditional CBD. Suburban employment centers, or "subcenters," sometimes rival the CBD in total employment and in the scope of industrial activity.

Measuring the effects of subcenters on the urban economy is currently a popular research topic. The first step is to identify the subcenters. To qualify as a subcenter, an area must have a large concentration of employment, and it must have a significant effect on the overall spatial distribution of activity. The second requirement makes subcenter identification an empirical question.

McDonald proposed the first rigorous procedure for identifying subcenters. He estimates a simple exponential employment density function and identifies subcenters by looking for clusters of significantly positive residuals. Giuliano and Small define a subcenter as a cluster of contiguous employment zones where each zone has at least 10 employees per acre while the entire cluster has at least 10,000 employees. These cutoff points for minimum density and total employment can be altered to suit conditions unique to a metropolitan area. More procedures have been proposed since, and a variety of urban areas has been analyzed. The studies share the finding that subcenters are common in large metropolitan areas, but the traditional CBD still tends to dominate the spatial distribution of urban employment.

Urban Economics and Social Problems

Aside from studying the spatial distribution of economic activity, the other major role of urban economists is to use economic tools to analyze urban social problems. Examples include pollution, crime, poverty, racism and discrimination, education, and local public finance. These issues are important everywhere, not just in urban areas. But their effects are often more pronounced in cities because one problem can affect hundreds of thousands people.

The Hedonic Model

The first tool that urban economists use to measure the effects of urban social problems is the hedonic price function. The idea is to infer the value of a good, such as clean air, from house prices. Housing is the single most important asset for most households, and people gather lots of information before they are willing to spend thousands of dollars on a house. Home prices clearly are higher for larger, higher quality homes. Importantly, prices also reflect the characteristics of a neighborhood. An identical-looking house will trade for less if it is located in a polluted area with high crime and poor schools, and the size of the price discount will depend on the severity of these problems. Thus, we can infer how much people value clean air, low crime, and good schools from the value of housing in a neighborhood, once we have controlled for the house size and other characteristics of the homes.

Let P_i be the sales price of house i. Let X_i represent the vector of characteristics describing the house—living area, lot size, number of rooms, and so on. Also, let A_i represent the vector of characteristics describing the neighborhood, such as the crime rate, school quality, and air quality. The hedonic price function is simply $P_i = h(X_i, A_i) + \varepsilon_i$, where ε_i is an error term that is always present when estimating empirical relationships. The implicit market price of neighborhood characteristic j is simply $\delta h(X_i, A_i)/\delta A_{ij}$. In the absence of full arbitrage there is no reason to expect these hedonic prices to be constant; nonlinearity is a fundamental part of the hedonic equilibrium.

The hedonic approach can be extended to include both housing and labor markets. Metropolitan-wide amenities such as weather may be discounted into wages as well as house prices. However, most applied work analyzes either the housing market or the labor market alone, rather than both markets simultaneously.

Urban Crime

Lynch and Rasmussen use the hedonic approach to measure the economic impact of crime in Jacksonville, FL. They match data on over 2800 house sales to measures of crime across police beats. Explanatory variables for the home prices include structural and lot characteristics, neighborhood characteristics, and alternative measures of crime. They find that the cost of property and violent crime has only a small effect on sales prices. However, most places have little significant crime. In high-crime areas, homes trade at nearly a 40% discount relative to comparable houses in other areas.

Pollution

Kiel and Zabel analyze the effect of proximity to two Superfund sites on home values in Woburn, MA. The

Superfund is a federal program that was established in 1980 to clean up hazardous waste sites. Controlling for variables such as living area, age, and housing style, Kiel and Zabel find that proximity to the superfund sites lowered home values by as much as 12% in the time before the cleanup. They estimate the benefits from cleaning up the sites are between $72 million and $122 million in 1992 dollars.

School Quality

A final example of the use of hedonic price functions is found in the work of Gibbons and Machin, who analyze the benefits of school quality in England. The measure of school quality is the average performance on tests that are administered to all English students at age 11. Although the house price regressions have few controls for structural characteristics, Gibbons and Machin control carefully for locational effects. They estimate that the social valuation of a 1% improvement in primary school performance is £90 if the improvement is sustained over time.

See Also the Following Articles

Location Analysis • Locational Decision Making • Spatial Externalities • Urban Studies

Further Reading

Clark, C. (1951). Urban population densities. *J. Roy. Statist. Assoc. Ser. A* **114,** 490–496.

Coulson, N. E. (1991). Really useful tests of the monocentric city model. *Land Econ.* **67,** 299–307.

Ellison, G., and Glaeser, E. L. (1997). Geographic concentration in U.S. manufacturing industries: A dartboard approach. *J. Polit. Econ.* **105,** 889–927.

Giuliano, G., and Small, K. A. (1991). Subcenters in the Los Angeles region. *Reg. Sci. Urban Econ.* **21,** 163–182.

Gibbons, S., and Machin, S. (2003). Valuing English primary schools. *J. Urban Econ.* **53,** 197–219.

Henderson, J. V. (2003). Marshall's scale economies. *J. Urban Econ.* **53,** 1–28.

Kiel, K., and Zabel, J. (2001). Estimating the economic benefits of cleaning up Superfund sites: The case of Woburn, Massachusetts. *J. Real Estate Fin. Econ.* **22,** 163–184.

Lynch, A. K. (2001). Measuring the impact of crime on house prices. *Appl. Econ.* **33,** 1981–1989.

McDonald, J. F. (1987). The identification of urban employment subcenters. *J. Urban Econ.* **21,** 242–258.

McMillen, D. P. (1996). One hundred fifty years of land values in Chicago: A nonparametric approach. *J. Urban Econ.* **40,** 100–124.

Mills, E. S. (1972). *Studies in the Structure of the Urban Economy.* Resources for the Future, Baltimore.

Muth, R. F. (1969). *Cities and Housing.* University of Chicago Press, Chicago.

Urban Studies

Paul A. Longley
University College London, London, UK

Glossary

geodemographic indicators Small area measures of social, economic, and demographic conditions.
lifestyles data Quantitative measures of the varied consumption choices, shopping habits, and practices of identifiable individuals.
urban Pertaining to towns and cities.
urban ecology The study of the social and spatial organization of urban society.

The adjective "urban" pertains to towns and cities, which are objects that, in measurement terms, are usually taken to be crisp, well defined, and clearly delineated in space. The umbrella term "urban studies" is commonly used to describe academic activities involving urban geography, economics, anthropology, politics, and sociology, predominantly but not exclusively focusing on developed Western cities. Since 1938, when sociologist Louis Wirth published his classic paper on urbanism, the focus of urban studies has been on spatial differentiation in the occurrences of particular lifestyles. However, the conceptual framework within which to set this differentiation, the attribute mixes that have been used to measure and describe lifestyles, and the scales at which they have been analyzed have changed during the evolution of the field. There is a strong and explicit spatial component to urban studies; this is manifest in the concerns of urban researchers with residential differentiation and residential segregation, often using informal and formal techniques of spatial pattern analysis. From a historical perspective, there has been an established focus on the ways in which the nature of differentiation has been conceived, the ways in which it has been measured, and some of the ways in which it has been analyzed. This will be used to account for current state-of-the-art practice, and the breadth of current applications in urban studies.

Definitions, Scales, and Terms of Reference

Urban studies are an interdisciplinary meeting place of enduring importance, not least because the 21st century is becoming increasingly urban. Following a half-century of astonishing demographic growth and change, about half the world's population is currently classified as urban. It has been helpful to differentiate between inter- and intra-urban studies in the sense of the analysis of the city as a system within a system of cities. In global terms, fast-growing cities (such as those of Latin America) contrast with the patterns of deconcentration and decentralization in more mature settlement hierarchies (such as those of Western Europe). This big picture of urban studies is very much driven by the dynamics of national and international economic development, technological change, and growth, whereas the focus in the developed areas of the world is more on processes of differentiation and change within much more established urban environments. What is clear, however, is that at most all geographical scales of measurement, the umbrella term "urban" accommodates ever-increasing diversity in lifestyles, and that effective measurement of urban lifestyles provides daunting challenges. There is inevitably an overlap between urban studies and what is described as "development studies," with the balance of activity in urban studies at the intra-urban scale, and this scale is the focus here.

The ways in which urban phenomena are conceived very much determine the ways in which they are measured and then subsequently analyzed. Studies concerned principally with urban extent (such as inventory analyses focusing on the rate at which open countryside is annexed by urban growth) tend to adopt definitions that focus on the geographic extent of irreversibly urban artificial structures on the surface of Earth. These different structures support a range of residential, commercial, industrial, public open space, and transport land uses. They can

provide simple, robust, and directly comparable measures that focus on the dichotomy of natural/artificial land cover, and may be measured using digital framework data or statistical classification of the reflectance characteristics of the land surface. The resultant urban development patterns may not be entirely contiguous, and techniques of geographic information systems (GIS) can be used to devise appropriate contiguity and spatial association rules. Such indicators provide useful and direct indicators of the physical form and morphology of urban land cover, and are very useful in delineating the extent of individual urban settlements and in generating magnitude-of-size estimates for settlement systems. In recent years, developments in urban remote sensing have led to the deployment of instruments that are capable of identifying the reflectance characteristics of urban land cover to submeter precision. In addition to direct uses, such measures are also of use in developing countries where socioeconomic framework data, such as censuses, may not be available. They also provide some useful indications of settlement size distributions within the global settlement system. For reasons that are explored later, improvements in the resolution of satellite images have not been matched by commensurate improvements in the details of socioeconomic data on urban distributions, and thus today's high-resolution urban remote-sensing data may also be used to constrain GIS-based representations of socioeconomic distributions. Although increasingly detailed and precise in spatial terms, such representations nevertheless tell us rather little about urban lifestyles, unless supplemented by socioeconomic data. If augmented with consistent population size data, such measures provide indirect estimates of lifestyle (e.g., sprawling low-density settlements suggest suburban lifestyles), but realistically they are only likely to satisfy the requirements of most intraurban studies, wherein lifestyles are, in any case, likely to be rather homogeneous over quite extensive areas (as, for example, in developing-country settings). Today there is no single urban way of life (if ever there was), and there is a need for sharper differentiation between lifestyles.

All of this does not amount to social measurement in a strict sense—precise measures of the extent and physical morphology of the carcasses of urban settlements provide rather few direct indicators of the lifestyles of those who work and live in the city. Urban studies in the broader sense require information on urban function, and thus research focuses on the creation of zones that may be uniform in socioeconomic characteristics (as in the delineation of areas that experience urban deprivation or hardship, as a precursor to neighborhood policy implementation), that may delimit a clearly defined urban function (as in empirical delineation of levels of a central place hierarchy, or identification of a travel-to-work zone), or that fulfill an administrative role. Analysis of urban function and lifestyle is fundamentally dependent on the collection and availability of data pertaining to socioeconomic characteristics, policy functions, and lifestyle activities. Where available, and notwithstanding some success at developing pan-European data sets, this often restricts the scope of urban studies to national units of analysis or to subsets of them. There are exceptions to this in the attribution of particular functions or interaction patterns to urban areas (for example, in recent analyses of international telecommunication data). Today, most developed and many developing countries have socioeconomic data infrastructures that are quite rich compared with the even quite recent past, although there is variability between countries in terms of content and in terms of the degree of reliance on national censuses, address registers, and public sector sample surveys. From the standpoint of the urban analyst, however, all data share the unfortunate characteristic that they are available only for aggregations of individuals, for reasons of ethics and confidentiality. The foundations to analysis are thus always artificial aggregations, such as census tracts. Data availability and measurement issues have driven much of the agenda of urban studies, but there has also been an element of choice in the variables selected, in turn driven by the social constructs that they are deemed to represent.

Historical Perspective on Social Measurement in Urban Studies

Studies by Charles Booth of the intraurban geography of poverty in Victorian London provided perhaps the earliest systematic intraurban social measurements of population characteristics. Yet the conceptual roots to urban studies are conventionally traced to the work of Robert Park, Ernest Burgess, and the other Chicago-based ecologists from 1916 onward. The root metaphor in this early work was that of vegetation competition and succession, and this was broadly deemed transferable to human communities by differentiating between the biotic and cultural levels of urban society. In this work, biotic forces were manifest at the level of the community, whereas social forces were deemed to be asserted only at the cultural level of society as a whole. In conceptual terms, urban society thus provided a superstructure above the more basic competitive level of the community, and neighborhood communities were fashioned by subsocial forces. Park and his colleagues saw the role of what would now be called urban studies as investigating the community level and the subsocial forces that acted within it. In 1923, Burgess envisioned the community-level forces in his famous concentric-zone model, in which ecological forces of competition for space were deemed to fuel a process of radial expansion. The resulting pattern of

urban land uses can thus be thought of as a snapshot of the ripple effects of urban growth.

Park was avowedly empiricist in his approach to measurement and analysis (one report notes that Parks once famously described his methodology as "walking the streets"), yet a central contradiction of the work of the Chicago school was that the biotic and social components of residential differentiation could never be observed and measured in isolation from one another. Nevertheless, an enduring appeal of the work lay in its reference to explicitly spatial terms such as "neighborhood" and "district," and these came to conjure up the image of the city as a mosaic (or perhaps, more dynamically, as a kaleidoscope) of subareas, defined not just in terms of built form but also in terms of the patterning of socioeconomic and demographic groups. In short, these became the major sociological and geographical structures that were seen by academics and policymakers as shaping urban form, although in subsequent work, the Chicago school's theoretical concerns with dominance, functional interdependence, and differentiation were to become decoupled from empirical classification of mosaics of social worlds.

Subsequent empirical work thus focused on the patterning of the tiles of the urban mosaic as the principal objective of analysis, and secondary data sources, specifically censuses of population, were used to attribute values to the tiles. What became known as "social area" analysis thus developed as *à la carte* selections from menus of available census variables, using the mosaic of zones used in census dissemination. The approach, epitomized by the work of Eshref Shevky, Marilyn Williams, and Wendell Bell in Los Angeles and San Francisco in the mid-1950s, essentially provided a descriptive empirical technique for reducing multivariate census tract returns to the tripartite constructs of social rank, urbanization, and segregation. These constructs were provided in Shevky and Bell's limited, *post facto*, rationalization of their initial work, and sought to embed social measurements in changes in the range and intensity of social relations (social rank), differentiation of industrial function (urbanization), and increasing complexity of the organization of society (segregation). However, by the 1960s and 1970s, the approach had become driven overwhelmingly by inductive generalization from secondary data, using the then fashionable analytical techniques of factor analysis in the developing field of computer analysis. Social measurement and data analysis thus became divorced from any substantial or coherent conceptual framework, and urban studies became submerged by the tide of inductivism that continues to run through science and social science to this day. Data were allowed, by and large, to speak for themselves—with increasingly occasional lip service paid to concepts and theories.

The tide of inductivism ran fast, but in urban studies it dissipated quickly, and by the mid-1970s, disenchantment with data-led empiricism had led to the demise of factorial ecology as a sustainable epistemology. In very general terms, the approach has been supplanted by successive waves of urban managerialism and Marxist sociology. The field of urban studies has, in important respects, become rather stronger in conceptual terms, yet work in these recent genres has by and large proved to be no more empirically verifiable than are the ecological metaphors and social constructs that preceded them. Urban geography appears to have been an obvious casualty, for most current research appears to lack an explicit spatial dimension. In an era in which there has been a resurgence of interest in the neighborhood as a unit of social action, economic planning, and policy implementation, the field of urban studies has very largely withdrawn from the quest to generalize about intraurban socioeconomic distributions. However, this is not to say that the significance of inductive generalization about urban social patterning has been entirely wasted. At the same time as factorial ecologies and multivariate classifications were becoming unfashionable in academia, Richard Webber (working at the London Center for Environmental Studies) developed the approach into the branch of urban studies that has become known as geodemographics. The earliest UK national classifications were developed from the 1971 Census of Population, at four scales, ranging from the enumeration district (census block) to the parliamentary (electoral) constituency. The classifications were initially intended to guide local government in neighborhood policy implementation, but, following Webber's move into the private sector in 1979, the basic approach was successfully developed into commercial applications through proprietary systems such as the MOSAIC and ACORN classifications. Today, these kinds of systems are a proved social measurement technology and enjoy repeat purchases by a wide variety of businesses and service organizations. Numerous different systems exist for general and niche markets in North America and Europe, and systems are under development in much of Latin America and China. More rudimentary approaches to multivariate socioeconomic classification remain an important area of activity in the public sector, as in the calculation of standardized composite indices of levels of hardship and deprivation.

Whither Social Measurement in Urban Studies?

Beyond the academy, there continues to be very strong interest in social measurement of urban socioeconomic distributions, and many interesting developments of

technique and application are underway. One important development has been the recognition that richer depictions of urban lifestyles can be assembled from diverse and under-used digital sources that are available at a range of spatial scales. These include the limited range of attributes that can be gathered for individuals from address registers, to georeferenced vehicle registrations, social surveys, shopping surveys, and guarantee card returns. Increasing numbers of commercial databases concerning lifestyles are now routinely used in commerce: many enumerate salient characteristics of tens of millions of individuals. Although they are rarely collected to the same exacting standards to which conventional public sector data sets are collected, they are updated continuously and provide very detailed snapshots of the diverse lifestyles that are often to be found in quite small urban areas. Some enlightened private sector data providers have been prepared to deposit these disaggregate data in research archives (subject to confidentiality constraints). Although these vastly enrich the potential content of social classifications, their use raises a number of profound and possibly frustrating issues of coverage and representativeness, not least because commercial organizations are unlikely to be motivated toward assembling detailed data inventories on the "have nots" of society. Companies that use lifestyle lists to identify potential mail-drop targets are unlikely to select households with low incomes. For this reason, most lifestyle operators have decided that it is more cost effective to target the blanket door drops of questionnaires to postcode sectors with higher rather than lower levels of affluence. However, coarse administrative zones (such as UK postcode sectors) are the lowest areal units that can be leafleted by distributors, and this is likely to improve the representativeness of response. There is a clear need to cross-validate small area lifestyle measures with respect to external data sources.

These issues illustrates a number of the tensions in current social research: digital data capture is now routine, but scientific surveys conducted to rigorous standards account for a diminished real share of available data; the scale and pace of change to urban systems now makes the decennial snapshots of censuses increasingly irrelevant to policy needs; and the fission of lifestyles among urban populations that are increasingly heterogeneous at fine scales of granularity makes the limited content (attribute base) of censuses increasingly limiting in analysis of urban systems. Taken together, this argues for the need to generate scientific findings in real time, and with frequent update cycles; the need for interdisciplinary approaches to urban studies that blend together salient indicator variables into comprehensive indicators; and the need for public–private partnerships in order to unlock the potential of the richest, most relevant, and most recent data.

In relating the success of social measurement through geodemographic applications in business, it is important to note some differences in the measurement goals of different academics and private sector practitioners. In general terms, it is worth distinguishing between two rather different applications. Direct marketers, who communicate with individuals rather than serve areas, desire measures that are optimally predictive at the person or household level, but are not materially concerned whether there is any geographically systematic error in this estimate. There are no inferential errors generated in one-to-one marketing applications. By contrast, retailers and other organizations that serve areas, rather than individuals, are not particularly concerned whether their social measures are accurate at the household level. They, like the academic or policy analyst, are more concerned that whatever inferential errors there may be are not systematic at the area level. Thus, the best social measurements for direct marketers are not necessarily the best measures for analysts concerned with geographic catchments.

Today's diminished academic interest in social measurement of urban systems is a pity for a number of reasons. In conceptual terms, the experience of factorial ecology did bring general recognition of the ways in which choice of classification method, choice of variables, and to some extent choice of data source would determine the outcome of the classification. Today, similar sensitivity to context is recognized in the measurement of local or regional effects. Second, in measurement terms, more data are collected about more aspects of our individual lifestyles than at any point in the past, through routine interactions between humans and machines. Enlightened approaches to public data access (especially through online portals) make wide dissemination of socioeconomic data a reality and the creation of general-purpose and bespoke data systems straightforward. Geodemographic systems based on socioeconomic framework data can be successfully fused to census sources to provide richer depictions of lifestyles. And third, in analysis terms, the toolkit of spatial analysis and GIS now make it easier than ever before to match diverse data sources and accommodate the uncertainties created by scale and aggregation effects.

Developments in computation, technique, and data analysis continue to offer incremental improvements in the ways that geodemographic representations are specified, estimated, and tested, but it is correct to suggest that it is repeat purchases of a core tried and tested technology that ensure retention of the approach as a mainstay of contemporary urban studies. Hitherto, the overwhelming majority of geodemographic applications has concerned tactical and strategic decision making in private sector applications (specifically retailing), and it is probably true to say that the clearest indicator of "success" is the way in which improvements in targeting of goods and service offerings improve measured profitability. One

of the interesting challenges of the coming years will entail use of these techniques in public service applications, given the pressures to demonstrate value for money in targeting public funds according to local needs. It is not possible in an article of this length to examine the various caveats to the geodemographic approach, but issues of the content and coverage of the data sources that are used to create and update geodemographic profiles are certainly likely to become important in developing and extending the realm of geodemographic applications. The approach has very important contributions to make to the developing rationalities, performance metrics, and change measures in the developing public policy debate. It is of strategic importance that academics and policymakers engage in these important measurement issues, and do not simply become passive consumers of geodemographic systems. In this context, it is important to return to the themes of conception, measurement, and analysis. In the early days of urban studies, the issue of empirical verification of ecological analogy was seen as a crucial issue. Much of the recent history of urban studies has also posited concepts and processes that can only rarely and unsystematically be observed. The geodemographic approach, by contrast, entails a return to the mosaic metaphor of urban structure, and provides robust, transparent, and disaggregate observations of what is going on in urban systems. In conceptual terms, it is founded on the basic theoretical premise that "birds of a feather flock together." This is by no means a trivial concept, whether viewed in the context of genetic selection and mapping of the human genome, or in the simulation of city evolution as the outcome of cellular interactions across a range of geographic scales. It has been suggested that the patterning of urban social areas may express more than the outcome of (unmeasurable) economic and social processes and, at a conceptual level, that there may be much that can be developed from the success in measuring urban phenomena through geodemographics. It is not just the urban ecology of the Chicago school that might be revitalized: neighborhood classification could also develop the ideas of Shevky and Bell in terms of classifying urban areas according to the range and intensity of social relations, differentiation of industrial function, and increasing complexity of the segregation of society.

Consolidation

The scale and pace of urban development in the 21st century is without historical precedent, and the patterns of functional interdependence within and between national settlement systems make it no longer sensible to envisage any direct correspondence between urbanism and a small range of lifestyles. We are all now urban in some senses, but so too we are increasingly differentiated from one another in terms of our lifestyles. The mosaic of urban areas provides the most obvious laboratory in which to study diversity in lifestyles. The fission of urban lifestyles, in terms of both conventional and other social indicators, presents profound challenges to urban studies, and engagement with the social measurement task is pivotal to progress. The success of geodemographics in commercial and (increasingly) policy settings provides testimony to the relevance of urban studies, and the rationale for the approach is by no means ill conceived or atheoretical. It also provides evidence that improved measurement guides the development of better theory in this important area of social science.

Finally, a broader aspect of this quest to reinstate social measurement at the core of urban studies entails redefinition of urban studies in more than conventional territorial terms. The emergent computer-linked "e-society" is defined in large part by the changing connectivity and interactions between individuals, and between individuals and the state (which is also changing profoundly as new information and communications technologies become ever more pervasive). As lifestyles and the institutions that shape them adapt to the technologies of the digital age, it remains clear that the urban nexus will remain pivotal. Social measurement is absolutely central to understanding changes in the modes of social and economic interaction, and hence the organization of society.

Acknowledgments

This work was funded under ESRC AIM Fellowship RES-331-25-0001 and ESRC research grant RES-335-25-0020 (E-Society Programme).

See Also the Following Articles

Built Environment • Census, Varieties and Uses of Data • Cognitive Maps • Geographic Information Systems • Socio-Economic Considerations

Further Reading

Batty, M., and Shiode, N. (2003). Population growth dynamics in cities, countries and communication systems. In *Advanced Spatial Analysis: The CASA Book of GIS* (P. A. Longley and M. Batty, eds.), pp. 327–343. ESRI Press, Redlands, CA.

Benenson, I., and Torrens, P. (2004). Geosimulation: Object based modelling of urban phenomena. *Comput. Environ. Urban Syst.* **28,** 1–8.

Berry, B. J. L., and Horton, F. E. (1970). *Geographic Perspectives on Urban Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Dawkins, R. (1976). *The Selfish Gene.* Oxford University Press, Oxford.

Donnay, J.-P., Barnsley, M. J., and Longley, P. A. (eds.) (2001). *Remote Sensing and Urban Analysis*. Taylor and Francis, London.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage, London.

Goodchild, M. F., and Longley, P. A. (1999). The future of GIS and spatial analysis. In *Geographical Information Systems: Principles, Techniques, Management and Applications* (P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, eds.), pp. 567–580. Wiley, New York.

Harris, R. J., and Longley, P. A. (2002). Creating small area measures of urban deprivation. *Environ. Plan. A* **34,** 1073–1093.

Johnston, R. J. (2000). Urban studies. In *The Dictionary of Human Geography* (R. J. Johnston, D. Gregory, G. Pratt, and M. Watts, eds.), 4th Ed., pp. 870–871. Blackwell, Oxford.

Saunders, P. (2001). Urban ecology. In *Handbook of Urban Studies* (R. Paddison, ed.), pp. 31–51. Sage, London.

Thurstain-Goodwin, M. (2003). Data surfaces for a new policy geography. In *Advanced Spatial Analysis: The CASA Book of GIS* (P. A. Longley and M. Batty, eds.), pp. 145–169. ESRI Press, Redlands, CA.

Wirth, L. (1938). Urbanism as a way of life. *Am. J. Sociol.* **44,** 1–24.

Utility

Michael Quinn Patton
Union Institute and University, Minneapolis, Minnesota, USA

Glossary

conceptual use Use of research or evaluation to influence thinking about issues, policies, or programs in a general way.
findings use Identifiable influence of research or evaluation conclusions on decisions, policies, programs, or thinking of people.
instrumental use Direct action or decision that follows, at least in part, from research or evaluation findings.
misuse Inappropriate, dishonest, and improper interpretations and applications of findings, both intentional and unintentional, that distort, confuse, or obfuscate.
non-use Ignoring research and evaluation conclusions.
process use Identifiable effects of participating in research or evaluation on the participants and/or their organizations or communities, for example, enhancing evaluation capacity or increasing research skills.
symbolic use Evaluation is undertaken simply to conform with funding mandates or to legitimate pre-existing conclusions or biases.
use In the context of research or evaluation, to employ inquiry processes and apply findings.
utility Actual or potential use of research or evaluation.
utilization-focused evaluation An approach to evaluation that focuses on facilitating and achieving intended use by primary intended evaluation users.

Utility concerns the extent to which research and evaluation are useful—and actually used—to inform action, decisions, policies, program improvements, and understanding. The primary proof of utility is actual use. Different research and evaluation purposes lead to varying kinds and degrees of utility and use. Some studies are designed to inform decisions and support taking action; such uses are instrumental. Other studies aim at enhancing understanding, but no specific actions or concrete decisions are expected. Moreover, utility concerns not only use of findings, but also the effects of participating in research or evaluation on the participants and/or their organizations or communities, for example, enhancing evaluation capacity or increasing research skills. The shadow side of use is misuse, that is, inappropriate, dishonest, and improper interpretations and applications of findings, both intentional and unintentional, that distort, confuse, or obfuscate. Some evaluation approaches, like utilization-focused evaluation, make monitoring and attaining utility the centerpiece of the design. The relative balance in social science between attention to utility versus attention to truth is controversial, as are definitions of and attempts to measure utility.

Utility as a Concern in Applied Social Science

Applied social science, as opposed to basic research, strives for utility. Concern about utility has emerged at the interface between science and action, between knowing and doing. Funders of applied social science expect that findings from research will be able to be applied in solving social problems. Funders of program evaluation expect results to be used to improve programs. Utility of research and evaluation is increased when the findings are viewed by potential users as understandable, relevant, and credible. Efforts to increase utility have raised fundamental questions about human rationality, decision-making, and knowledge applied to creation of a better world.

Conclusion-Oriented versus Decision-Oriented Inquiry

Not all social science strives for utility. A classic distinction contrasts decision-oriented inquiry with conclusion-oriented inquiry. Conclusion-oriented social

measurement aims at the creation of knowledge as a good in and of itself, for example, understanding how the brain functions when learning to read. In contrast, decision-oriented inquiry aims to inform and influence decision makers, for example, evaluating reading programs to determine which are most effective. In decision-oriented inquiry, concern about utility comes to the fore.

The Challenge of Knowledge Use

Getting people to use what is known has become a critical concern across the different knowledge sectors of society. A major specialty in medicine (compliance research) is dedicated to understanding why so many people do not follow their doctor's orders. Common problems of information use underlie trying to get people to use seat belts, quit smoking, begin exercising, eat properly, and pay attention to evaluation findings. In the fields of nutrition, energy conservation, education, criminal justice, financial investment, human services, corporate management, international development—the list could go on and on—a central problem, often the central problem, is getting people to apply what is already known. In agriculture, a major activity of extension services is trying to get farmers to adopt new scientific methods.

These examples of the challenges of putting knowledge to use set a general context for the specific concern of research utility: narrowing the gap between generating research findings and actually getting those findings used.

Problems of Underuse and Nonuse

Historically, a great many studies have documented and examined the problems of underuse and nonuse. Both the scholarly literature on the subject and various commission reports on the use of social scientific knowledge reached a decidedly gloomy conclusion that instances where social science research had a clear and direct effect on policy were rare.

Government Underutilization of Knowledge

Research and evaluation have become part of political debates and government decision-making around the world. A study in 1995 by the U.S. General Accounting Office (GAO) raised questions about how well governments use research and evaluation. Entitled "Improving the Flow of Information to Congress," analysts concluded that research and evaluation findings sent to Congress seldom reached the right committees and people, that reports were poorly organized and communicated ineffectively, and that much data was too highly aggregated to be useful or was too difficult to digest. It seems unlikely that this American example is unique.

Many factors affect research and evaluation use in government, but politics is the overriding factor. Ideological conflicts overwhelm empirical evidence. This is true at state and local levels of government as well.

Problems of Misuse

The other side of the coin when looking at utility is misuse. Results from poorly conceived studies have frequently been given wide publicity while findings from good studies have been improperly used, selectively applied, and distorted in political debates. One form of misuse is called symbolic use referring to the selective and distorted use of empirical findings to legitimate pre-existing positions or biases, for example, attempting to enhance the credibility of a political stance by selectively citing evaluation findings that support that stance. Undertaking evaluation simply to conform with funding mandates with no intention to use the results is also a form of symbolic use. Social scientists face a dual challenge then: supporting and enhancing appropriate uses of knowledge while also working to eliminate improper uses.

Evaluation Standards and Utility

In the past, many researchers took the position that their responsibility was merely to design studies, collect data, and publish findings; what decision makers did with those findings was not their problem. This stance removed from the social scientist any responsibility for fostering use and placed all the blame for nonuse or underutilization on decision makers. While the role of social scientists in policy-making has long been debated, the emergence of program evaluation as a distinct field of professional practice led to an explicit focus on utility as a criterion of excellence.

Before the field of evaluation identified and adopted its own standards, criteria for judging evaluations could scarcely be differentiated from criteria for judging research in the traditional social and behavioral sciences, namely, technical quality and methodological rigor. Use was ignored. Methods decisions dominated the evaluation design process. By the late 1970s, however, it was becoming clear that greater methodological rigor was not solving the utility problem. Program staff and funders were becoming openly skeptical about spending scarce funds on evaluations they could not understand or found irrelevant. Evaluators were being asked to be "accountable" just as program staff were supposed to be accountable. How would evaluation be evaluated? It was in this context that professional evaluators began discussing standards.

Standards were hammered out over five years by a 17-member committee appointed by 12 professional organizations, with input from hundreds of practicing evaluation professionals. The standards published by the Joint Committee on Standards in 1981, and revised in 1994, dramatically make utility a primary criterion of

excellence. The other three general criteria in the standards framework are feasibility, propriety, and accuracy. (For the detailed standards see www.eval.org.)

Utility Standards

The Utility Standards are intended to ensure that an evaluation will serve the information needs of intended users. The standards call for clear identification of primary stakeholders, assuring evaluator credibility, collecting information responsive to the needs and interests of stakeholders, making value judgments clear, writing clear reports submitted on time, and planning, conducting, and reporting studies in ways that encourage follow-through by stakeholders to increase the likelihood of use.

Types of Utility

Alternative research and evaluation purposes intend and support varying uses. Applied research and evaluation are characterized by enormous diversity. From large-scale, long-term, international comparative designs costing millions of dollars to small, short evaluations of a single component in a local agency, the variety is vast. Thus, reducing the complexity of utility types to a few major categories will inevitably oversimplify. Yet, three major purpose categories capture the primary alternatives with regard to using empirical findings: (1) generating useable, generalizable knowledge, (2) making overall evaluative judgments about the merit, worth, value, and effectiveness of specific policies or programs, and (3) facilitating improvements in policies and programs.

These are not mutually exclusive purposes and some studies strive to incorporate all three approaches, but one is likely to become the dominant motif and prevail as the primary purpose informing design decisions and priority uses; or else, different aspects of a study are designed, compartmentalized, and sequenced to address these contrasting purposes. Confusion among these quite different purposes, or failure to prioritize them, is often the source of problems and misunderstandings when it turns out that different intended users had different expectations and priorities about utility.

Generating Useable, Generalizable Knowledge

Generalizations and ideas that come from research and evaluation help shape the development of policy and program theory. Generalizations from evaluation can become part of the knowledge that influences future decision-making about programs generally. Understandings gleaned from evaluations can contribute to applied social science theories about how to produce social change as well as inform and test implementation theory to better understand variations in program delivery and outcomes. Such knowledge-generating efforts focus beyond the effectiveness of a particular program to future program designs and policy formulation in general.

Building General Program Theory about Effectiveness

As the field of evaluation has matured and a vast number of evaluations has accumulated, evaluation researchers look across findings about specific programs to formulate generalizations about effectiveness. This involves synthesizing findings from different studies. These kinds of "lessons" constitute accumulated wisdom—principles of effectiveness or "best practices"—that can be adapted, indeed, must be adapted, to specific programs or organizations when disseminated.

Making Overall Evaluative Judgments about Merit and Worth

Evaluation research aimed at determining the overall merit, worth, or value of a program or policy derives its utility from being explicitly judgment-oriented. Merit refers to the intrinsic value of a program, for example, how effective it is in meeting the needs those it is intended help. Worth refers to extrinsic value to those outside the program, for example, to the larger community or society. A welfare program that gets jobs for recipients has *merit* for those who move out of poverty and *worth* to society by reducing welfare costs. Judgment-oriented evaluation approaches include performance measurement for public accountability; program audits; summative evaluations aimed at deciding if a program is sufficiently effective to be continued or replicated; and quality control and compliance reports.

Facilitating Improvements in Policies and Programs

Using evaluation results to improve a program turns out, in practice, to be fundamentally different from rendering judgment about overall effectiveness. Improvement-oriented forms of evaluation include formative evaluation aimed at identifying a program's strengths and weaknesses, quality enhancement efforts, and learning organization approaches. What these approaches share is a focus on making things better rather than rendering overall judgment. Judgment-oriented evaluation requires preordinate, explicit criteria, and values that form the basis for judgment. Improvement-oriented approaches tend to be more open-ended, gathering varieties of

data about strengths and weaknesses with the expectation that both will be found and each can be used to inform an ongoing cycle of reflection and innovation.

Instrumental Utility

Both judgment-oriented and improvement-oriented studies involve the instrumental use of results. Instrumental use occurs when a decision or action follows, at least in part, from the study. Specifically, instrumental utility is knowledge used for action in contrast to conceptual utility, which is knowledge for understanding.

Conceptual Utility

With conceptual use, no decision or action is expected; rather, a study is used to influence thinking about or understanding of issues. Evaluation findings can contribute knowledge by clarifying a program's model, testing theory, distinguishing types of interventions, or elaborating policy options. In other cases, conceptual use is more vague, with users seeking to better understand the program; the findings, then, may reduce uncertainty, offer illumination, inform funders and staff about what participants really experience, enhance communications, and facilitate sharing of perceptions. In early studies of utilization, such uses were overlooked or denigrated. In recent years, they have come to be more appreciated and valued. Studies can be designed with conceptual use in mind or conceptual use may be a by-product or unintended secondary effect of efforts at instrumental use.

Enlightenment as a Form of Utility

Conceptual use is sometimes described as "enlightenment use." Evaluation theorist Carol Weiss first used this term to describe the effects of evaluation findings being disseminated to the larger policy community where they may affect the terms of debate, the language in which debate is conducted, and the ideas that are considered relevant in to resolve.

Process Utility

Process use refers to and is indicated by individual changes in thinking and behavior, and program or organizational changes in procedures and culture, that occur among those involved in evaluation or participatory forms of research as a result of the learning that occurs during the inquiry process. An example of process utility is the following kind of statement after an evaluation: "The impact on our program came not so much from the findings but from going through the thinking process that the evaluation required." Attention to process use is evident in approaches to organizational development that emphasize action research, capacity-building, and organizational learning.

Issues and Controversies Regarding Utility

Defining and Studying Utility

It has proved difficult to identify, measure, and substantiate how particular findings from research or evaluation directly affect decisions. Decisions result from multiple influences and flow from some combination of values, politics, social interactions, personal relationships, perceived risks, perceived benefits, perceived incentives, and knowledge. How much to attribute a decision to knowledge (empirical findings) in any complex decision-making process has proved speculative at best. Moreover, since both knowledge and changes accumulate incrementally over time, disentangling cause and effect relationships within the vagaries of varying organizational memories makes measurement of either instrumental or conceptual use imprecise and subjective.

Definitional Disagreements

Alkin and Hofstetter, in a comprehensive review, concluded that despite 30 years of research on knowledge and evaluation utilization, the definition of "utilization" has never been agreed on. Those who study utility disagree about what should be considered instances of use.

An Interactive Perspective on Utility

Most research on utility has been based on a linear model in which knowledge is generated, is then disseminated, and then influences action, decisions, and thinking. An alternative is an interactive perspective in which researchers and policy makers, or evaluators and information users, interact together and mutually inform one another.

Connecting New Knowledge with Prior Knowledge

Most research on utility focuses on research and evaluation findings as new knowledge. Theories of human cognitive processing add a layer of complexity to studies of how empirical findings are used by positing that individuals use their prior knowledge to assess how to interpret new information. The more persuasive the new information is, in relation to existing conceptions and information, the greater the level of agreement between the two. The more contrary the pieces of information are, the lower the level of agreement. The cognitive reactions, separate from the actual information itself, influence if

and how new information is used. Further, the user's motivation and ability to understand the information are also influential. Motivated, able users who are open to different ideas are more likely to cognitively incorporate new information than are those who are passive, closed-minded, and resistant. Thus, what is convincing and useful social evidence to one person may not be to another.

Factors that Explain and Predict Utility

Factors identified in the research literature on utility generally fall into three main categories. Though conceptually distinguishable, these factors generally interrelate across categories and at different times in affecting utility and actual use.

Human Factors

Human factors affecting utility include user and researcher characteristics, such as people's attitudes toward and interest in the inquiry, their backgrounds and organizational positions, and their professional experience levels.

Context Factors

Context factors include the social, organizational, and political climate for the inquiry, fiscal constraints, and how much empirical findings are valued.

Inquiry Factors

Inquiry factors include how the study is undertaken, its timeline and budget, the procedures and methods used (including their rigor), the information collected including its perceived relevance and credibility), and how findings are reported.

Utility Tests and Truth Tests

In a classic study of prevailing influence, Weiss and Bucuvalas found that decision makers apply both "truth tests" (whether data are believable and accurate) and "utility tests" (whether data are relevant) in deciding how seriously to pay attention to findings. Decision makers want highly accurate and trustworthy data, and they want those data to be relevant to their interests and concerns. The ideal, then, is both truth and utility. In the real world, however, there are often choices to be made between the extent to which one maximizes truth and the degree to which data are relevant.

The simplest example of such a choice is time. The timelines for evaluation can be ridiculously short. A decision maker may need whatever information can be obtained in three months, even though the researcher insists that a year is necessary to get high quality data. This involves a trade-off between truth and utility. Highly accurate data available in a year are less useful to this decision maker than data of less precision and validity obtained in three months.

Decision makers regularly face the need to take action with limited and imperfect information. They prefer more accurate information to less accurate information, but they also prefer some information to no information. This is why research quality and rigor are often only modestly important in determining utility.

Methodological Quality and Utility

The effects of methodological quality on utility must be understood in the full context of a study, its political environment, the degree of uncertainty with which decision makers are faced, and thus their relative need for any and all clarifying information. If information is scarce, then new information, even of dubious quality, may be somewhat helpful. Less rigorous conclusions provided on time can be more useful than highly rigorous findings supplied after a decision has had to be taken. Those are real world trade-offs. In contrast, debates about technical quality are more likely to be center stage in national policy evaluations than in local efforts to improve programs where the stakes are lower.

Efforts to Increase Utility

Much attention has been paid to strategies for increasing utility. Preskill and Caracelli surveyed evaluation practitioners and found several strategies that may enhance evaluation utility: planning for use at the beginning of an evaluation; identifying and prioritizing intended users and intended uses of evaluation; designing the evaluation within resource limitations; involving stakeholders in the evaluation process; communicating findings to stakeholders as the evaluation progresses; and developing a communication and reporting plan.

The Personal Factor

The "personal factor," that is, the extent to which specific decision makers or other primary intended users care about findings, appears from research on utility to be the most important determinant of what impact as well as the type of impact a given study will have. The importance of the personal factor has led evaluators to attempt to enhance use by engaging and involving intended users early in the evaluation, ensuring strong communication between the producers and users of evaluations, reporting evaluation findings effectively so users can understand and use them for their purposes, and maintaining credibility with the potential users. Utilization-focused evaluation is an example of a comprehensive strategy that incorporates the knowledge that has emerged from studying utility.

Utilization-Focused Evaluation

Utilization-focused evaluation is an approach that begins with the premise that evaluations should be judged by their utility and actual use; therefore, evaluators should facilitate the evaluation process and design any evaluation with careful consideration of how everything that is done, from beginning to end, will affect use.

Intended Use by Intended Users

The focus in utilization-focused evaluation is on intended use by intended users. Since no evaluation can be value-free, utilization-focused evaluation answers the question of whose values will frame the evaluation by working with clearly identified, primary intended users who have responsibility to apply evaluation findings and implement recommendations. Utilization-focused evaluation is highly personal and situational. The evaluation facilitator develops a working relationship with intended users to help them determine what kind of evaluation they need. This requires negotiation in which the evaluator offers a menu of possibilities within the framework of established evaluation standards and principles.

Utilization-focused evaluation does not advocate any particular evaluation content, model, method, theory or even use. Rather, it is a process for helping primary intended users select the most appropriate content, model, methods, theory and uses for their particular situation. Utilization-focused evaluation is a process for making decisions about these issues in collaboration with an identified group of primary users focusing on their intended uses of evaluation.

Psychology of Utility

A psychology of use undergirds and informs utilization-focused evaluation: intended users are more likely to use evaluations if they understand and feel ownership of the evaluation process and findings; they are more likely to understand and feel ownership if they have been actively involved; and by actively involving primary intended users, the evaluator is training users in use, preparing the groundwork for utility, and reinforcing the intended utility of the evaluation every step along the way.

See Also the Following Articles

Bentham, Jeremy • Edgeworth, Frances Ysidro • Morgenstern, Oskar

Further Reading

Alkin, M. C. (1985). *A Guide for Evaluation Decision Makers.* Sage, Beverly Hills, CA.

Alkin, M. C. (1990). *Debates on Evaluation.* Sage, Newbury Park, CA.

Alkin, M. C., and Hofstetter, C. H. (2002). Evaluation use revisited. *International Handbook of Educational Evaluation* (D. Stufflebeam and T. Kellaghan, eds.). Kluwer Academic, Boston.

Bearn, A. G. (ed.) (1999). *Useful Knowledge*. American Philosophical Society, Philadelphia, PA.

Havelock, R. (1968). *Bibliography on Knowledge Utilization and Dissemination.* Center for Research on Utilization of Scientific Knowledge, University of Michigan, Ann Arbor.

Nowotny, H. (1990). *In Search of Usable Knowledge: Utilization Contexts and the Application of Knowledge*. Public Policy and Social Welfare, Vol 3. Westview Press, Boulder, CO.

Patton M. Q. (1997). *Utilization-Focused Evaluation: The New Century Text*, 3rd Ed. Sage, Thousand Oaks, CA.

Preskill, H., and Caracelli, V. (1997). Current and developing conceptions of use. *Eval. Practice* **18,** 209–225.

Shulha, L. M., and Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *Eval. Practice* **18,** 195–208.

Weiss, C. H. (ed.) (1977). *Using Social Research in Public Policy Making.* Lexington Books, Lexington, MA.

Weiss, C. H. (1998). Have we learned anything new about evaluation use. *Am. J. Eval.* **1,** 21–34.

Weiss, C. H., and Bucuvalas, M. (1980). Truth tests and utility tests: Decision makers' frame of reference for social science research. *Am. Soc. Rev.* **45,** 302–313.

Validity Assessment

Edward G. Carmines
Indiana University, Bloomington, Indiana, USA

James A. Woods
West Virginia University, Morgantown, West Virginia, USA

Glossary

construct validity A type of validity that is concerned with the relationship between the measure under consideration and theoretical expectations on other measures.
content validity A type of validity that focuses on the extent to which a particular empirical measure reflects a specific domain of content.
criterion-related validity A type of validity that concerns the correlation between a measure and some criterion variable of interest.
measurement The process of linking abstract concepts to empirical indicators of those concepts.
random measurement error All of the chance factors that confound the measurement of any phenomenon.
reliability The extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials.
systematic (nonrandom) measurement error An error that has a systematic biasing effect on a measuring instrument.
validity The extent to which an indicator of some abstract concept measures what it purports to measure.

Validity is concerned with whether a measure actually measures the concept that it is being used to represent. In science, this relationship between the theoretical and the observable is crucial. The unobservable concept is the quantity of interest; how it is measured, or represented, is fundamental to any understanding of the inferences concerning the relationships among the various theoretical concepts.

Measurement

Measurement focuses on the representation of abstract concepts by empirical indicators. Thus, measurement concerns the relationship between abstract, theoretical, and unobservable concepts proposed in a theory and empirical indicators of those concepts for which there are direct observations. As such, measurement involves both theoretical as well as empirical considerations. Empirically, the focus is on the observable response—answers on a questionnaire, observed behavior, answers given to an interviewer, etc. Theoretically, the interest is in the underlying unobservable (and directly unmeasurable) concept that is represented by the response.

Measurement allows the scientist to move from the purely abstract to the empirical and testable. It is centered on the relationship between the empirically grounded indicator, or the observable response, and the underlying unobservable concept. When this relationship is strong, analysis of empirical indicators can lead to useful inferences about the relationships among the underlying concepts. Social scientists can evaluate the empirical applicability of theoretical propositions. If there are no empirical indicators of the theoretical concepts, then the empirical tenability of the theory remains unknown. In situations when the relationship between concept and indicators is weak or faulty, analysis of the indicators can lead to incorrect inferences and misleading conclusions concerning the underlying concepts. Research based on such inadequate measurement models does not result in a greater understanding of the phenomenon under investigation. From this perspective, the auxiliary theory specifying the relationship between concepts and

indicator is as important to social research as is the substantive theory linking concepts to one another.

Basic Properties of Measurement

There are two basic properties of measurement: validity and reliability. Reliability is the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials. It is concerned with the consistency of a measure over repeated observations. Reliability is driven by random measurement error—all of the chance factors that confound the measurement of any phenomenon. If an indicator is a reliable indicator of the theoretical concept, that indicator will produce consistent results on repeated observations because the random error is not great enough to cause notable fluctuation from one observation to the next. Thus, the greater the random error, the less reliable is the measure. For example, if a bathroom scale shows a weight as 5 pounds greater than the true weight on the first reading, 8 pounds greater on the second reading, and 10 pounds less on the third reading, the readings are being affected by random error and the reliability of the scale is low.

Nonrandom, or systematic, error does not detract from a measure's reliability. All of the error in a measure could be nonrandom and it would not affect that measure's reliability. For example, if a bathroom scale always shows a weight as 5 pounds over the true weight, it is a reliable, or consistent, measure. The error in this case is not random; indeed, it is entirely nonrandom—it is 5 pounds too high every time. Thus, the measure is perfectly reliable, but it is not a valid measure. The scale does not measure what it is intended to measure.

Types of Validity

Validity is the extent to which any measuring instrument measures what it purports to measure, rather than reflecting some other phenomenon, i.e., nonrandom measurement error. In testing, the purpose of the test must always be considered when assessing validity. That is, a driver's test may be valid for determining if a person knows how to drive, but it may not be valid at all for determining intelligence. Validity is always an argument between competing theoretical claims. Because of this, what is validated is not the instrument, but rather the instrument in relation to the purpose for which it is being used.

Several types of validity are appropriate in social science research. Each has a slightly different approach in assessing the degree to which a measure is valid. In the body of measurement literature, there are references to internal validity, statistical validity, construct validity, convergent validity, discriminant validity, cross-validation, face validity, concurrent validity, external validity, content validity, sampling validity, criterion validity, predictive validity, and empirical validity. Some of these types of validity overlap. Face validity is sometimes discussed as a separate type of validity and sometimes as a subtype of content validity. Some types of validity are the same, only with different names, e.g., criterion-related and empirical validity are used to mean the same thing. Some are used to denote subtypes of a main type of validity. For example, both concurrent validity and predictive validity are subtypes of criterion-related validity. Convergent validity, discriminant validity, and cross-validation are used to denote types of construct validity. The following discussions cover the three most basic types of validity—content validity, criterion-related validity, and construct validity—and the relevance of each to social science.

Content Validity

Content validity focuses on the extent to which a particular empirical measure reflects a specific domain of content. That is, does the indicator adequately and comprehensively represent what it is supposed to measure? The indicator is said to be content-valid if it reflects the full domain of content. For example, a driver's test that consisted only of right turns and excluded left turns, parking, stopping, and an understanding of traffic signals would not be content-valid. Similarly, and more relevant for the social sciences, a measure of political ideology that included only a question about support for or opposition to government-sponsored medical care would not be valid because it excludes many other policy preferences relevant to ideology.

Obtaining content validity involves two interrelated steps. First, the researcher must be able to specify the entire domain of content that is relevant to a particular measurement situation. In the example of the driver's test, everything that is necessary to know to operate an automobile safely and legally is contained in the state's driver's manual. This is the domain. Specifying the full domain for social science concepts is much more difficult. Minimally, the researcher would need to include questions about a wide array of political issues, certainly issues that go well beyond health care. The second step involves selecting or constructing the specific indicators that are used in the measure. For example, a written driver's test contains a sample of indicators from the driver's manual. In this example, specification and selection procedures are relatively straightforward. However, in the social sciences, this may be quite complex. Specification of the domain of content for abstract concepts such as ideology or alienation is a formidable task. A beginning would be to consult the literature on the subject to gain an understanding of the concept. Once a general understanding of the abstract concept is obtained, indicators that reflect the meaning

of particular aspects of the phenomenon would then be selected or constructed. It is impossible to state a general rule concerning the number of indicators that are necessary to represent any particular domain of content. However, it is always preferable to begin with too many indicators rather than too few, because deficient items can be dropped, but it is much harder to add new or better items at a later stage in the research.

Establishing a content-valid measure of concepts used in the social sciences, such as ideology or alienation, is a very difficult task; indeed, it is much more complex than developing a content-valid measure of driving proficiency. When dealing with abstract concepts, it is difficult to establish concurrence about a domain of content relevant to the phenomenon, because most theoretical concepts in the social sciences have not been described with the required exactness. Further, when measuring most concepts in the social sciences, it is impossible to sample content. A researcher uses one or more indicators that are intended to reflect the content of a given theoretical concept. Without a random sampling of content, however, it is impossible to ensure the representativeness of the particular indicators.

Thus, there are two fundamental limitations of content validity as applied to the social sciences. First, the chosen indicators must reflect the full domain of content relevant to a particular theoretical concept. However, as easy as this may be to achieve with regard to some tests, such as proficiency tests, it is extremely difficult to accomplish for the more abstract phenomena that tend to characterize the social sciences. The second limitation of content validity is the lack of concurrence concerning the criterion for determining the extent to which a measure has attained content validity. This leaves the researcher with the task of having to provide a plausible rationale for accepting a version of what constitutes the appropriate domain of content and for establishing that the indicators included in the measure have been satisfactorily sampled from that domain. Because of these limitations, content validity is not a fully satisfactory means of assessing the validity of social science measures.

Criterion-Related Validity

A second type of validity, more closely related to what is usually meant in everyday usage of the term, is criterion-related validity. This type of validity concerns the relationship or correlation between a measure and some criterion variable of interest. Using criterion-related validity, a driver's test can be validated by demonstrating that the test is a good predictor of the ability of a well-defined group of individuals to drive a car. Criterion-related validity is fully determined by the degree of correspondence between the measure, or test, and its criterion. If the correlation is high, the measure is valid for that criterion. If the test does not correlate significantly with the criterion, it is not valid for that criterion and, thus, is useless for that particular purpose.

The higher the correlation, the more valid is a measure for a specific criterion. For criterion-related validity, this is all that matters. It is the only evidence that is relevant. It does not matter if the test makes no theoretical sense as a predictor of the criterion. If the accuracy of horseshoe pitching, for example, is found to be highly correlated with college success, then horseshoe pitching would be a valid measure for predicting success using criterion-related validity. There is also no single validity coefficient. There are as many coefficients as there are criteria for a particular measure. Technically, criterion-related validity can be differentiated into two types. If the criterion exists in the present, then concurrent validity can be assessed by correlating the measure and the criterion at the same point in time. For example, a verbal report of voting behavior could be correlated with participation in an election, as revealed by official voting records. Predictive validity, on the other hand, concerns a future criterion that is correlated with the relevant measure. Using the Scholastic Aptitude Test (SAT) as a predictor of success in college is an example. Scores on the SAT could be correlated with students' subsequent performance in college to demonstrate the predictive validity of the SAT. The logic of concurrent and predictive validity is the same. The only difference between them concerns the current or future existence of the criterion variable.

What is sometimes overlooked in assessing criterion-related validation procedures is that the scientific and practical utility of criterion validity depends as much on the measurement of the criterion as it does on the quality of the measuring instrument. For example, in many different types of training programs, much effort and expense goes into the development of a test for predicting who will benefit from the program in terms of subsequent job performance. However, the subsequent performance, the criterion, is often given very little attention. Job performance is very difficult to assess. Thus, those using criterion-related validation procedures should provide independent evidence of the extent to which the measurement of the criterion is valid. Although criterion validation is intuitively appealing, it has a major limitation in regard to the social sciences. For many if not most measures in the social sciences, there simply do not exist any relevant criterion variables. For example, to return to the earlier example, it is not clear what an appropriate criterion variable would be for political ideology. Thus, criterion-related validation has limited usefulness in the social sciences. Further, the more abstract the concept, the more difficult it is to find an appropriate criterion for assessing a measure of it.

Construct Validity

The third basic type of validity is construct validity. Construct validity is important when there is no universal agreement concerning the domain of content for the phenomenon, and no relevant criteria. This type of validity focuses on the theoretical expectations surrounding a particular empirical indicator. Thus, construct validity is theory driven. Using theory, the researcher formulates theoretical predictions about the existence, direction, and extent of relations among empirical indicators of different theoretical concepts. If the empirically observed outcomes are consistent with these theoretical predictions, then the measure is said to be construct valid. This type of validity is more pertinent in the social sciences than is either criterion-related validity or content validity.

Construct validation involves three distinct steps. First, the theoretical relationship between the concepts must be specified. Second, the empirical relationship between the measures of the concepts must be examined. Finally, the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure. Again, to return to the earlier example, a researcher might specify that political ideology is related to social class, such that those respondents with lower social status would more likely be leftist in political orientation, whereas those with higher status would more likely be on the right. If income was employed as the measure of social status, and attitudes toward the government provision of health care were used as the measure of political ideology, then the researcher could calculate the relationship between these variables. If this relationship was significant and in the expected direction, this evidence would constitute one piece of evidence supporting the construct validity of the measure of political ideology.

The fundamental feature of construct validation is theory. There must be a theoretical framework about the concept or it will be impossible to validate the measure. Without this theoretical framework, it is impossible to generate theoretical predictions that, in turn, lead directly to empirical tests involving measures of the concept. What is required, then, is to be able to state several theoretically derived hypotheses involving the particular concept. Construct validity is not established by confirming a single prediction on different occasions or confirming many predictions in a single study. Instead, construct validation ideally requires a pattern of consistent findings involving different researchers across a significant portion of time and with regard to a variety of diverse but theoretically relevant variables. Only if and when these conditions are met can the construct validity of a particular measure be spoken of with confidence.

A problem exists if the theoretically derived predictions and the empirical relationships are inconsistent with each other; that is, there is a problem when the evidence relevant to construct validity is negative. The most typical conclusion to draw from negative evidence is that the measure lacks construct validity. That is, the indicator does not measure what it purports to measure—the construct of interest. The accumulation of negative evidence leads to the interpretation that the measure is not construct-valid and should not be used as an empirical manifestation of that concept in future research. Previous research using that particular measure of the concept is also called into doubt. There are other conclusions that are consistent with this sort of negative evidence, however. First, the interpretation could be made that the theoretical framework used to generate the empirical predictions is incorrect, i.e., the theory is wrong. Another interpretation is that the method or procedure used to test the theoretically derived hypotheses is faulty or inappropriate. That is, the statistical technique used in the test could be inappropriate and the researcher could be using it incorrectly. The final interpretation regarding negative evidence is that it is due to the lack of construct validity or due to the unreliability of some other variable in the analysis. It is a very subtle point, but when the construct validity of the measure of interest is assessed, the construct validity of measures of the other theoretical concepts is also being evaluated simultaneously. Thus, it could be the case that the construct validity of the measure is quite high, but the measure hypothesized to correlate with that measure is invalid. In the example of social status, if social status turned out to be unrelated to political ideology, perhaps the problem is that income is not a valid measure of social status.

There is no foolproof procedure for determining which one (or more) of the interpretations of negative evidence is correct in any given instance. The first interpretation, that the measure lacks construct validity, becomes increasingly compelling as grounds for accepting the other interpretations become less tenable. To the extent possible, the construct validity of a particular measure should be assessed in situations characterized by the use of strong theory, appropriate methodological procedures, and other well-measured variables. Only in these situations can it be confidently concluded that negative evidence is probably due to the absence of construct validity of a particular measure of a given theoretical concept. It is apparent that construct validity is the most appropriate and generally applicable type of validity used to assess measures in the social sciences. The researcher can assess the construct validity of an empirical measurement if the measure can be placed in theoretical context. That is, this type of validity, unlike other types, focuses on the extent to which a measure performs in accordance with theoretical expectations.

See Also the Following Articles

Content Validity • Measurement Error, Issues and Solutions • Validity, Data Sources

Further Reading

Adcock, R., and Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *Am. Polit. Sci. Rev.* **95,** 529–546.

Blalock, H. M. (1982). *Conceptualization and Measurement in the Social Sciences.* Sage Publ., Beverly Hills, CA.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* Wiley, New York.

Carmines, E. G., and Zeller, R. A. (1979). *Reliability and Validity Assessment.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-017. Sage, Newbury Park, CA.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Houghton-Mifflin, Boston, MA.

DeVellis, R. D. (1991). *Scale Development: Theory and Applications.* Sage Publ., Newbury Park, CA.

Nunnally, J. C. (1978). *Psychometric Theory.* McGraw-Hill, New York.

Spector, P. E. (1992). *Summated Rating Scale Construction: An Introduction.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-082. Sage, Newbury Park, CA.

Validity, Data Sources

Michael P. McDonald
George Mason University, Fairfax, Virginia, USA

Glossary

concept An abstract object or thought.
contextual specificity Different meanings for the same measure in different contexts.
construct validity The degree of how well a measure fits within existing hypothesized relationships with other measures.
convergent validity The comparison of a measure against one or more measures that are also measures of the same concept, but none holds the distinction of being considered as a direct measure.
criterion validity The comparison of a measure against a single measure that is supposed to be a direct measure of the concept under study.
definition A concept placed into words.
external validity The generalizability of the relationship between two concepts beyond the research question under study.
internal validity The robustness of the relationship of a concept to another internal to the research question under study.
measure Data related to a concept.
operational definition Classification rules for a concept's definition.
reliability The degree by which repeated scoring of a measure provides consistent values.
scoring Gathering information for a measure based on the rules of classification.
validity The degree that an abstract concept is accurately measured.

Validity is the degree by which an abstract concept is accurately measured. Validity is best thought of as a degree, since no variable completely captures an abstract concept. Although this may be a discouraging limitation, much of social science research is driven by the quest for more valid measures. This chapter describes many of the threats to the validity of a measure, tests of validity, and the relationship of validity of a measure to research questions.

Introduction

Hypotheses are tested by projecting abstract concepts onto the real world through measurement and observing the strength of the hypothesized relationships. Validity concerns the component of projecting abstract concepts to the real world through measurement. This projection is best described as an approximation, as there is rarely a one-to-one correspondence of clean abstract concepts to the dirty world of reality. In this context, validity plays an important role in theory building. As subfields within the social sciences mature, competing hypotheses linking concepts arise. Often there are multiple ways of measuring different aspects of the same concept, and these measures may not commonly be used among studies within a subfield. The hypothesis that is ultimately accepted as the best explanation of observed causal relationships may depend upon the widely agreed degree of validity of the measures of the concepts researchers use in their studies.

Validity from the abstract to the real world may be thought of as having four stages, as presented in Fig. 1. In the abstract is the *concept*, the underlying theoretical construct that is to be studied. The *definition* gives the concept meaning though a concrete description. The *operational definition* provides rules of classification to distinguish cases. *Measures*, also called indicators or variables, are generated through the process of *scoring*, or gathering data by following the rules of classification.

The validity of the concept of democracy serves as an example that may help readers grasp the concept of validity itself. Many readers will have an internal

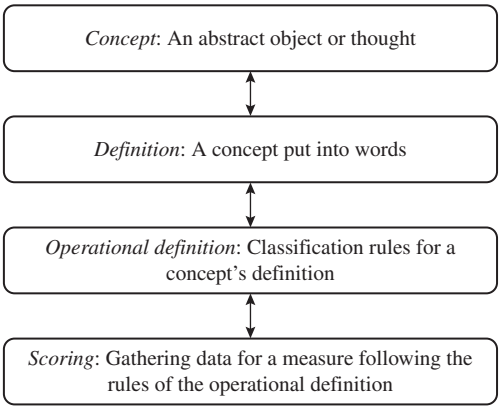


Figure 1 Four stages of validity of measurement, linking an abstract concept to scoring of data.

conception of democracy, even if they have not thought much about it. One widely agreed upon necessary component of the definition of democracy is participation of a people in choosing their government. A measure of democracy within the context of participation might be the percentage of people who choose to vote in elections. Scoring the measure of the voter turnout rate measure requires gathering election returns and the number of persons eligible to vote.

Threats to validity may arise at any of the links between the four stages of validity. First, the concept is defined. Although participation may be a common aspect of a minimal definition of democracy, it is not the only component. Indeed, there is a spirited debate among cross-national researchers as to the minimal definition of a democracy. In an example drawn from Adcock and Collier, political scientists who study Latin America noted that some Central and South American countries scored high on measures of the widely agreed minimal definition of democracy even though the scholars themselves did not believe these countries were fully democratic. Despite participation of citizens in relatively free elections, Latin American democratic governments were not effective because the military still controlled "reserved domains of power." This inconsistency led researchers to amend their definition of democracy to include the requirement that the civilian government must have an effective power to rule.

The measurement of the definition of participation may also be challenged. American scholars have oft-lamented the on-going decline in one measure of participation, voter turnout rates, over the last half of the 20th century. McDonald and Popkin show that the perceived decline in United States turnout rates in recent United States elections turns on the operationalization of the measure of eligible voters. Measures of turnout rates in these previous studies are defined as dividing the number of persons who voted by the voting-age population (VAP), a Bureau of the Census estimate of all persons age 18 and older residing in the United States. The VAP includes persons ineligible to vote, most importantly noncitizens and ineligible felons, who as a percentage of the VAP have increased considerably since 1972. When United States turnout rates are recalculated for those eligible to vote, the decline of turnout rates is relocated to the 1960s, with a relatively flat "trend" since 1972.

The protracted recount of the Florida presidential vote in 2000 exposed the entire country to the reliability of scoring election returns. The Florida presidential election of November 7 was so close, less than 0.5% of the vote, that it invoked an automatic recount of the presidential vote in that state. The overall outcome of the Electoral College, and who would be president, rested in the balance of the Florida recount. Presidential candidate Al Gore contested election returns in certain Florida counties, dragging out the counting and recounting of ballots until the United States Supreme Court decision to halt recounting on December 12, 2002. Most disturbing to the legitimacy of the outcome for either candidate was that each new recount produced a new vote total. "A confounding array of vague laws (and) arbitrary local decisions . . . resulted in turmoil across the state—from the way voters were treated to how ballots were designed and counted." Afterward, with time to carefully reexamine ballots, newspaper organizations hired accounting firms to classify and tally ballots and found that the election outcome still depended on which standard was used to score ballots.

Validity also works in the direction from the bottom of scoring up to the concept, and for this reason the arrows connecting the stages of validity in Fig. 1 point in both directions. Validity may be seen as a dynamic process, whereby definitions of a concept and the rules of measurement are fine-tuned as cases are observed. In the proceeding example, Latin Americanists found the scoring of the concept of democracy did not conform to the perceived degree of democracy in Latin American countries. Researchers reevaluated the minimal definition of democracy and a new definition emerged that included the effectiveness of the democratic government to govern policy areas. Here, the process of scoring led researchers to formulate a new definition.

Validity covers a broad scope linking concept to data, only some of which is touched upon in the proceeding example of democratic participation. With such a broad scope, it is not surprising that one study found 37 different adjectives to describe aspects of validity. Adcock and Collier argue that differentiating the different types of validity detracts from the overriding goal of validity, "... validity must be seen, not as establishing multiple independent *types of validity*, but rather as providing *types of evidence for validity*" (emphasis in original).

Since no abstract concept may be perfectly projected onto the real world, the quest for valid measurement of concepts in the social sciences is one of degree, not of absolutes, where different aspects of validity evidence serve to strengthen the overall validity of a measure, but can never guarantee with absolute certainty the perfect validity of any measure.

Explanation of the different aspects of validity will proceed with this admonition in mind. The discussion is organized around three aspects of validity. In the first section, threats to validity of a measure that may occur between the four stages of concept, definition, measure, and scoring are discussed. In the second section, validity of a measure in relation to other measures of the same concept, or proposed causal relationships between concepts is discussed. These are called *criterion validity*, *convergent* or *divergent validity*, *content validity*, and *construct validity*. It is important to note that with these tests of validity, observed relationships between measures of the same concept or the observed casual relationship between two measures depends on how well the measures are themselves valid measures of the concept they seek to measure. In the final section, validity of data in terms of hypothesis testing, *internal* to the research question and *external* to the application of a hypothesis outside of the research question under study is discussed.

Threats to Validity

Because social science measures do not correspond neatly to the latent concepts to be measured, social scientists are confronted continually with validity issues. Threats to the validity of a measure are posed at the three linkages of a measure to the latent concept with which it is associated. Below, the three points where threats to the validity of a measure may occur are described. A measure may not be valid if (1) an incomplete definition of the concept is proposed, (2) if an inadequate measure is used, or (3) if the scoring of a measure is unreliable.

Formulating the Definition of a Concept

Adcock and Collier distinguish the link between the concept and its definition from the links between rules of measurement and actual scoring of cases; the latter they refer to as measurement validity. They argue that the two have little in common with one another. Formulating a definition of a concept is a theoretical exercise while measurement is a process that lends itself to explicit tests of validity. Indeed, much of what is discussed in this article is within the sphere of what they term measurement validity. Still, a complete discussion of validity warrants a discussion of threats to the link between concept and definition.

In defining a concept, researchers make the first step in approximating the abstract to the real. Some concepts are simple enough that a single, universally accepted definition exists, such as the colors black and white. More often, social scientists operate in a gray area wherein multiple definitions are posited for "contested concepts," such as gender and race.

At first, one may be discouraged to contemplate that there are competing definitions of concepts. Ironically, " . . . it is fortunate that we cannot in reality achieve widely accepted definitions of most constructs." The different definitions provide researchers with the opportunity to tackle the same concept from different angles. The validity of measures, and the validity of causal relationships between measures, may thus be "triangulated" by comparing measures of the same concept derived from different definitions of that concept. These sorts of tests of validity are considered in the next section on validity tests.

Within a contested concept, a consensus definition may emerge among researchers. A proper balance must be struck, such that the scope of the definition is broad enough to quantify the meaning of a concept, but must not be so broad as to include irrelevancies. Well thought out parsimony is the key to the development of a good definition.

Operationalizing Measures

Formulating or operationalizing measures of concepts is the next stage in creating a valid measure of a concept. The relationships between concept and definition, and between definition and measure are very similar. Whereas a definition of a concept is a concept put into words, the measure of a definition is the definition projected onto data, following the rules devised to create the measure. From the introductory example, democracy might be simply defined as persons choosing the government, while a measure of democracy might be the proportion of people who participate in the electoral process.

A primary difficulty in developing measures is context. Consider a cup of coffee. A cup served in European restaurants is very different than what is served in the United States. Americans ordering their first European coffee will be tempted to eat the sludge they are served with a spoon, while Europeans will simply conclude that Americans do not know how to make a strong drink, be it coffee or beer. *Contextual specificity* refers to a measure possessing different meanings in different contexts. Two cases of a measure may possess the same score, but the score may have entirely different meanings in different contexts. Contextual specificity may arise in cross-national

studies, in comparisons across subgroups, in panel studies on polls, and in historical research.

Survey research is particularly sensitive to contextual specificity. Respondents may assign different meanings to the same question. Survey methodologies (beyond random sampling) such as sample design, weighting, interviewing techniques, question wording, and question ordering may vary between surveys. Even seemingly trivial issues such as the order in which questions are asked of respondents may affect their answers. Compilations of surveys, such as the American National Election Survey Cumulative Dataset or the Cumulative General Social Survey, provide illusions of consistency by assembling answers to similar questions asked over time into one variable within a data file, but careful reading of codebooks reveal many of the minutia that may pose a threat to the validity of comparing answers across time and between respondents.

Contextual specificity may arise whenever comparisons are made across cases, be it persons, groups, countries, or time. Since these categories cover virtually all social science research, it is important for researchers to be aware of the potential for contextual specificity and take corrective measures, if warranted and possible. For example, American politics researchers who make comparisons between U.S. states often distinguish the special nature of Southern politics by including a dummy variable for that region in their statistical analyses. (The concept of the "South" is itself a contested concept.) This definition is operationalized and then interacted with other variables. In doing so, these researchers attempt to "establish equivalence" of their measures across regions.

Equivalence may also be established when the measures themselves are operationalized and scored, by carefully considering what important components are necessary to measure the concept to be studied, making sure that the full scope of components are indeed measured. For example, a study of civic participation that neglects activities available through the internet would be deficient, especially one that compares participation across time.

Content validity refers to the scope of the measure accurately capturing the definition of a concept. Just as with developing a definition for concept, researchers must strike a compromise between completeness and parsimony when developing measures. While it is desirable to capture the full scope of the definitional concept, including the beating wing of a butterfly into the operationalization of a measure is too broad, and thereby the meaning of the concept one is trying to capture with the measure is blurred. The continual drive of social science research is toward completeness, as most concepts have at some point in time been already measured at least once. The challenge for future researchers is to refine the measurement of concepts without overly increasing their complexity.

The content validity of social sciences measures is a subjective matter. As a branch of study develops, widely agreed upon measurement of the definitional concept may arise. That does not mean that there is not room for improvement. By cataloging the operationalization of the concept of democracy used by leading researchers, Paxton noted that the scholars used male suffrage as the operationalization of universal suffrage. Paxton reformulated the operationalization of suffrage to include female suffrage and was rewarded with different results. This is a lesson that young researchers should heed, as finding these sorts of errors of omission often prove fertile ground for groundbreaking research.

Reliability of Scoring

Validity is often distinguished from *reliability*, the degree by which repeated scoring of a measure provides consistent values. Alwin provides a detailed treatment of reliability in the *Encyclopedia for Social Measurement* but a brief discussion is warranted here because reliability is a necessary condition for validity. If repeated attempts to score the cases of a measure yield dramatically different results, then validity may be impossible to ascertain.

Virtually all social science data have measurement error. The naïve researcher accepts all data at face value while the cynical researcher never fully trusts that any data source reports the exactly correct values for all cases. A classic example of measurement error is the statistical "margin of error" of polling percentages associated with random sampling of a population. Because of random selection of respondents, no two surveys will produce the same result, and no survey is guaranteed to represent the true value. Instead, margins of error are reported to indicate where we have confidence in where the true value may lie. Measurement error extends well beyond surveys to any setting where human error in coding and classifying cases may occur. Some cases of a measure may even be missing entirely.

The reliability of a measure is determined by the size and bias of the measurement error. If measurement error is large, then the noise in the scoring of the measure may make any meaningful interpretation of the measure in relationship to the concept, or other concepts, impossible to determine. In the context of polling, the margin of error of a poll is inversely related to the sample size of the survey. If the sample size is small, then the margin of error will be large, so large that the location of the true value will fall within such as large range as to be virtually unknown. In a similar vein, if the poll result is within the margin of error of the thing that one wishes to determine,

such as the winner of an election, then the outcome is said to be within the margin of error of the poll, and impossible to determine with confidence.

Measurement error may be small but still have bias, systematic measurement error that may also invalidate a measure's usefulness. For example, in 1932, *Literary Digest* magazine conducted a poll of over two million respondents selected from telephone directories and automobile registrations. The poll predicted a landslide victory for Alf Landon, but the landslide that year went to Franklin D. Roosevelt. Even with two million respondents, the poll was unrepresentative of the population, since only the most affluent owned phones and cars at that time. Thus, even though the margin of error of this poll would have been relatively smaller than usual surveys, because of the extremely large sample size, the survey was still unreliable because it was biased.

Biases may appear in many guises. In polling, it is well known that respondents may lie to the interviewer to provide the socially correct response. For example, more respondents consistently report voting than official government figures indicate. Sociologists term the dual nature of bias that enters when humans observe human behavior as the insider versus the outsider biases. Sociologists who seek an insider perspective attempt to gain the trust of those that one studies by becoming active participants in the subject they are studying. As an insider, the sociologist hopes that the subjects are more willing to provide their true feelings on a subject to the researchers. However, the gap between researcher and subject may never be fully gulfed, as subjects will always know that they are being studied. Furthermore, a researcher may never be able to achieve insider status, such as a person studying different races. In becoming an insider, the sociologist gains the trust of the study subjects, but at the same time loses the perspective of detachment. As an outsider, the sociologist is able to see the large picture and place the observations into a larger meaningful context. A careful balance of insider insight and outsider perspective is needed to produce meaningful sociological research.

Since polls rely on random sampling, a poll may, by just sheer bad luck of random chance, draw an aberrant sample. The margin of error refers to the 95% confidence interval of a poll, and provides a false sense of reliability of a poll since one out of twenty times the true value will lie outside the confidence interval. During the course of a presidential election, many polls, much more than 20, are conducted. Unfortunately, the surprise poll, the one with a surprising result that may be the result of a bad sample, is the one that receives the most attention. To turn statistics on its head, 1 out of every 20 statistical results may incorrectly reject the null hypothesis merely as a consequence of random chance.

Although polls provide classic examples of reliability issues, by no means is reliability an issue restricted to polling. For example, Ward chronicles how research into the retirement of U.S. Supreme Court justices was fatally flawed because the first study of retirement incorrectly reported the text of a 1937 law that set the age at which judges may retire with benefits. Subsequent researchers cited the original study, without checking the statute itself, and propagated the error. The lesson is that it is important to check the reliability of secondary sources.

In the course of data entry, inevitably cases will be incorrectly scored through human error. These sorts of data entry errors are may appear as *outliers*, atypical cases with values far from other cases. Outliers are not necessarily the product of an error, but they often are, and they can severely affect observed relationships between measures. Researchers should carefully check the extreme and highly implausible values of a measure, as these cases are typically the result of data entry errors. For example, in analyzing election data for a project, one election outcome had an exact 50–50 split of the vote between two candidates. Since an exact tie is unlikely, checking the observation revealed the vote total for one candidate had been incorrectly entered twice, for both candidates. Some of the same techniques to test the validity of a measure may also reveal outliers. For example, plotting two related measures against one another will reveal deviant cases. To improve the reliability of a measure, when a new data entry project has been completed a careful researcher will perform checks for outliers and check their validity. Just as importantly, data obtained from even the most respected outside sources should not be blindly accepted as error free, and should be similarly scrutinized.

Data may also be missing: randomly or with bias. Statisticians have developed methods to impute, or fill in, missing data and incorporate the statistical error of imputation into results. Imputation of missing data is a statistical guess for the true values, and thus it inherently contains measurement error in the resulting scores for missing cases. The loss of reliability associated with imputation will be proportional to the number of missing cases and the reliability of the imputation procedure, which itself may also contain bias.

Reliability is also an issue for statistical software. The world of pencil and paper does not directly translate into the binary arithmetic of computers. Silent, small measurement error is introduced when numbers, particularly fractions, are entered into statistical software programs. Even for perfectly valid measure, if such a thing existed, these errors may propagate through statistical algorithms to produce wildly inaccurate statistical results. An understanding of the limits of the computers can help researchers from avoiding these undesirable results.

Summary

As researchers move from the abstract concept to the actual scoring of measures they operationalize, threats to the validity of the measure arise. No social science measure is a perfect indicator of the latent concept it purports to measure. Validity should be considered as a degree, not a dichotomous "valid" or "invalid" label. Validity of measures is constrained by degree of abstractness of the concept. Validity of measures for extremely abstract concepts, such as "love," may be low, while concepts that are well defined and linked to empirical data will have a greater degree of validity.

The flow from the abstract to the real through the process of measurement is not a one-way street, either; it is a dynamic process. An observant researcher will, in the course of scoring cases, refine the operationalization of measures to account for cases that do not neatly fit the typology of the rules of the measure. This may even lead researchers to refine the working definition of the concept. Since many measures that have been formulated by researchers are available in electronic form, the careful process of scoring through data entry may be deemed inefficient by the impatient researcher. But, it is an art form that should not be undervalued, especially in the early stages of research.

Validity Tests

In a perfect world, perhaps, there would be one measure that would be universally accepted as the valid measure for each concept. Inevitably, social science concepts are open to the threats to validity discussed in the previous section. Fortunately, the limitation of imperfection is leveraged into an opportunity in the social sciences to improve research by developing tests for the validity by comparing measures of the same concept or related concepts.

In isolation, the validity of a single measure cannot be guaranteed. This poses problems for emerging research, such as polling of the attitudes of minority groups such as Hispanics and Asians, who until only recently, had a large enough population for which to formulate sampling schemes. The first studies of Hispanic attitudes found less support for immigration policies than their political leaders had expected, leading them as question the validity of the polling. Subsequent polls found the same results.

As a field of research evolves, researchers begin to formulate more indicators to measure the concept they are studying. When subsequent polls of Hispanic attitudes found similar results to the first poll, the overall validity of this finding was strengthened. Similarly, as new measures are developed in other fields, the validity of the new measures, and even remeasurement of existing indicators, may be ascertained by comparing measures against one another.

Three types of validity tests will be considered in this section. *Criterion validity* is the comparison of a measure against a single measure that is supposed to be a direct measure of the concept under study. *Convergent validity* is the comparison of a measure against one or more measures that are also measures of the same concept, but none holds the distinction of being considered as a direct measure. A measure is considered to be valid if, by these tests, the measure is positively correlated with the other measures. If a negative correlation is found, the test offers existence of discrimination. *Construct validity* is final evolution in the tests of validity; it is the degree of how well a measure fits within existing hypothesized relationships with other measures. Here, the normal context of hypothesis testing in turned on its head. The hypothesis is assumed to be correct, and the measure itself is validated by how well it fits within the existing hypothesized theoretical framework.

Criterion Validity

Criterion validity is the comparison of a measure against a single measure that is supposed to be a direct measure of the concept under study. Perhaps the simplest example of the use of the term validity is found in efforts of the American National Election Study (ANES) to validate the responses of respondents to the voting question on the post-election survey. Surveys, including the ANES, consistently estimate a measure of the turnout rate that is unreliable and biased upwards. A greater percentage of people respond that they voted than official government statistics of the number of ballots cast indicate.

To explore the reliability of the measure of turnout, ANES compared a respondent's answer to the voting question against actual voting records. A respondent's registration was also validated. While this may sound like the ideal case of validating a fallible human response to an infallible record of voting, the actual records are not without measurement error. Some people refuse to provide names or give incorrect names, either on registration files or to the ANES. Votes may be improperly recorded. Some people live outside the area where surveyed and records were left unchecked. In 1984, ANES even discovered voting records in a garbage dump. The ANES consistently could not find voting records for 12–14% of self-reported voters. In 1991, the ANES revalidated the 1988 survey and found 13.7% of the revalidated cases produced different results than the cases initially validated in 1989. These discrepancies reduced the confidence in the reliability of the ANES validation effort and, given the high costs of validation, the ANES decided to drop validation efforts on the 1992 survey.

The proceeding example is of *criterion validity*, where the measure to be validated is correlated with another measure that is a direct measure of the phenomenon of concern. Positive correlation between the measure and the measure it is compared against is all that is needed for evidence that a measure is valid. In some sense, criterion validity is without theory. "If it were found that accuracy in horseshoe pitching correlated highly with success in college, horseshoe pitching would be a valid measure of predicting success in college" (Nunnally, as quoted in the work of Carmines and Zeller). Conversely, no correlation, or worse negative correlation, would be evidence that a measure is not a valid measure of the same concept.

As the example of ANES vote validation demonstrates, criterion validity is only as good as the validity of the reference measure to which one is making a comparison. If the reference measure is biased, then valid measures tested against it may fail to find criterion validity. Ironically, two similarly biased measures will corroborate one another, so a finding of criterion validity is no guarantee that a measure is indeed valid.

Carmines and Zeller argue that criterion validation has limited use in the social sciences because often there exists no direct measure to validate against. That does not mean that criterion validation may be useful in certain contexts. For example, Schrodt and Gerner compared machine coding of event data against that of human coding to determine the validity of the coding by computer. The validity of the machine coding is important to these researchers, who identify conflict events by automatically culling through large volumes of newspaper articles. As similar large-scale data projects emerge in the information age, criterion validation may play an important role in refining the automated coding process.

Convergent Validity

Frequently, there are a number of competing measures that are posited as viable indicators of a given concept, but none is given the special status as a direct measure that may be used as a reference to make use of criterion validation. *Convergent validity* is the comparison of different measures for the same definitional concept or different definitional concepts, respectively. When a measure correlates with other measures that have been posited to measure the same definitional concept as the measure in question, this is taken as evidence for convergent validity. Conversely, lack of correlation between measures of different concepts is evidence of discrimination that the tested measure is not related to the other measures analyzed.

Tests of convergent (or divergent) validity will only be as good as the validity of the measures that are being tested. If the measures are weakly valid, then correlations among them may also be weak, unless by chance the biases or randomness in the measures align. Tests of convergent validity are particularly difficult for comparative researchers, who often have only a small number of cases to analyze, and thus the variability of measurement is not evened out by the law of large numbers.

Often, it is understood that a single measure is not a perfect indicator of a concept. Instead, a bundle of indicators is used to develop a measure of a latent concept. For example, the measurement of the concept of political knowledge would not rely on the answer by a survey respondent to a single question. A more accurate measure would incorporate answers to several questions, summed to form an index of political knowledge.

The problem with indexes is that all items in the index may not be equal. Should the answer to "Who is President?" be given the same weight as the question "Who is Postmaster General?" A related concern is that some of the proposed indicators used to construct an index may discriminate from one another. Respondents may or may not have general political knowledge as well as specialized knowledge in specific subjects that matter to them. Further, the list of measures used in such an index must have content validity, lest an important indicator is missing from the measurement. This problem applies beyond survey research, as researchers are often faced with the task of condensing several measures into a single measure fitting on a scale, such as how democratic a country is based on several measures of democracy.

A number of methods have been proposed to determine the convergent validity of measures, and to determine how measures that correlate should be weighted, for example, Chronbach's alpha. Beyond simply constructing indexes, researchers often rely upon factor analysis to estimate the convergent validity between different indicators and develop measures of latent concepts among those indicators that correlate. Factor analysis essentially attempts to identify similar measures and fit them onto a single quantifiable measure of a latent concept. Factor analysis may find more than one grouping and combine discriminated groups of indicators into multiple "dimensions" of different latent concepts. More sophisticated methods exist, such as Poole and Rosenthal's NOMINATE procedure to estimate measures of political ideology of legislators from legislative roll call votes. Similar Bayesian methods have been proposed to cover a wider scope by combining multiple measures into one measure of a latent concept.

Construct Validity

Tests of convergent validity assume that the measures under study are related to the same concept. It may be that two measures correlate not because the two capture the same concept, but that casual relationships exist that drive the correlation. For example, Poole and Rosenthal's

measure of the ideology of members of Congress based on their roll call votes is highly correlated with measures of party unity of the members on roll call votes. The most liberal and conservative members of the legislative parties tend to vote strongly with their party while those in the ideological center may occasionally cross over and vote with the other side. Do these measures capture the same concept of ideology, does one cause the other, or are both caused by some other factor?

Construct validity refers to how well a measure is associated with measures of other latent concepts that are theorized to have causal relationships, or constructs, with one another. The normal rules of inference are turned on their head. Instead of assuming the measure is correct and testing the causal hypothesis, the causal hypothesis is assumed to be correct and the validity of the measure is tested. If the relationships among the measures perform as expected according to the construct, then the measure is deemed to possess construct validity.

An example of construct validation is drawn from Adcock and Collier. Lijphart uses construct validation to justify his classification of India as a consociational democracy, a democracy where all ethnic groups are provided with institutional roles in the government. He first classifies India as a consociational democracy using the criteria he associates with his definition of the concept of consociationalism. He then identifies causal factors that are hypothesized to produce consociational democracies, and notes their presence in India, providing evidence of the construct validity of his classification of India.

The information age provides the fingertips of researchers with numerous measures developed by other researchers developed in the course of their inquiry. Care should be taken in importing precompiled measures into a new line of research, as a measure developed by one researcher may not have construct validity in another line of inquiry. For example, Epstein and Mershon show how Segal and Cover measures developed to represent the political preferences of Supreme Court justices in civil liberties cases may not be valid for judicial processes outside of civil liberties.

Summary

There are limits to the testing of measures against one another to ascertain their validity. First, the measures must be valid indicators of the concepts they purport to measure. Unfortunately, this is exactly the question one is trying to determine through validity tests. If the measures one is testing against are flawed, then the test itself will usually be similarly flawed. Second, if statistics are used, the proper statistical assumptions must be made commensurate with data and hypothesized relationships, and the statistical tools must be reliable enough to correctly estimate the relationships. Finally, for construct validity, the hypothesized theoretical constructs must also be "valid" in the sense that they hypothesize the correct relationships.

Internal and External Validity

The causal relationship of one concept to another is sometimes also discussed in terms of validity. *Internal validity* refers to the robustness of the relationship of a concept to another internal to the research question under study. Much of the discussion in the section under threats to validity and the tests for validity is pertinent to the internal validity of a measure, vis-a-vis another concept with which it is theoretically correlated. *External validity* refers to the greater generalizability of the relationship between two concepts under study. Is the uncovered relationship applicable outside of the research study?

The relationship between one measure and another may be a true relationship, or it may be a spurious relationship that is caused by invalid measurement of one of the measures. That is, the two measures may be related because of improper measurement, and not because the two measures are truly correlated with one another. Similarly, two measures that are truly related may remain undetected because invalid measurement prevents the discovery of the correlation. By now, the reader should be aware that all measures are not perfectly valid, the hope is that the error induced in projecting theory onto the real world is small and unbiased so that relationships, be they findings that two measures are or are not correlated, are correctly determined.

All of the threats to validity apply to the strength of the internal validity of the relationship between two measures, as the two measures must be valid in order for the true relationship between the two, if any exists, to be determined. Much of the discussion of tests of content and convergent validity also applies to internal validity. In addition, researchers should consider the rules of inference in determining if a relationship is real or spurious. Are there confounding factors that are uncontrolled for driving the relationship? A classic example in time-series analysis is cointegration, the moving of two series together over time, such as the size of the population and the size of the economy, or any other measure that grows or shrinks over time. In the earlier example of voter turnout, the confounding influence of a growing ineligible population led researchers to incorrectly correlate a largely invalid measure of decreasing voter turnout to negative advertising, a decline of social capital, the rise in cable television, campaign financing, the death of the World War II generation, globalization, and decline in voter mobilization efforts by the political parties.

External validity refers to the generalizability of a relationship outside the setting of the study. Perhaps

the most distinguishing characteristic of the social sciences from the hard sciences is that social scientists do not have the luxury of performing controlled experiments. One cannot go back in history and change events to determine hypothetical counterfactuals, while physicists may repeatedly bash particles together and observe how changing conditions alter outcomes. The closest the social sciences come to controlled experiments is in laboratory settings where human subjects are observed responding to stimuli in controlled situations. But are these laboratory experiments externally valid to real situations?

In a classic psychology experiment, a subject seated in a chair is told that the button in front of them is connected to an electric probe attached to a second subject. When the button is pushed an increasing amount of voltage is delivered. Unknown to the subject, the button is only hooked to a speaker, simulating screams of pain. Under the right circumstances, subjects are coerced into delivering what would be fatal doses of voltage.

Such laboratory experiments raise the question as to whether in real situations subjects would respond in the similar manner and deliver a fatal charge to another person, i.e., is the experiment externally valid? Psychologists, sociologists, political scientists, economists, cognitive theorists, and others who engage in social science laboratory experiments painstakingly make the laboratory as close to the real world as possible in order to control for the confounding influence that people may behave differently if they know they are being observed. For example, this may take the form of one-way windows to observe child behavior. Unfortunately, sometimes the laboratory atmosphere is impossible to remove, such as with subjects engaged in computer simulations, and subjects are usually aware prior to engaging in a laboratory experiment that they are being observed.

External validity is also an issue in forecasting, where relationships that are based on observed relationships may fail in predicting hypothetical or unobserved events. For example, economists often describe the stock market as a random walk. Despite analyst charts that graph levels of support and simple trend lines, no model exists to predict what will happen in the future. For this reason, mutual funds come with the disclaimer, "past performance is no guarantee of future returns." A successful mutual fund manager is likely to be no more successful than another in the next business quarter.

The stock market is perhaps the best example of a system that is highly reactionary to external shocks. Unanticipated shocks are the bane of forecasting. As long as conditions remain constant, modeling will be at least somewhat accurate, but if the world fundamentally changes then the model may fail. Similarly, forecasts of extreme values outside the scope of the research design may also fail, or when the world acts within the margin of error of the forecast then predictions, such as the winner of the 2000 presidential election, may be indeterminate.

Conclusion

It is worth emphasizing again that no indicator is a perfectly valid indicator of the concept one is attempting to measure. Abstract thought is not cleanly translated to the complexities of the real world. Awareness of the potential pitfalls of validity, outlined in the first section of this article, help researchers to avoid these problems. Still, no measure will be perfect. The tests in the second part of this article help researchers to have some confidence that their measures are indeed valid.

Ironically, the social science profession exists, partially, because no measure is a perfect measure. This limitation of completely valid measurement provides researchers with opportunities to attack theories from different angles, rather than settling on one absolute truth. A concluding word of advice: a healthy dose of skepticism regarding the validity of the measures that one may use in the course of research may unexpectedly open new pathways to viewing the world one studies and hopes to explain.

See Also the Following Articles

Reliability • Validity Assessment

Further Reading

Altman, M., Gill, J., and McDonald, M. P. (2003). *Statistical Computing for the Social Sciences.* Wiley, New York.

Anderson, B. A., and Silver, B. D. (1986). Measurement and mismeasurement of the validity of the self-reported vote. *Am. J. Polit. Sci.* **30,** 771–785.

Babington, C. (2000). Bush claims presidency as Gore concedes. *The Washington Post*, December13.

Burden, B. C. (2000). Voter turnout and the national election studies. *Polit. Anal.* **8,** 389–398.

Campbell, A., Converse, P. E., Miller, W. E., and Stokes, D. E. (1960). *The American Voter.* University of Chicago Press, Chicago.

Carmines, E. G., and Zeller, R. A. (1979). *Reliability and Validity Assessment.* Sage, Beverly Hills, CA.

Clausen, A. R. (1967). Response validity: Vote report. *Public Opin. Quart.* **32,** 1–38.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-experimentation: Design and Analysis Issues for Field Services.* Houghton-Mifflin, Boston.

Cox, G. W., and McCubbins, M. D. (1993). *Legislative Leviathan: Party Government in the House.* Cambridge Press, New York.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* **16,** 297–334.

Delli Carpini, M. X., and Keeter, S. (1993). Measuring political knowledge: Putting first things first. *Am. J. Polit. Sci.* **37,** 1179–1206.

Epstein, L., and Mershon, C. (1996). Measuring political preferences. *Am. J. Polit. Sci.* **40,** 261–294.

Gill, P. E., Murray, W., and Wright, M. H. (1981). *Practical Optimization.* Academic Press, San Diego.

Herbert, A. (1998). *Polling and the Public: What Every Citizen Should Know,* 4th Ed. Congressional Quarterly Press, Washington, DC.

Keating, D., and Mintz, J. (2001). From election audit, mostly uncertainty: Miami Herald review shows results hinges on standard used in recount. *The Washington Post,* April 5.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton University Press, Princeton, NJ.

Lijphart, A. (1996). The puzzle of Indian democracy: A consociational interpretation. *Am. Polit. Sci. Rev.* **90,** 258–268.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

McDonald, M. P. (2003). On the over-report bias of the national election study. *Polit. Anal.* **11**.

McDonald, M. P., and Popkin, S. (2001). The myth of the vanishing voter. *Am. Polit. Sci. Rev.* **95,** 963–974.

Merton, R. K. (1972). Insiders and outsiders: A chapter in the sociology of knowledge. *Am. J. Sociol.* **78,** 9–47.

Milgram, S. (1965). Some conditions of obedience and disobedience to authority. In *Current Studies of Social Psychology* (I. D. Steiner and M. Fishburn, eds.), pp. 243–262. Holt Rinehart and Winston, New York.

Mintz, J., and Slevin, P. (2001). Human factor was at core of vote fiasco: Decisions and leadership were erratic, arbitrary. *The Washington Post*, June 1.

Poole, K. T., and Rosenthal, H. (1997). *Congress: A Political-Economic History of Roll Call Voting.* Oxford University Press, New York.

Schrodt, P. A., and Gerner, D. J. (1994). Validity assessment of a machine-coded event set data for the Middle East, 1982–1992. *Am. J. Polit. Sci.* **38,** 825–854.

Segal, J. A., and Cover, A. D. (1989). Ideological values and the votes of U.S. Supreme Court justices. *Am. Polit. Sci. Rev.* **83,** 557–565.

Traugott, M. W., and Katosh, J. P. (1979). Response validity in surveys of voting behavior. *Public Opin. Quart.* 359–377.

Traugott, M. W., and Presser, S. (1992). *Revalidation of Self-Report.* American National Election Study Technical Report No. 42.

Traugott, S. (1989). *Validating Self-Reported Vote: 1964–1988.* American National Election Study Technical Report No. 34.

Trier, S., and Jackman, S. (2002). Beyond factor analysis: Modern tools for social measurement. Presented at the 2002 Midwest Political Science Association, Chicago.

Uhlaner, C. J., and Garcia, F. C. (2002). Latino public opinion. In *Understanding Public Opinion* (B. Norrander and C. Wilcox, eds.), 2nd Ed., pp. 77–102. Congressional Quarterly Press, Washington, DC.

Ward, A. (2002). How one mistake leads to another: A research note on the importance of verification/replication. Presented at the 2002 Midwest Political Science Association Meeting.

Weaver, C. N., and Swanson, C. L. (1974). Validity of reported date of birth, salary, and seniority. *Public Opin. Quart.* **38,** 69–80.

Web Hyperlink Analysis

Liwen Vaughan
University of Western Ontario, London, Ontario, Canada

Glossary

external link (also called external inlink) An inlink coming from a Web site outside the site being linked to. See the example of Total Link below for further explanation.
inlink (also called backlink) A link coming into a Web page. See the example of Outlink below for further explanation.
internal link (also called as selflink) An inlink coming from a Web page that is on the same site as the page being linked to. For example, the "back to home" link from one of the pages on a site to the home page of the site. See the example of Total Link below for further explanation.
outlink (also called outgoing link) A Link going out from a Web page. Suppose that Web page X has some kind of relationship with Web page Y so that page X has a link pointing to page Y. This link will be an inlink for page Y and an outlink for page X.
total link (also called all link) All inlinks to a Web page regardless of the origin of the links. It is the sum of internal link and external link (see above). For example, page X receives 50 inlinks from other Web sites (external links). Meanwhile, there are 20 internal links (links coming from other pages of this site to page X). Then there are 50 external links, 20 internal links, and 70 total links.

One of most important characteristics, in fact the defining feature, of the Web is the hyperlinks embedded in Web pages. It is these hyperlinks that join otherwise separated Web sites into an interconnected Web. Given their importance, Web hyperlinks have been studied by researchers from various disciplines including computer science and sociology, as well as information science. This article focuses on the social science aspects of Web hyperlinks. It traces the history of hyperlink analysis to citation analysis, a mature area of information science. It then introduces data collection methods for hyperlink analysis. Finally, it reports important findings from hyperlink studies, including both link count and link topology studies. Research in this area is developing rapidly as the Web is still evolving. Findings and theories reported in this article are not as "time tested" as those in other, more mature, areas of social sciences. However, this does not undermine the significance of what is reported, it only highlights the need for an open mind. The constantly changing nature of the Web provides both challenges and opportunities for future research in this area.

From Citation Analysis to Web Hyperlink Analysis

The astonishing development of the World Wide Web (the Web for short) that started in the 1990s and is still continuing in the 2000s propelled the research into this increasingly important area of the modern society. One of the key features that sets Web pages apart from traditional documents is the hyperlinks (links for short) embedded in the Web pages. Indeed, it is these hyperlinks that tie individual Web sites into an interconnected Web. Users of the Web follow hyperlinks to visit related sites while search engines use these hyperlinks to discover new Web sites. Web hyperlinks are significant in a number of ways. The more links to a Web site, (1) the more visible the site is on the Web; (2) the more potential traffic there will be to the site; (3) the more likely it is that the site will be covered by search engines; and (4) the higher the site will be ranked in search results. In addition, there is the "success breeds success" effect in that a site with more links will potentially attract even more links because of its visibility. In short, Web hyperlinks can have social, political, and even economic power.

Since hyperlinks play such an important role on the Web, researchers from a number of disciplines have studied them. For example, computer scientists have investigated ways to use link information for search engine development. Google pioneered the link-based ranking algorithm in which link counts are used in ranking search results. Other major search engines adopted the concept and used link counts in various ways. Meanwhile, social researchers such as sociologists and interdisciplinary researchers such as information scientists examined hyperlinks as a social phenomenon. This article naturally focuses on the social science aspects of Web hyperlinks. It should be noted that research efforts from various disciplines are not mutually exclusive but in fact are mutually beneficial and can inform one another.

While Web hyperlink analysis is a new research area given the relatively short history of the Web, it has roots in citation analysis, a fairly mature area of information science that has regular coverage by major academic journals such as *Nature*. Citation analysis studies the nature of citations that appear in scientific or scholarly literatures. These citations, in the form of references cited in papers, show the explicit linkages between current research and prior work. The main data sources for citation analysis are citation indexes that are produced by the Institute for Scientific Information founded by Eugene Garfield. The commercial success of citation indexes is evidence of the widespread utility of citation-based information. A compelling body of research findings supports the theory that citations can be used as valid indicators of quality, utility, or impact of cited papers although, like other social measurements, they are not a perfect measurement. Citation data have been used to successfully (1) identify influential scholars; (2) map relationships among academic fields; and (3) predict movements of fields and trends in research.

Even in the early years of the Web, researchers pointed out the similarity between citations and Web hyperlinks. The first use of the neologism "sitation" was probably made by Gerry McKiernan on his Cited Sites Web page http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm. Rousseau defined the term (a combination of the words "citation" and Web "site") more explicitly in the context of Web hyperlink analysis. Cronin analyzed in detail the relationship between citation analysis and Web hyperlink analysis and discussed the validity and reliability issues surrounding the link analysis. Ingwersen proposed a new measure, the Web Impact Factor, modeled on the Journal Impact Factor in citation analysis. The Journal Impact Factor is a measure of the frequency with which the "average article" in a journal has been cited in a particular year or period. It is calculated to be the number of current year citations to the citable items published in that journal during the previous two years divided by the number of those items. The Web Impact Factor of Web site X is defined as the number of Web pages with at least one link pointing to Web site X divided by the number of pages in Web site X. The parallel between the Journal Impact Factor and the Web Impact Factor is clear: the former measures the impact of a journal while the latter does the same for a Web site.

Many studies of Web hyperlinks, including those in computer science, draw the analogy between citations and hyperlinks and acknowledge the intellectual inspiration from citation analysis. However, there are also differences between citations and Web hyperlinks. One major difference is that an academic citation represents a more formal connection between the citing and cited articles. Citations, especially those in refereed journals, have quality control. In contrast, Web links can be created by anyone for any reason without any control. It may seem that Web links are random and may not contain much useful information. However, studies have shown that the number and the pattern of Web links can reveal useful information when aggregated over large area of the Web, as described below.

Data Collection for Web Hyperlink Analysis

The two main tools used in data collection for quantitative analysis of Web hyperlinks are Web crawlers and commercial search engines. Qualitative studies of Web hyperlinks use data collection methods that are commonly used in social science research. For example, personal interviews have been used to investigate link creation motivations, while content analysis has been used to classify types of Web links. These methods are covered elsewhere in the Encyclopedia so they will be not repeated here. The two methods for collecting quantitative data are described below.

Using a Web Crawler

A Web crawler is a computer program that automatically traverses Web hyperlinks by retrieving a Web page, then recursively retrieving all Web pages to which that page is linked. A Web crawler can be used to find and record the hyperlinks on the set of Web sites that are being studied. These recorded links can then be analyzed in various ways to discover patterns of links on these sites, e.g., what kind of sites are being linked to (commercial sites vs government sites; domestic sites vs international sites). You can also find out how the sites in the study interconnect to each other. While generic Web crawlers are available for downloading free of charge, specialized Web crawlers have been developed for research projects.

Using Commercial Search Engines

Major commercial search engines, such as Google (www.google.com) and AllTheWeb (www.alltheweb.com), have the capability of searching for links pointing to a particular Web site. For example, the search query "link:www.abc.com" in Google will find links pointing to Web site www.abc.com (i.e., inlinks to site www.abc.com). The search result not only shows the total number of links (e.g., 100) pointing to the site in question but also lists all these 100 sites so that users can visit each site to see under what circumstances the links are being made. Different search engines use different query syntax for link searching. The syntax can be easily found on the search engine's FAQ sheet or "help" page. Some search engines such as AllTheWeb and AltaVista (www.altavista.com) can further distinguish between external links and total links (see Glossary), which is an extra function that is very useful for Web hyperlink analysis.

Comparison of the Two Data Collection Methods

The main advantage of using commercial search engines for data collection is that these engines typically cover a much larger area of the Web (i.e., index more Web sites) than a Web crawler for a particular research project can do. However, the crawling algorithms of commercial search engines are typically proprietary and are not revealed to the general public. The crawling algorithm affects what Web sites get covered and what portion of a site (what pages on a site) gets covered. In contrast, a Web crawler designed for research purposes can be programmed to crawl in the way that is most appropriate for the study. In addition, commercial search engines can only search for inlinks while a research Web crawler can record both inlinks and outlinks (see Glossary). However, if a study only needs inlink data, then it is appropriate to use only commercial search engines. Studies that used both a research crawler and commercial search engines found that data collected from these two sources are correlated.

Validity and Reliability of Data Collected from Commercial Search Engines

Web hyperlink data collected from different search engines will differ because different search engines employ different crawling algorithms and thus index different Web sites. One of the ways to overcome the possible bias of a particular search engine and to improve the quality of data is to use multiple search engines. Data collected from different search engines can be pooled or averaged to achieve a more reliable count of links. Studies that used multiple search engines for data collection have found that the number of links to Web sites reported by different search engines are highly correlated, providing some assurance the quality of data collected from even a single search engine. However, the use of multiple search engines will provide an extra safeguard to the quality of data.

Results from commercial search engines were found to be volatile in that the same search query retrieved different results on different days or even during different times of the same day. This calls into question the suitability of commercial search engines for data collection. Due to the improvement of search engines in general, recent (from late 2001 on) data collection experiences show that search engine performance is fairly stable. However, the stability of a search engine used in a research project still needs to be monitored in data collection.

Web Hyperlinks as an Indicator of the Calibre of Hosting Organization

Numerous studies have been carried out to determine if Web hyperlinks do contain useful information. The main method of this line of investigation is to compare the number of links to Web sites with existing measures of the organizations operating the Web sites. Significant correlations between the two types of data have been found, providing evidence that the hyperlink is a new source of information and Web data mining using this source could be fruitful. This type of research has so far been conducted on the following three types of Web sites, academic Web sites, commercial Web sites, and scholarly journal Web sites, as summarized below. There are also studies that analyze links between universities and other sectors of society, such as industry and government. More of this kind of cross-sector research is needed.

Academic Web Sites

Studies on academic Web sites attempted to establish a positive correlation between the number of links to university Web sites and the scholarly activity of the universities such as research and teaching. Thelwall collected data on the external link counts of university Web sites in the United Kingdom using both a specialized Web crawler and the commercial search engine AltaVista. He then compared the link data with the official U.K. government measure of university research called the Research Assessment Exercise (RAE) and found a positive correlation between the two measures. Studies that investigated university Web sites of other countries such as Australia, Canada, and China also

found significant relationships between inlinks and various university quality measures. Similar correlations were also found when the unit of analysis is individual university departments rather than the whole university.

Commercial Web Sites

There are fewer studies of Web hyperlinks on commercial Web sites than those on academic sites. Reid developed a method for analyzing the Web's hyperlink structure for gathering business intelligence information. The first step of the method is to conduct a link search to identify the implicit business community surrounding a particular company or product. The Web sites of this community can then be analyzed to obtain business information.

Vaughan and Wu investigated the Web sites of China's top 100 information technology companies and found significant correlation between counts of links (both external link and total link) to a company's Web site and the company's business performance measures such as revenue and profit. A follow up study that applied the same research process to top U.S. information technology companies found the same correlation. Even more remarkable is that the two sets of correlation coefficients for the two countries are very close although the two groups of Web sites are very different in terms of characteristics such as age. Further analysis that examined Web site age as a possible confounding variable confirms that the correlation found is genuine rather than spurious.

Journal Web Sites

Since Web hyperlink analysis is rooted in citation analysis and there is a direct analogy between Web hyperlinks and citations in journal articles, as described above, it is natural to analyze links to journal Web sites. While early studies did not establish a relationship between links to journal Web sites and the quality of the journal, probably due to limitations of the methodology employed and the lack of maturity of the Web sites, more recent studies did find a significant correlation between the two, where the quality of journal is measured by the Journal Impact Factor described in the first section of this article. Furthermore, the Vaughan and Thelwall study found that the ages of Web sites correlated with the numbers of inlinks in that older Web sites received more links to them. The study also found that sites containing more detailed content attracted more links.

Reasons for Web Linking

All the correlations reported above, although statistically significant, were not very high and outliers that did not fit the general pattern existed. It is very important to study the reasons or motivations for linking to gain a better understanding of Web hyperlink phenomenon. Such studies are qualitative in nature, which complement the quantitative link count studies reported above and shed light into the correlations found. However, link creation studies have lagged behind although most correlation-based quantitative studies did some basic qualitative analysis of anomalies. The most comprehensive link creation motivation study so far was conducted by Wilkinson, Harries, Thelwall, and Price. A random sample of 414 links among university Web sites in the United Kingdom was taken and classified into various types based on the reasons for linking. The study found that over 90% of links were related to academic activity in some way, including research and teaching. There are other studies on the reasons for Web linking. Combining results from these studies with the link count correlation studies, it can be said that Web hyperlinks could be an indicator of the caliber of the organization operating the Web site. Links to academic sites indicate the quality of the universities while links to commercial sites reflect the business performance of the companies. However, the word "indicator" here should be interpreted in the context of correlation rather than causation. Studies so far have demonstrated correlation but no causation has been established.

Analyzing the Topology of Web Hyperlinks

While the studies reported above focused on the number of links to Web sites, the topology (i.e., structure) or the pattern of Web hyperlinks has also been examined resulting in important findings.

Social Network Analysis Using Web Hyperlink Data

Social network analysis, a research method developed primarily in sociology and communication science, focuses on patterns of relations among people and among groups such as organizations and states. As the Web connects people and organizations, it can host social networks. Therefore, social network analysis has been used to study Web hyperlinks. Garrido and Halavais studied the networks of support for the Zapatista movement, a contemporary social movement in which the Internet played a central role. The study collected data on links to the Zapatista Web site and mapped these links into a Zapatista network on the Web. This network of Web sites provided a unique insight into the character of the Zapatista's phenomenal success. Park, Barnett, and Nam examined Korea's 152 most popular Web sites and defined the affiliation among the Web sites based on

interdomain hypertext links. They then applied cluster analysis to reveal a hyperlinked network in which financial Web sites were found to be in the most central position.

The "small-world" theory, which stems from research in social network analysis over 30 years ago, deals with the short distances between two arbitrary persons through intermediate chains of acquaintances. The theory has been applied to the Web environment where the number of intermediate acquaintances between two persons is replaced by the number of links along directional link paths between two Web sites or Web pages. In a "small-world" network, such as in a Web network, it is sufficient to have a very small percent of links to function as "shortcuts" connecting "distant" parts of the network. Many studies have been conducted on the subject including a Ph.D. thesis. There are a number of applications of the "small-world" theory on the Web including analyzing Web communities and examining cultural and social currents and formation. For example, Björneborn used the "small-world" theory to identify central connectors, gatekeepers, and cross-disciplinary contacts in academic Web spaces. Furthermore, the "small-world" phenomena affects the speed and exhaustivity with which Web crawlers can discover and retrieve Web pages when following links from Web pages to Web pages.

"Success Breeds Success" Phenomenon of Web Hyperlinks

Numerous studies have discovered and confirmed the "success breeds success" (also called "preferential attachment") phenomenon of Web hyperlinks. When the number of inlinks is plotted on the Y axis and the number of Web sites plotted on the X axis, the result is a hyperbolic curve with very long tails (see the work of Rousseau for an example). This means that a small number of Web sites receive many links while the majority of Web sites attract very few links. The underlying cause of this phenomenon has been explained by the fact that the more inlinks a Web site has, the more visible the site is, and thus the more likely it is that the site will receive further links. This "success breeds success" or "cumulative advantage" phenomenon is similar to the rich-get-richer phenomenon in the society in general.

Identifying Web Communities Based on the Hyperlink Structure

The interconnected Web pages can be viewed as a "graph" in mathematical graph theory where the Web pages can be seen as "nodes" and the hyperlinks among pages seen as "edges" in the graph theory terminology. It is beyond the scope of this article to discuss the details of the graph theory. Instead, a brief description of the work on community identification based on hyperlink structure is provided to illustrate the potential of Web hyperlink analysis. A Web community can be defined, in nontechnical terms, as a group of people whose Web pages share a common interest. The identification of Web community is useful not only for improving Web searching by displaying similar pages together but more importantly, from a social science point of view, to study social communities. Web link data have been used successfully to identify Web communities and a faster algorithm has been developed recently to make this method more viable.

See Also the Following Articles

Computerized Record Linkage and Statistical Matching • Internet Measurement

Further Reading

Björneborn, L. (2003). *Small-World Link Structures across an Academic Web Space—A Library and Information Science Approach*. Ph.D. thesis. Royal School of Library and Information Science, Copenhagen, Denmark.

Cronin, B. (2001). Bibliometrics and beyond: Some thoughts on web-based citation analysis. *J. Inform. Sci.* **27,** 1–7.

Garfield, E. (1994). The impact factor. *Current Contents,* June 20 (www.isinet.com/isi/hot/essays/journalcitationreports/7.html).

Garrido, M., and Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying social network analysis to study contemporary social movements. *Cyberactivism: Online Activism in Theory and Practice* (M. McCaughey, and M. Ayers, eds.). Routledge, New York.

Ingwersen, P. (1998). The calculation of web impact factors. *J. Document.* **54,** 236–243.

Leydesdorff, L., and Curran, M. (2000). Mapping university–industry–government relations on the Internet: The construction of indicators for a knowledge-based economy. *Cybermetrics* **4** (www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html).

Milgram, S. (1967). The small-world problem. *Psychol. Today* **1,** 60–67.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web (citeseer.nj.nec.com/page98pagerank.html).

Park, H. W., Barnett, G. A., and Nam, I. (2002). Hyperlink-affiliation network structure of top web sites: Examining affiliates with hyperlink in Korea. *J. Am. Soc. Inform. Sci. Technol.* **53,** 592–601.

Reid, E. (2003). Using web link analysis to detect and analyze hidden web communities. In *Information and Communications Technology for Competitive Intelligence* (D. Vriens, ed.). Ideal Group, Hilliard, OH.

Rousseau, R. (1998/99). Daily time series of common single word searches in Alta Vista and Northern Light. *Cybermetrics* **2/3** (www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html).

Rousseau, R. (1997). Sitations, an exploratory study. *Cybermetrics* **1** (www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html).

Thelwall, M. (2001). Extracting macroscopic information from web links. *J. Am. Soc. Inform. Sci. Technol.* **52,** 1157–1168.

Vaughan, L., and Thelwall, M. (2003). Scholarly use of the web: What are the key inducers of links to journal web sites? *J. Am. Soc. Inform. Sci. Technol.* **54,** 29–38.

Vaughan, L., and Wu, G. (2003). Link counts to commercial web sites as a source of company information. In *Proceedings of the 9th International Conference on Scientometrics and Informetrics, Beijing, China, Aug. 25–29, 2003* (G. Jiang, R. Rousseau, and Y. Wu, eds.), pp. 321–329. Dalian University of Technology Press, Dalian, China.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* **393,** 440–442.

Wilkinson, D., Harries, G., Thelwall, M., and Price, E. (2003). Motivations for academic web site interlinking: Evidence for the web as a novel source of information on informal scholarly communication. *J. Inform. Sci.* **29,** 49–56.

Web-Based Survey

R. Michael Alvarez
California Institute of Technology, Pasadena, California, USA

Carla VanBeselaere
Mount Allison University, Sackville, New Brunswick, Canada

Glossary

framing (or priming) Activating different mental constructs to see how they influence responses. In surveys, this usually involves altering question wording or the information accompanying questions.
Internet Term used to capture both e-mail and World Wide Web (WWW) applications.
random digit dialing (RDD) A method of number selection that includes all possible telephone numbers including unlisted numbers and new numbers. RDD often uses information about which interchanges will likely contain residential numbers as a basis for selecting a sample of telephone numbers.
sample The units from the population that are drawn from the sampling frame to be included in the survey.
sampling frame A list of sampling units from which the sample can be selected.
sampling units Groupings of the target population that cover the whole population but do not overlap—every element of the population belongs to one and only one sampling unit.
self-completion survey A survey completed by the respondent without assistance from an interviewer.
target population The entire set of units (whether individuals, households, organizations, institutions, geographic entities, or others) for which the researcher wants to make generalizations or inferences.
World Wide Web (WWW or Web) A system of extensively interlinked hypertext documents.

The World Wide Web (WWW) or Internet has recently been recognized as a valuable instrument for conducting surveys. Low costs, rapid turn around, access to a vast geographically diverse pool of potential respondents, and the ability to present complex graphical material make the Web appealing as a new survey mode. While this new survey mode may offer many opportunities, its strengths and weaknesses are still being studied. As the Internet develops—especially as Internet access widens to include a more representative cross-section of the adult population—the applications for Web-based surveying are likely to flourish. The future of Web-based surveys will undoubtedly be hotly debated, but an understanding of the fundamentals of Internet surveying is a prerequisite for such a debate.

Introduction

What Are Web-Based Surveys?

Web-based surveys are not a new creation. Rather, the World Wide Web simply provides a new medium through which survey data can be collected. Surveys published on the Web can be accessed by potential respondents who have a computer with an Internet connection and Web-browsing software. A variety of computer programs and languages, such as the Hypertext Markup Language (HTML), are used to present surveys and collect data. To participate in these surveys, respondents are usually required to visit a particular Web address or universal resource locator (URL). Once respondents complete the survey and submit their responses, data is transmitted electronically to the researcher for analysis. Early Internet surveys were generally distributed by e-mail; but, while e-mail facilitates the distribution of surveys, e-mail responses cannot be automatically submitted to a database. This article focuses primarily on Web-based surveys,

since this new survey mode appears to offer more opportunities that e-mail surveys.

The flexibility of HTML and related computer languages allows the deployment of diverse types of surveys. Exploiting the graphical capabilities of Web browsers, such as Internet Explorer or Netscape, Web-based surveys can incorporate images and multimedia material. Web-based surveys can also be programmed to provide respondents a tailored experience. Like other types of computer-assisted surveys, Web-based surveys can be designed to automatically skip questions that are irrelevant based on responses to previous questions, or they can randomize the order of questions or response options.

A key feature of Web-based surveys is that they do not require interviewers. Like mail surveys, Internet surveys are self-administered. Respondents complete the survey at their leisure and transmit their responses electronically. This process makes Internet surveying relatively low cost, reduces data entry requirements, eliminates the possibility of transcription or data entry errors, and greatly accelerates survey administration. As a result, Internet surveys are proliferating at an amazing rate. In 1999, Kaye and Johnson identified over 2000 Web-based surveys using an informal Yahoo! (www.yahoo.com) search.

How and Where Web Surveys Are Used

Web-based surveys have been used for a variety of different purposes; these surveys have been used to collect opinion data, demographic information, and purchasing behavior, to name a few. Web-based surveys are used by academics, market researchers, corporations, the media, and many others. Topics range from political polls to surveys of illicit drug dealing behavior. Web-based surveys are used for marketing purposes for public opinion polling and to study the behavior and beliefs of Internet users.

Internet searches reveal a large number of companies that offer Web-based survey services and software. Existing research companies expanded their services by offering Internet-based surveys while other new companies entered the field offering unique Web-based technology for surveys. Academic institutions are also using the Internet to contact respondents. Some prominent Internet survey groups include Harris Interactive, Knowledge Networks, the Internet Survey of American Opinion, the GVU WWW User survey, and Greenfield On-line. Although coverage error and nonresponse bias are a concern for Web-based surveys, several surveys have performed well on the objective measure of election forecasting; in the 2000 presidential election, the Harris Interactive poll did better at predicting state level presidential votes than similar telephone surveys.

Since Web-based surveys generate a large number of responses and can be easily programmed to provide respondents a tailored experience, they are optimal venues for testing how the wording of questions affects responses. By providing different respondents with different question wording or accompanying text, researchers can examine how framing or priming affects responses. Rapid turn around, low costs, and high response volumes make Web-based surveys an attractive medium for this type of survey experimentation.

Web-based surveys also promise to democratize the process of survey data collection: "Not only can researchers get access to undreamed numbers of respondents at dramatically lower costs than traditional methods, but members of the general population too can put survey questions on dedicated sites offering free services and collect data from potentially thousands of people" (from Couper). Given the popularity and accessibility of Web-surveying, it is important that we understand the fundamental issues related to Internet surveying. Not all Web surveys are equal. The validity and value of Web surveys will depend on how the survey is designed and implemented. Recognizing the strengths and weaknesses of Internet surveys will ensure that Web surveys are designed appropriately and that results are considered carefully.

Article Overview

This article examines some of the fundamental issues about using the Internet as a survey tool. In addition to considering the practical issues involved in implementing Web-based surveys, this paper presents an overview of the different types of Web-based surveys. Methodological issues such as coverage and sampling are examined. In an attempt to facilitate the task of evaluating and improving Web-surveys, a typology of Web-surveys is also offered. The last section contains some final details about implementing Web surveys.

Methodological Issues

While the Internet or WWW offers a new and exciting mechanism for the collection of survey data, this new survey mode faces many important methodological issues. The validity of survey results depends on how well the survey is designed and the techniques used to obtain the sample. Cochran lists the following principle steps that should be considered in any sample survey:

- Determine the objectives of the survey.
- Define the population about which information is wanted (*target population*).
- Determine the relevant data to be collected.
- Specify the degree of precision wanted from the sample.

- Determine format of the survey to be implemented.
- Define the sampling frame from which a sample is to be drawn.
- Divide the population into sampling units and select a sample from among these units.
- Organize the survey administration
- Summarize and analyze the data.

Careful implementation of these steps helps to ensure that survey results are valid and generalizable. In the context of Internet surveying, there are three components of survey sampling that may introduce problems. These issues are coverage error, sampling issues, and nonresponses bias. Recognizing the concerns surrounding these issues, adjustments in survey objectives, and administration may be necessary.

Coverage Error

Coverage error is the deviation between the sampling frame and the target population. The degree to which coverage error is an issue depends on the population about which we wish to make general statements. For example, if we are interested in sampling from a population for whom we have e-mail addresses there is no coverage error because we can use the list of e-mail addresses as the sampling frame. If, however, we are interested in surveying a large group, such as all eligible voters in the United States, the coverage error is a significant concern because not all voters have Internet access, nor is there a list of e-mail addresses for this population.

In order to develop an appropriate sampling frame for Web-based surveys, we must be able to identify the units to include. In some cases, the population and sampling frame are known with certainty: examples of these types of situations usually involve smaller populations, like groups with known e-mail addresses, visitors to particular Web-sites, or the like. In these circumstances, it is possible to easily identify all members of the target population and ensure that they have a positive probability of being sampled.

Unfortunately, there are many circumstances where the population is not known with certainty, especially if we are interested in studying large populations like the universe of Americans adults. For many large populations of interest a simple list—for the sampling frame—does not exist. For example, there is no general list of e-mail addresses for the adult American population. Web-based surveys of large groups, like all eligible voters, is further complicated by the fact that not all potential respondents have computers or Internet access. According to NUA Ltd. (www.nua.ie/surveys/how_many_online/index.html), in February 2002, approximately 544.2 million people (8.96% of the population) had Internet access throughout the world. Internet penetration is the highest in North America; in the United States, approximately 164.4 million people (58.5% of the population) had Internet access. While Internet penetration rates continue to grow, we are a long way from being able to access all potential respondents if the target population is the entire American population.

Sampling

Despite the difficulties in developing a complete sampling frame of Internet users, researchers have pursued a variety of ways to obtain Internet survey responses. Since no established methods exist for recruiting survey respondents, a variety of approaches have been considered to obtain Web-based survey samples. Couper identifies two basic approaches to Web-survey recruitment: probability and nonprobability surveys.

Probability approaches involve the researcher identifying the population, developing a sampling frame, and using the sampling frame to generate a random research sample. Using this approach, the probability that any unit of the population will be sampled is known and thus the sampling error can be calculated. Probability-based Web-surveys can be used to make generalizations about the population upon which they are based.

Two basic approaches have arisen to conduct probability-based surveying on the Web. One is to restrict the population to only Internet users and to devise methods of randomly selecting Internet users into a sampling frame. The second is to use other approaches to contact a broader spectrum of potential respondents (i.e., telephone) and then recruit them into a pool or panel of potential survey respondents.

The other types of Internet surveys, based on nonprobability approaches, are probably the most ubiquitous surveys on the Internet. These surveys make no attempt to identify the sampling frame or randomly select respondents. These types of Web-survey are frequently used when it is difficult to identify members of the target population or contact a probabilistic sample from the population. Any inferences made about population parameters from nonprobability surveys are potentially problematic.

For surveys of the general American population, probability-based sampling over the Internet is complicated. While the same problem arose in telephone surveying, the development of technologies like random digit dialing (RDD) enabled researchers to approximate random samples of the American population. Assuming that most households have a telephone and given that RDD techniques ensure that each residential phone number has an equal probability of being drawn, RDD generates a random sample of potential respondents. Unfortunately, this method does not generalize directly to the Internet.

First, over 95% of households have telephones but less than 50% of households have Internet access (from the U.S. Department of Commerce). In addition, since e-mail addresses (the Web equivalent of telephone numbers) involve more than simple seven-digit numerical combinations, it is extremely difficult to randomly generate valid e-mail addresses. Furthermore, even if random generation of e-mail addresses were possible, sending large quantities of unsolicited e-mail (spam) is frowned upon. As a result, contacting a random sample from a large population for which no contact list exists is difficult and may not generate representative samples even if we condition on having Internet access.

Until Internet access becomes universal and all-inclusive e-mail directories are developed, obtaining probability samples using only Web-based tools will be difficult. To overcome this, many researchers have implemented multimode surveys. Respondents are often contacted by telephone and asked to participate in a Web-based survey. Knowledge Networks claims to have solved the problem of sample representativeness by providing prerecruited pools of survey respondents Internet access in exchange for completing regular Web-based surveys. Alternatively, researchers have used telephone surveys to supplement Web-based surveys. For example, Harris Interactive undertakes a telephone survey in order to develop appropriate weights for their Web-based survey responses.

Nonresponse Bias

The methodological concerns do not end once a sample of potential respondents has been contacted. Error or nonresponse bias may also be introduced because some members of the selected sample are unable or unwilling to complete the survey. The extent of bias depends on both the incidence of nonresponse and on how nonrespondents differ from respondents on variables of interest. The effect of nonresponse is to confound the behavioral parameters of interest with parameters that determine response. Nonresponse bias is not unique to Internet surveys but the potential problem is quite severe for Web-based surveys that have low response rates and nonrandom recruitment procedures.

Web-survey nonresponse might be aggravated because potential respondents encounter technological difficulties. Internet respondents need to have basic literacy skills, know how to surf the Web, be able to use the mouse to select response options from menus, and know how to type answers in the fields provided. Furthermore, technological hurdles, such as browser incompatibility and slow Internet connections, will influence whether a potential respondent completes a survey. Since Internet access tends to be correlated with demographic characteristics such as income and age, Internet survey data will provide biased results if these demographics affect the variables of interest.

Several methods exist to account for selection bias in survey samples but these corrections are complicated by the fact that Web-based surveys provide very little information about nonrespondents. Techniques such as propensity weighting or other simple weighting schemes may be useful in improving the representativeness of Internet survey samples. Simple weighting schemes may be useful in minimizing these biases and errors if there is a strong relationship between the weighting variable and the data in the survey. Supplementing Web surveys with telephone surveys can be used to develop appropriate weighting schemes.

While nonresponse bias is a significant concern for Internet surveys, recent research makes apparent the fact that traditional methodologies, like RDD telephone surveys, may also be problematic. Alvarez *et al*. report data from a telephone survey in which they began with 13,095 residential telephone numbers to obtain 1500 complete interviews. Of these, 3792 phone numbers were bad in some way, 5479 produced no answer or complete interview, and 1469 produced a valid contact but the survey interview was refused. As few telephone survey studies report statistics like these, it is impossible to characterize the extent to which contemporary telephone survey techniques produce representative samples. The Alvarez *et al*. evidence suggests that RDD techniques do not necessarily provide truly random samples. Obtaining random samples from large populations may be difficult over the Internet but telephone surveys are not a panacea.

Web-Survey Typology

Probability vs Nonprobability Surveys

As discussed above, Web surveys can be classified based on how they generate respondent samples. The two basic ways of recruiting respondents involve probability or nonprobability approaches to Web surveying.

Couper identifies at least four different types of probability-based Web surveys:

1. Intercept-based surveys of visitors to particular Web-sites.
2. Known e-mail lists.
3. Prerecruited panels.
4. Mixed-mode survey designs.

The first, the intercept-based approach, is based on interview techniques used in exit poll surveys or many types of market research. With a sampling frame being all visitors or users of a particular Web site, the sample is some randomly selected set of visitors who are asked to participate in some form of survey. Known e-mail lists are

a second form of probability-based Web survey. When the population is one that has universal Internet access and for which a directory of e-mail addresses is available, Web-based surveys can be extremely useful; student, university faculty, or employee surveys are examples of known e-mail list surveys. These two types of surveys can minimize sampling and coverage errors.

The remaining approaches for probability-based Web surveys are based on already having a probability sample and then using this sample to obtain Web-survey subjects. In the prerecruited panel approach, researchers use other techniques of probability sampling, like RDD telephone surveys, to recruit Web-survey samples; such an approach works well for studies of the Internet-using population if respondents with Internet access are willing to provide their e-mail addresses and participate in subsequent Web surveys. Knowledge Networks (www.knowledgenetworks.com) extended this concept by offering a random sample of respondents' Internet access in exchange for a commitment to participate in on-going Web-based surveys. Finally, mixed-mode approaches simply offer Web surveys as one of a multitude of modes for their participants to use (in addition to telephone or other modes).

Nonprobability Web surveys are probably the most ubiquitous surveys on the Internet. There are three types of nonprobability Web surveys identified by Couper:

1. Entertainment surveys.
2. Self-selected surveys.
3. Volunteer survey panels.

The first, entertainment surveys, are found all over the Internet. Generally they are not intended for scientific surveying, but for the entertainment of visitors to Web sites. Self-selected surveys are those on the Internet that give visitors to a Web site the opportunity to participate in a survey; thus, only visitors to the site are possible subjects and only if they actively initiate the interview. The third type of nonprobability Web survey is volunteer survey panels, where respondents are recruited on the Internet through advertisements of various types. Harris Interactive (vr.harrispollonline.com/register/main.asp) and Greenfield Online (greenfieldonline.com) are perhaps the best known volunteer panels, but the technique is used by many other survey researchers. Volunteer panels require that prospective respondents go to a particular Web site and provide some information about themselves (including their e-mail address). This data is then maintained in a database from which respondents can be sampled for participation in subsequent Web surveys. Although volunteer panels are not based on probability sampling they are more likely than the other nonprobability surveys to attract a representative sample.

Nonprobability Internet surveys are not based on rigorous sampling procedures, raising concerns about the validity of inferences drawn from them. However, nonprobability Internet survey samples can and are being used in situations where researchers desire to exploit within-sample variance in a situation where statistical power can be maximized. For example, Internet survey samples can be used to examine priming or framing, especially studies that might involve graphical or multimedia materials. In such designs, thousands or tens of thousands of subjects might be included in a potential study and, as long as these subjects are assigned to control and experimental groups using some type of probability assignment protocol, this could produce powerful experimental results.

Web-Survey Formats

Examining the different Web-survey formats is also enlightening. While Web surveys involve many different topics, there are only two main formats for presenting a Web survey: interactively or passively. These two formats are aesthetically different and have distinct advantages and disadvantages.

Figure 1 contains an example of a typical interactive survey. As illustrated in this figure, interactive surveys are presented screen-by-screen. By clicking on a button, like the "to continue" button in Fig. 1, respondents can go to a new question on a new screen. This allows the data from the question to immediately be electronically transmitted to the surveyor ensuring that data from partially completed surveys is maintained. However, this format may make it difficult for respondents to review and correct their answers. Interactive surveys can also automatically skip questions which are determined to be irrelevant to the respondents based on how they answer previous questions. For example, if respondents indicate that they do not have children, all subsequent questions related to children can be automatically skipped. A drawback of this design is that respondents do not see the survey in its entirety and therefore cannot easily determine its length. To compensate, a progress indicator can be used to inform the respondent how much of the survey remains to be completed. Another difficulty with interactive surveys is that they may require special software, such as Java, potentially making it difficult for respondents with older, less powerful computers and Web browsers to respond.

Passive survey designs involve presenting the entire survey at once. Figure 2 displays the first part of a passive survey. The bar on the right-hand side of this figure indicates that respondents can scroll down on the page to view the rest of the survey. The data from passive surveys is transmitted once the respondent has completed all questions and clicked a submit button. An advantage of passive surveys is that respondents can easily browse through questions and review their responses before

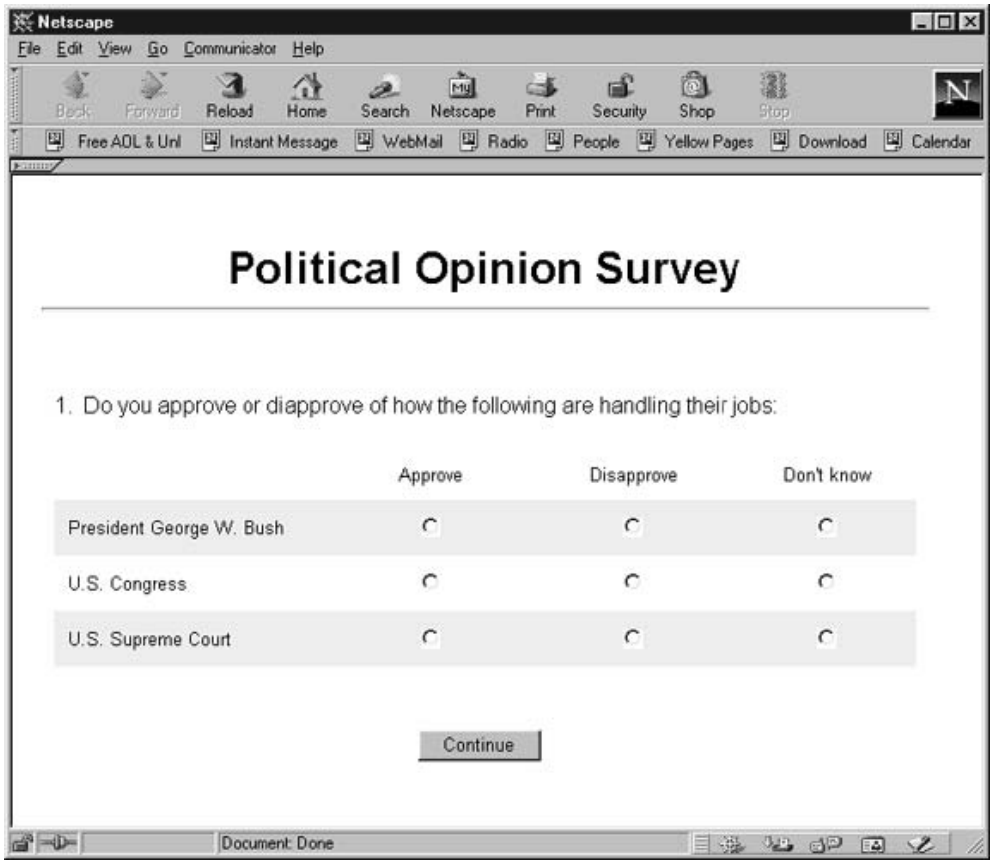


Figure 1 Example of an interactive style Web survey.

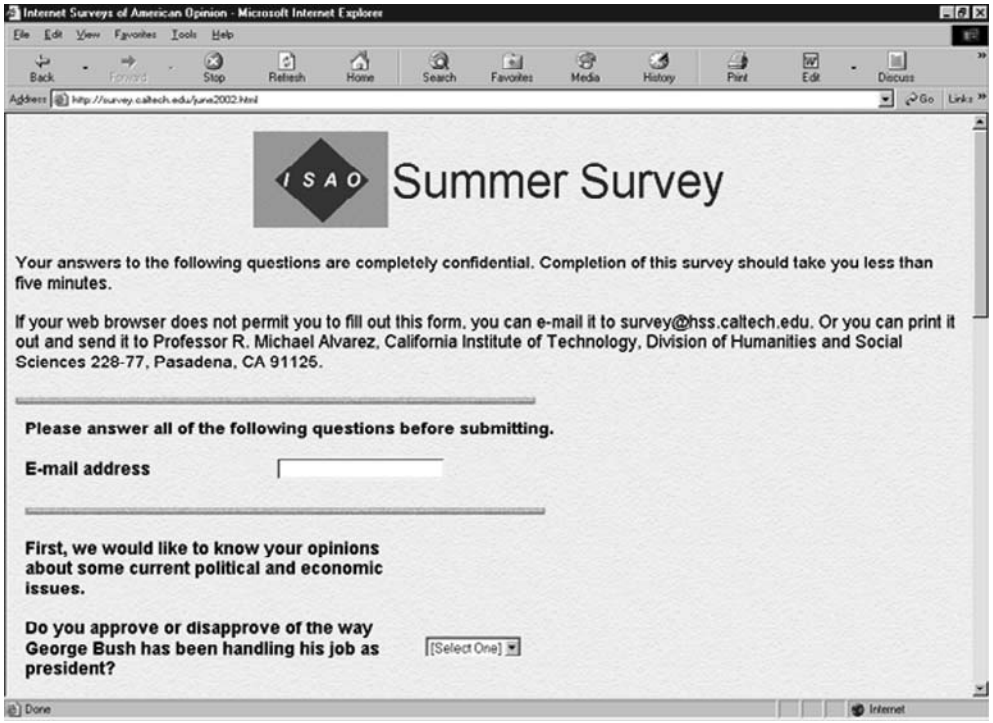


Figure 2 Example of a passive style Web survey.

submitting. These types of Web surveys are also easy to produce and easy to access so technical difficulties are less likely.

In addition to these two basic formats, the appearance of a Web-based survey also depends on how question response options are presented. There are four distinctly different ways to present response options: drop-down boxes, radio dials, check boxes, and open-ended boxes. Figure 1 contains an example of radio dials while Fig. 2 illustrates both open-ended boxes and drop-down boxes. Drop-down boxes appear in the questionnaire as a box with a downward pointing arrow—clicking on this arrow displays the list of response options and allows respondents to select from the list provided. Drop-down boxes are convenient for long lists of items since the response options are hidden. Radio dials, on the other hand, display all the responses options and require the respondent to click in the circle corresponding to their choice. Both drop-down boxes and radio dials are usually used when only one response must be selected from among the options provided. When respondents are allowed to select more than one option from a list, check boxes are the appropriate question format—respondents click on all the boxes that correspond to their answers. Finally, open-ended boxes allow respondents to type their responses in the space provided. As with other survey formats, open-ended question boxes can be useful when the associated question does not have responses that can be conveniently listed.

Developing Web Surveys

Respondents

Web-based surveys are only useful if they actually generate data, thus recruiting respondents is a priority. As discussed in the section on Web-survey typology, there are many different ways to recruit subjects. The choice of recruitment method will of course depend on the objectives of the survey. If the target population can be identified and easily contacted then producing probability-based Web surveys should be feasible. If, however, the intended target population is not well defined or readily contactable over the Internet, nonprobability respondent recruitment methods may be necessary. Inferences made about population parameters from nonprobability surveys are potentially problematic although several techniques have been proposed to improve the representativeness of Internet surveys.

Alvarez *et al.* discuss two prominent methods for recruiting respondents over the Internet. The first involves Web advertisements. Advertisements on various Web sites or newsgroups encouraging people to complete the Web survey is a fairly effective way of obtaining a large nonprobability sample of respondents. Another method to recruit respondents is through subscription or coregistration procedures. This involves asking individuals registering for another service whether they would like to provide their e-mail address and participate in Internet surveys. Once respondents provide their e-mail addresses, they can be contacted by e-mail to participate in Web-based surveys.

Web-Based Survey Panels

Because recruiting respondents over the Internet can be somewhat complicated, survey panels are popular. Rather than asking respondents to complete a single survey, Web-based survey panels recruit subjects to participate in a series of surveys. In order to obtain probability-based subjects, potential respondents are often initially contacted by telephone. Once respondents agree to participate in a survey panel, they are contacted by e-mail when they are required to complete a new survey. Using panels, researchers can draw samples from the registered respondents in order to undertake studies of specific subpopulations. Knowledge Networks claims that their Web-based survey panels is particularly effective for market research. Panels also offer the opportunity to examine temporal changes in respondent behavior and beliefs.

Researchers are currently studying the long-term effectiveness of Web-survey panels. Although the concept is relatively new, studies to date do not indicate that extended participation in Internet panels affects respondent behavior. However, preliminary data indicates that response rates do tend to decline with panel tenure. A significant problem for many Internet panels is that participants are frequently unreachable by e-mail because they have changed e-mail addresses or there are technical problems. These issues need to be continually examined especially for long-standing panels.

Creating the Survey

An advantage of Web-based surveys is that they are relatively easy to conduct. All that is needed is a Web site and some basic Web programming skills. Many surveys are created simply using Hypertext Markup Language (HTML); there are dozens of HTML editors available and they are becoming increasingly sophisticated and easy to use. Data from surveys can be captured either by programming the form to e-mail the data to a specified address or through a common gateway interface (CGI) script. Several HTML development packages automate the process of developing CGI scripts necessary to capture data from HTML forms. Internet survey companies have even developed computer programs that automatically create surveys.

Despite the fact that Web-based surveys are easy to implement, their effective use requires an understanding

of the methodological issues presented above. While Web surveys have many potential uses, making general statements about large populations based on Internet survey results is currently problematic. The Web opens up a whole new realm of survey possibilities, but it is important to evaluate surveys based on the fundamental criteria outlined in this article.

See Also the Following Articles

Internet Measurement • Survey Design • Surveys • Telephone Surveys • Total Survey Error

Further Reading

Alvarez, R. M., Sherman, R. P., and VanBeselaere, C. E. (2003). Subject acquisition for web-based surveys. *Political Anal.* **11,** 23–43.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., and Weimer, D. L. (2003). The advent of Internet surveys for political research: A comparison of telephone and internet samples. *Political Anal.* **11,** 1–22.

Cochran, W. (1977). *Sampling Techniques.* 3rd Ed. Wiley, New York.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opin. Quart.* **64,** 464–494.

Couper, M. P., Traugott, M. W., and Lamias, M. J. (2001). Web survey design and administration. *Public Opin. Quart.* **65,** 230–253.

Dillman, D. A. (2000). *Mail and Internet Surveys: The Tailored Design Method.* 2nd Ed. Wiley, New York.

Kaye, B. K., and Johnson, T. J. (1999). Research methodology: Taming the cyber frontier. *Soc. Sci. Comput. Rev.* **17,** 323–337.

Knowledge Networks (2002). *Decoding the Consumer Genome.* (www.knowledgenetworks.com/press/presskit.html).

Soloman, D. J. (2001). Conducting web-based surveys. In *Practical Assessment, Research and Evaluation,* Vol. 7.

U.S. Department of Commerce (2000). Falling through the Net: Toward digital inclusion (www.ntia.doc/ntiahome/fttn00/falling.htm).

Weber, Max

Anthony Oberschall
University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Glossary

empirical social research Obtaining and analyzing data on social topics using a variety of observational techniques and quasi-experimental study designs.
methodological individualism A theoretical principle that holds that groups, institutions, collective beliefs, and supra-individual entities can be explained only with reference to the actions, beliefs, attitudes, and motivations of individual human beings.
social survey Quantitative inquiry on a social topic using a standardized instrument or method applied to a sample of units, respondents, or informants.
verstehen (understanding) Weber's method of explanation in the human sciences.

In several episodes over his intellectual career, Max Weber engaged in empirical social research, experimented with a variety of techniques of inquiry, and reflected on the importance of individual and group research by scholars to supplement data from census bureaus and from historical studies. His most concentrated effort came in 1908 when he investigated industrial labor in a textile factory and explained the output of workers in terms of psychophysics (e.g., fatigue), cultural influences on work habits, social dynamics (e.g., peer solidarity), and instrumental goals (e.g., wages and earnings). His empirical work and methodological writings did not achieve the influence that his comparative historical and cultural studies and his theory of action did.

Overview of Weber's Work

Max Weber was born in 1864 and died in 1920. He was an influential comparative historian, legal scholar, economic historian, sociologist, historian of religion, and social scientist. Most of his professional life he was a private scholar in Heidelberg, although for short spells he lectured at various German universities. He was active in a number of German scholarly and policy associations, and he co-edited the *Archiv fur Sozialwissenschaft*, where he also published many of his essays.

Weber took on the big questions and themes in the social and historical sciences of the late 19th century. The biggest question was the origins of capitalism in early modern Europe and, more broadly, the distinctive causes and trajectory of Western rationalism that enabled the West to exit from tradition into modernity and to dominate non-Western civilizations. Though he remained indebted to Marx in many ways, he challenged the dominant Marxian historical materialism and argued for the unintended but powerful consequences of religion for other, including economic, institutions. Using the comparative historical method and relying on the specialized historical scholarship of his time, he also wrote specific essays and monographs on the evolution of the city as a distinct institution, the decline of the social and economic system of the Roman–Mediterranean world, the Puritan religion's impact on the saints' worldly activities (which he termed "inner worldly asceticism"), the economic ethic of world religions, political sociology, social stratification, nationalism, the Russian revolution of 1905, the stock market, socialism, the sociology of music, and other topics.

Second in importance was Weber's effort to create the foundations for a single social science, and to counteract the differentiation of the social and historical sciences into specialized and splintered disciplines. In *Wirtschaft und Gesellschaft* he anchored that venture on a theory of action whose primitive elements describing individual human actions are combined and elaborated into

aggregate level and relational concepts such as authority, legitimacy, state, and bureaucracy, using the principles of methodological individualism.

Third, he wrote several methodological essays on concept formation; explanation in the human sciences ("verstehen"); theoretical concepts and models ("ideal types"); the relation of unique events and individual actions to general explanations; the fundamental chasm between fact and value and between science and politics; the role of physical, biological, and hereditary factors in explanations of human behavior and institutions; the impact of great men ("charismatic leaders") on social change; the problem of intersubjectivity in understanding and interpreting human actions; causation in human affairs, and other topics in the philosophy of social science and the logic of social inquiry.

The fourth and the least known, indeed sometimes ignored, part of Weber's intellectual output was his empirical studies dealing with social policy issues such as the condition of farm workers in East Prussia, and the condition of life and productivity of industrial workers, together with methodological guidelines on how to design and conduct such empirical research. German social scientists in the Verein fur Sozialpolitik came to the conclusion that the facts and figures churned out by statistical agencies, which had stimulated the field of "moral statistics," based as they were on official documents such as birth and death registration, crime and suicide records, trade and production figures, public health records, and the like, were not always suited to answering questions scholars and policy analysts were interested in. They organized social surveys for obtaining additional data. Weber was an active participant in a major 1891–1892 Verein social survey on agricultural labor in East Prussia and a follow up survey in 1893 by the Evangelical Social Congress. He was the principal intellectual inspiration behind the Verein's 1909–1911 survey of industrial workers which failed in the end because of a trade union boycott. In preparation of the survey, in the summer of 1908, Weber undertook his most intense empirical study at the textile factory of a relative where he had complete access to personnel, production, and earnings records as well as to the workers themselves in the plant as a participant observer. At the founding of the German Sociological Society in 1910, which Weber conceived as an association for coordinating collective research projects, he outlined plans for a sociology of the press using content analysis and an empirical investigation of voluntary associations "from the bowling club ... to the political party and to the religious, artistic and literary sect." However, these plans were not implemented.

Weber's commitment to empirical social research ran deep. He labored to teach himself Russian just so he could read the Russian press on the revolution of 1905. He spent hours doing statistical calculations on factory workers' production records, and told university students in his well-known "Science as a Vocation" lecture that "No sociologist ... should think himself too good, even in his old age, to make tens of thousands of quite trivial computations ... perhaps months at a time. One cannot with impunity try to transfer this task entirely to mechanical assistants if one wishes to figure out something ..." He declined to join the Heidelberg Academy of Sciences in 1909 because he believed its resources would be better spent on a social science research institute for undertaking social surveys and postdoctoral field work.

Weber's Logic of Social Inquiry

In his methodological essays, "verstehen" (usually translated as "understanding" but sometimes better rendered as "interpretation") plays a central role. It is Weber's term for what the Enlightenment philosophers and Scottish moralists called sympathy or fellow feeling. Weber put it most succinctly when he wrote that "one does not have to be Cesar in order to understand Cesar, otherwise the writing of history would make no sense at all." "Verstehen" solves the problem of intersubjectivity between the historian–social scientists and their human subjects, and also allows them to make sense of the transactions among the human subjects themselves.

"Verstehen" is also the justification for methodological individualism in social science, because the cognitive, emotional, attitudinal, and motivational processes of human action are attributes of individual human persons and not of collective entities such as armies, states, social classes and the like, except in a metaphorical sense: "Interpretative [verstehende] sociology considers the individual and his action as the basic unit, its 'atom' ... the individual is the sole carrier of meaningful conduct ... for sociology, such concepts as 'state', 'association', 'feudalism', and the like, designate certain categories of human interaction. Hence it is the task of sociology to reduce these concepts to'understandable' action, that is, without exception, to the actions of participating individual men" (Weber, 1956 [1921]). For purposes of simplification, Weber recognized that collective entities acting in a coordinated and homogeneous manner could be attributed an interest, mentality, motive, or disposition. As well, if collective entities have meaning for the minds of individual persons, as something actually existing, these collective representations can have causal influence on their actions.

In keeping with his methodology, Weber explains the stability of custom, a group property, in terms of individual actions and interactions: "the stability of custom rests on the fact that those who non-conform are maladapted, i.e. they have to put up with inconveniences and disadvantages, so long as the actions of the majority in their

social milieu expect the persistence of the custom and act in conformity with it." This mode of analysis is a hallmark of contemporary game theory and Weber's example is an instance of the Prisoner's Dilemma paradigm. By unpacking group attributes with methodological individualism, Weber links agency to structure, and avoids causal explanations based on supra individual phenomena, e.g., traditional mentality and collective consciousness, that themselves need explanation.

"Verstehen" is also central in Weber's notion of causality in human affairs. Multiple outcomes from similar initial conditions are common. The intervening processes that account for the variance are the different cultural meanings that humans attach to the same events and actions. Therefore, the appearance of a comet can be viewed as a natural event obeying the Newtonian laws of motion, and it can also be viewed as a manifestation of divine wrath and a warning to humankind to change its sinful ways. Intervening cultural variables thus play a key role in human causation. As an illustration, Weber describes how a belief in predestination can give rise to fatalism in everyday behavior or, on the contrary, to active ethical action, such as the inner worldly asceticism of the Puritans. A causal chain by way of a cultural intervening variable is for Weber a typical means by which nonrational beliefs and motives (e.g., predestination) give rise to institutionalized instrumental actions (work as a calling).

Weber applied his causality model in his empirical sociology of industrial work as well as in the sociology of religion and historical writings. Some causal relations, such as between fatigue and work output among textile workers, can be explained as a physiological reaction, without intervening mental or cultural variables. He observed some workers' output over the course of the day, as they get more tired, and over the course of the week, and inferred two fatigue cycles. Yet, superimposed on the physiologically determined production cycles, for some workers he found other intervening cultural processes that modified physical output: some socialist workers purposely reduced output ("bremsen') toward week's end to avoid becoming norm busters out of solidarity with their peers, and some Pietist women who had been socialized for work as a value and not just a means for making a living exceeded output projected from the fatigue cycle of typical female workers.

As these examples from Weber's historical writings and empirical research show, the principle of "verstehen" which justifies intersubjectivity, methodological individualism, the unity of the social and historical sciences, and causal analysis with intervening cultural variables, enabled Weber to navigate confidently across history and cultures, past and present, qualitative and quantitative data, from contemporary textile workers to the 17th century Puritans, from the prophets of ancient Israel to the mandarins of the Ming dynasty, from the condottieri of medieval Italian cities to the members of revolutionary assemblies in 1905 Russia. His scholarly reputation grew on his unique talent for such explanations.

Study Design and Measurement: The Psychophysics of Industrial Work

Weber did not write explicitly about the techniques of social research, e.g., sampling, study design, measurement, and data analysis, as he did about the logic of social inquiry. We can however analyze how he handled these techniques by examining his most comprehensive empirical research on industrial labor in a relative's textile factory, and his research guide to the collaborators in the larger Verein fur Sozialpolitik survey of industrial workers. In preparation, Weber carefully studied the physiological and psychological experiments on work, especially those of his Heidelberg colleague Kraepelin and of the industrialist–innovator Abbe and reflected on how to transfer and modify laboratory techniques to a survey instrument and field work in the factory. The dependent variables, productivity and earnings of particular workers, could be measured from registers and pay records. The explanatory variables were of three types: physiological (fatigue, gender-linked skills) and physical (different machinery) variables could be measured quantitatively as in the laboratory; motivational and cognitive variables based on instrumental rational reasons (working for piece rate or fixed rate) could be "interpreted pragmatically" by talking to workers; and mentality and attitudes (work habits from upbringing and religious milieu) could be "reconstructed introspectively," i.e., using "verstehen." Among his findings were the fatigue cycles, the effects of Sunday alcohol consumption upon reduced productivity on Monday, the socialist workers' goal of limiting output for maintaining peer solidarity and equality, and the Pietist women's excellent attendance and superior work habits.

Although Weber believed cultural variables (e.g., workers' beliefs, attitudes, political orientation) were important, he was uncertain whether quantitative methods based on questionnaires could be used to measure them. He relied on participant observation and conversations with the workers on the factory floor. When Adolf Levenstein, a self taught worker, succeeded in carrying out a survey of steel, textile, and mine workers, with over 5000 respondents, about their hopes and aspirations, political views and "weltanschauung," he was at a loss of how to analyze and present his findings, and turned to Weber for advice. Weber recommended coding each question and running frequency counts, then cross tabulating with background variables, then creating a typology of worker

mentalities based on the quantitative analysis of responses, much as such data are currently analyzed.

If one classifies the measurement techniques used by Weber and his German contemporaries along two dimensions, unobtrusive to obtrusive techniques, and qualitative to quantitative, one will find that the bulk of it was based on published statistics of state agencies and census bureaus, e.g., the field of moral statistics, which was unobtrusive and quantitative. Another way to describe it is that it was armchair research. The Verein fur Sozialpolitik surveys did use intermediaries and informants (e.g., ministers, school teachers) and eventually also the subjects themselves, for studies of farm laborers and industrial workers. That was an attempt to overcome the limitations of armchair research by somewhat more obtrusive and more qualitative techniques. Weber himself was quite open to a variety of techniques: the factory records were unobtrusive and quantitative; but he also gathered qualitative information on disposition and mentality with participant observation. His plan for the sociology of the press combined quantitative and qualitative data analysis. Despite some promising beginnings, by and large the academic social scientists remained comfortable with armchair research.

The lack of influence of Weber's empirical and methodological writings on his contemporaries was due to the devastating impact of the First World War on Germany, the lack of students he mentored, dominance of the historical method in German academia (lone scholar working on archival, literary, and archeological data), and the absence of any social science research institute for collaborative, continuous, cumulative, and large-scale empirical undertaking, which Weber and some others wished to found but were not successful funding. It may be somewhat of a paradox that it is also the success and virtuosity of the other parts of the Weber oevre that contributed to the neglect of his episodic but only partially successful, yet determined efforts, in empirical social science.

See Also the Following Articles

Commensuration • Typology Construction, Methods and Issues

Further Reading

Gerth, H., and Wright Mills, C. (1958). *Max Weber: Essays in Sociology.* Galaxy, New York.

Oberschall, A. (1965). *Empirical Social Research in Germany 1848–1914.* Mouton, Paris.

Oberschall, A., and Lazarsfeld, P. F. (1965). Max Weber and social research. *Am. Sociol. Rev.* **30**(2).

Weber, M. (1908). Methodologische einleitung fur die erhebungen des vereins fur sozialpolitik. In *Gesammelte Aufsatze zur Soziologie und Sozialpolitik.* Mohr, Tubingen.

Weber, M. (1908). Zur psychophysik der industriellen arbeit. In *Gesammelte Aufsatze zur Soziologie und Sozialpolitik.* Mohr, Tubingen.

Weber, M. (1909). Zur methodik sozialpsychologischer enqueten und ihrer bearbeitung. *Arch. Sozialwissen. Sozialpolit* **29,** 949–958.

Weber, M. (1913). Kategorien der verstehenden soziologie. In *Soziologie, Analysen, Politik.* Kroner, Stuttgart.

Weber, M. (1956 [1921]). *Wirtschaft und Gesellschaft.* part I. Mohr, Tubingen.

Weighting

Peter J. Lynn
University of Essex, Colchester, United Kingdom

Glossary

adjusted sampling weight Synonymous with combined weight.
adjustment cell Synonymous with weighting class.
basic sampling weight Synonymous with design weight.
combined weight A weight that combines the design weight with weighting adjustments to reduce error due to under coverage, sampling and nonresponse. Techniques such as poststratification, response propensity modeling, and calibration may be used to derive the adjustments.
design weight A weight that is inversely proportional to the selection probability of a sample unit, as determined by the sample design (ignores the impact of nonresponse or coverage errors).
nonresponse weighting A class of methods that adjust the distribution of the responding sample, through weighting, to match that of the selected sample in terms of variables believed to be associated with propensity to respond.
population-based weighting A class of methods that involves comparing the responding sample to external population data.
post-strata Comprehensive and mutually exclusive subgroups for which population counts are known and to which survey respondents can be allocated. Post-strata are used in post-stratification and do not typically coincide with sampling strata (pre-strata).
post-stratification A technique that adjusts the sample distribution, through weighting, to match the population distribution over post-strata.
raking A method for deriving weights by iteratively weighting a sample to two or more sets of comprehensive and mutually exclusive classes for which population numbers or proportions are known. A form of population weighting.
raking ratio estimation Synonymous with raking.
rim weighting Synonymous with raking.
sample-based weighting A class of methods that involves defining weights based solely on information available for the selected sample units.
sampling weight Synonymous with design weight.
trimming The restriction of extreme weights, typically to some arbitrary maximum value, in an attempt to minimize the variance-inflation effect of weighting.
weight A numeric value associated with a responding sample unit, representing the relative importance of that unit in analysis that aims to makes inferences from sample to population.
weighting class A set of sample units meeting some criterion, each of which is assigned the same weight.

Weighting is a process by which the units in a survey sample are assigned different numeric values (weights), representing the contribution that they will make to estimates based upon the survey data. The weights are designed to make the sample representative of the population from which the sample has been drawn. The weight for any particular responding unit may be interpreted as the relative number of population units that it represents. The calculation and application of weights is part of the process of statistical inference, by which conclusions can be drawn about a population of interest based upon knowledge of a sample drawn from that population.

Basic Principles

The idea of weighting is conceptually quite simple. In order to make statistical inference from sample to population, the sample units must in some defined sense represent the population. As there are more units in the population than in the sample, each sample unit must represent at least one, and on average usually considerably more than one, population unit. The number of population units represented by a sample unit is the weight

of that sample unit. In the simplest case, if a simple random sample of n units is selected from a population of N units (and there is no nonresponse), each sample unit can be thought to represent N/n population units. Thus, the weight for sample unit i ($i, = 1, \ldots, n$) is: $w_i = N/n$. The weight does not vary with i, all sample units receive the same weight. However, there are many situations in which it will be recognized that each sample unit does not necessarily represent the same number of population units. These situations are outlined below. In consequence, weights usually *do* vary across sample units.

Typically, the weight for each sample unit will be added to the survey data set as a variable. Then, the weight variable will be specified in analysis as an indicator of the contribution that will be made by that unit to an estimate. For example, a simple weighted mean (of a variable y, which takes the value y_i for unit i) will be calculated as:

$$\overline{y}_w = \frac{\sum_{i=1}^{n} w_i \times y_i}{\sum_{i=1}^{n} w_i}.$$

Reasons for Weighting

Disproportionate Stratification

A situation in which sample units obviously do not all represent the same number of population units is when the sample units have not been selected with equal probabilities. If disproportionate stratification has been used (such as oversampling units that belong to a small but important subgroup), then those sample units that had higher selection probabilities represent fewer population units. If there were H strata ($h = 1, \ldots, H$) and a simple random sample of n_h units were selected from stratum h, then each sample unit i in stratum h should receive the weight: $w_h = N_h/n_h$. Now, the weights vary across strata (in inverse proportion to the sampling fractions) but not between units within a stratum. Weights that are in inverse proportion to selection probabilities are often referred to as "design weights" or "sample weights."

Sampling Variance

With any kind of random sampling, units are selected on the basis of the outcome of some chance mechanism. In consequence, a sample may turn out to contain a higher proportion of one type of unit than another, just due to the play of chance (sampling variance). This may be known by reference to the characteristics of all units on the sampling frame, or by reference to some external source of data about the population. If the definition of "type" here is thought to be of relevance to the analysis, weights may be applied to correct the chance imbalance. For example, if the study population is known to contain 6000 women and 6000 men, but a simple random sample selected from it happens to contain 285 women and 315 men, it might be appropriate to give sample women a weight of 6000/285 and men a weight of 6000/315. These weights are identical in form to design weights, being $w_g = N_g/n_g$, where g ($g = 1, 2$) are the strata defined by sex. However, here the strata are not sampling strata (pre-strata). They are referred to as *post-strata* and weighting of this kind is known as *post-stratification* weighting. Weights are not inverse selection probabilities, but ratios of population size to sample size within post-strata. If disproportionate sampling has been used, post-stratification weights must be defined as the ratio of population size to design-weighted sample size, i.e.,

$$w_g = \frac{N_g}{\sum_{h=1}^{H} \left((N_h \times n_{gh}) / n_h \right)},$$

where n_{gh} is the sample size in the intersection of post-stratum g and pre-stratum h. In this case, the combined weight for each unit i in post-stratum g and pre-stratum h is $w_{gh} = w_g \times w_h$.

Nonresponse Error

Most surveys experience some greater or lesser degree of nonresponse. Nonresponse may cause the distribution of the responding sample (i.e., the sample for which data are available for analysis) to differ from that of the selected sample. This will happen if nonrespondents differ from respondents in terms of relevant characteristics. For example, suppose that a selected sample contains 400 employed people and 200 who are not employed (students, retired, etc.). If the response rate is 75% among the employed people but 90% among the others, the responding sample will contain 300 employed people and 180 others. The proportion of employed people is therefore lower in the responding sample (62.5%) than in the selected sample (66.7%). To correct for this differential nonresponse, employed respondents could be given a weight of 400/300 and other respondents a weight of 200/180. These weights are of the form $w_f = n_f/m_f$, where f ($f = 1, 2$) are the classes defined by employment status, and n_f is the number of units selected and m_f the number of units responding in class f. The classes are known as *nonresponse classes* and the weights are referred to as *nonresponse weights*. Nonresponse weights are ratios of selected sample size to responding sample size within classes, i.e., the reciprocal of the within-class response rate.

A combined weight for a unit in nonresponse class f, post-stratum g, and pre-stratum h is simply $w_{fgh} = w_f \times w_g \times w_h$.

Coverage Error

Sampling error and nonresponse error are two of the three main sources of errors of nonobservation on surveys. The third is coverage error. There are two types of such errors, under-coverage and over-coverage. Over-coverage exists when a sampling frame includes units that are not members of the study population. For any contacted sample unit, membership of the population can usually be established, so such units do not enter the responding sample (they are categorized as "ineligible"). Under-coverage exists when a sampling frame has imperfect coverage of the study population. If the omitted units have relevant characteristics that differ from those of the included units, then under-coverage introduces error. Weighting may be used to address the impact of under-coverage error, if population data is available superior to that available from the sampling frame. For example, if it is known that 1% of units are omitted from the frame in one geographic region, but 6% in another, then sample units in the two regions could be given weights of 100/99 and 100/94, respectively, to correct this imbalance. This might be thought to be an appropriate procedure if region is likely to be correlated with important survey measures.

Methods of Calculating Weights

Assumptions

Design weights, to deal with disproportionate stratification, are in some respects different in nature to the other three kinds of weights described in the previous section. Design weights are a direct consequence of the choice of sample design. In order to calculate them, it is necessary only to know the relative selection probability of each sample unit. With a well-defined probability sample design, it should not be problematic to establish the selection probability of each unit (though care may be needed if the design is multi-stage). Once the selection probabilities are known (an essential requirement for a probability sample design), the method of calculation of the weights is not controversial. No assumptions are required.

The other three reasons for weighting all aim to adjust for the potentially undesirable impacts of sources of survey error (coverage, sampling, nonresponse). By definition, it is not possible to know the impact of these error sources on survey estimates (because survey data is not obtained for units that are excluded from the frame, not sampled, or nonresponding). Instead, the impacts must be estimated using some kind of model. The model may be simple (as in the examples above) or complex, but any model involves assumptions. It is helpful to recognize the nature of assumptions that are typically made when calculating survey weights, as consideration of these assumptions should influence the choice of calculation method.

Most methods of calculating weights proceed in one of two ways. The first involves splitting the sample into weighting classes (also known as adjustment cells), calculating a weight for each class, and applying that weight to each sample unit in the class. The second involves estimating a propensity (to be included on the frame, to respond to the survey) using, for example, a logit or probit regression model and calculating a weight as the inverse of the model-estimated probability (of being included on the frame, or of responding to the survey). In either case, the weighting relies upon an assumption that the responding sample units within a class (or with a given set of characteristics that determine the estimated propensity) are similar in terms of survey measures to unobserved units in the class whose absence the weighting is intended to address (e.g., the nonresponding units in the case of nonresponse weighting). To the extent that the assumption is true, the weighting will tend to reduce the error of survey estimates. But, to the extent that differences remain between the observed units and corresponding unobserved units, error will remain in the survey estimates. Weighting therefore reduces, but does not remove entirely, the errors in question. The effectiveness of weighting depends on the extent to which the underlying assumptions are realistic.

A second assumption in the case of weighting class approaches for nonresponse is that units within a class all have the same underlying response probability and that these probabilities are independent. This is sometimes referred to as the response homogeneity group (RHG) model. That is, conditional upon membership of the weighting class, the data are missing completely at random (MCAR).

It should be noted that the nature of the fundamental assumption is different in the case of post-stratification, compared with weighting for under-coverage or nonresponse. Any differences between sampled units and not-sampled units within post-strata are solely due to random sampling variation. In other words, any errors present are variable errors rather than systematic ones. For any survey measure, the expected value is by definition the same for both sampled and nonsampled units. Post-stratification alone cannot therefore affect the bias of survey estimates, but it will tend to reduce standard errors (to the extent that post-strata are correlated with survey measures). Conversely, the processes that cause under-coverage and nonresponse are likely to be systematic (or at least to have substantial systematic components). Thus, the nature of the survey error that the weighting aims to address is a bias rather than variance. For the assumption to be realistic, the weighting classes should therefore account for the systematic aspects of the relevant process.

A final point to note about weighting to correct the effects of sampling variance, coverage bias or nonresponse bias, is that an assumption is made that the effect is the same on all survey estimates. Indeed, one of the main advantages of weighting is that a single solution can be used for all subsequent analysis and estimation, rather than having to deal with each estimate separately. This contrasts with imputation techniques, where the missing data mechanism can be specified differently for each data item treated. The limitation of this assumption, however, tends to become apparent at the stage of defining weighting classes or calibration constraints. What appears to be a good choice for one survey estimate may be clearly suboptimal for another. Compromises are necessary.

Sample-Based Methods

Sample-based methods are those that rely solely upon information internal to the sampling process and the selected sample. Design weights are by definition sample-based: they are constants determined by the design. Sample-based methods are also often used for nonresponse weighting. They cannot be used to weight for under-coverage or for post-stratification, as these require external population data (see the next section).

Sample-based nonresponse weighting methods typically involve splitting the selected sample into classes, observing the number of responding and nonresponding units in each class and then calculating a weight $w_f = n_f/m_f$ to be applied to each responding unit in class f, as described above. This will ensure that, after weighting, the distribution of the responding sample across the classes is the same as that of the selected sample. Classes can only be defined by variables that are available for all selected sample units. Such variables may come from the sampling frame (e.g., a register or administrative file), may be collected in the course of the survey process (e.g., by interviewer observation in the case of a personal interview survey), or may be linked to the sample from some external source (e.g., Census small-area data in the case of an address-based household survey). A key factor in the success of sample-based survey weighting is the identification and collection of data that may be useful for weighting at the sample selection and field work stages.

Sample-based weighting will be successful at reducing nonresponse bias if three criteria are met:

- Response rates (and therefore weights) must vary over the classes;
- The values of survey estimates (e.g., means, proportions, and regression coefficients) must vary over the classes;
- The values of survey estimates must be similar for both respondents and nonrespondents within each class.

The construction of the classes is therefore important. There are many different methods that can be used to construct sample-based classes. Often the classes are created based purely on intuition or some very limited theory of the likely correlates of nonresponse. Empirical methods can be used to construct classes that provide good discrimination in terms of response rates. Segmentation algorithms (exploiting tree-based methods such as classification tree and regression tree) can be used, with a dichotomous response/nonresponse indicator as the dependent variable. Alternatively, logistic regression can be used and classes created defined by bands of model-predicted response propensity. Sometimes, these methods are used in conjunction with analysis of the survey data that aims to assess the second of the three criteria listed above. The process can be iterative, with classes initially constructed based on the response rate criterion, then amended in the light of analysis of the survey data using the initial classes as a covariate, then perhaps amended again in the light of a reassessment of response rates based on the amended classes.

Other sample-based methods do not involve the construction of explicit classes. For example, a sample unit's weight might be defined as the reciprocal of its predicted response propensity from a logistic regression model of nonresponse. It is also possible to model response propensity using relevant survey process data, such as the number of calls needed to make contact with a sample member, the amount of elapsed time before a postal questionnaire was received, or some indicator of the amount of persuasion that was needed.

Population-Based Methods

Other weighting methods involve creating classes for which the number of population units is known, or can be well estimated. Each responding sample member can then be assigned a weight, $w_g = N_g/n_g$, as described above. It will be noted that if there is any frame under-coverage, this weight will adjust simultaneously for both under-coverage and sampling variance. In fact, it is more common for the weight to be calculated as $w_g = N_g/m_g$, thus also adjusting simultaneously for nonresponse.

Aside from the definition of the weights, two important distinctions can be made between sample-based and population-based methods. The first is the ability of population-based methods to simultaneously adjust for multiple error sources (sampling error, coverage error, nonresponse error). This is usually viewed as an advantage. The second is that sample units are allocated to classes on the basis of their survey responses (or other unit-level data, such as information from the sampling frame), whereas the numerators for the weights usually come from an external source. This is usually viewed as a disadvantage of population-based methods.

The disadvantage lies in the potential for misclassification. Any differences between the survey data and the population data, in the definitions used, the point in time to which the data refer, the method of collecting the data, etc., could potentially result in units appearing in one class in the numerator and a different class in the denominator. If such misclassifications are systematic in any way, then the weighting can actually increase bias rather than reduce it (or, introduce one sort of bias while reducing another sort). It is necessary to consider carefully the potential scope for misclassification and its likely nature. For example, it may be preferable to create classes based on a variable that is only weakly related to survey variables but believed to be insensitive to differences in the population and survey data collection methods (e.g., age and gender for a survey of individuals) rather than one which might be much more strongly related to survey variables but liable to be sufficiently sensitive to data collection methods to cause systematic misclassification (e.g., ethnic group).

Calibration Methods

Calibration refers to a class of methods that produce weights that obey some calibration constraint(s). Understanding of the statistical properties of these methods was developed during the 1990s. During this same period, the development of powerful and flexible software to implement the methods contributed to a rapid growth in the popularity and use of calibration methods.

Post-stratification is one example of a calibration method, the constraint being that the weights must result in weighted post-stratum sample sizes equalling post-stratum population sizes. Thus, if there are H post-strata, there are H calibration constraints. Another possibility is to define the constraints in terms of continuous variables, such as stratum population totals, X_h, or means, $\bar{X}_h$. A common example occurs in business surveys, where the $\{X_h\}$ may be total production (or turnover, or employment), in a stratum perhaps defined by industry sector, the data typically coming from a business register. This is equivalent to ratio estimation (where the weights are typically implicit to the estimation process and the choice of auxiliary variables may be estimate-specific), which can therefore be thought of as a type of calibration. Raking (also known as the raking ratio method, or rim weighting) also falls within the class of calibration methods. Raking is an iterative procedure whereby the responding sample is weighted in turn to two or more different sets of comprehensive and mutually exclusive classes. A stopping rule, usually based upon changes in the weights or in the weighted sample sizes, determines when the iterations are complete. Typically, the different sets of classes are defined as categories of a number of different variables. Thus, the method ensures that the marginal distributions of each variable (but not necessarily the joint distributions) equal the population distribution. Raking enables calibration to multiple variables in situations where post-stratification to the full cross-classification of the variables would be undesirable due to small cell sizes and in situations where the joint population distribution is unknown.

For any type of calibration, some auxiliary data are required to determine the calibration constraints. One advantage of calibration methods for large survey organizations—especially national statistical institutes—is that the use of the same constraints across multiple surveys ensures coherence of the outputs from these surveys, at least in terms of the variables that define the constraints. For example, all national social surveys can be seen to be based on the same age × sex × region breakdown.

The calibration approach requires two main inputs. The first is a set of pre-weights. These are typically the design weights, but could equally be combined design and nonresponse weights, for example. The second is the set of calibration constraints. An algorithm is then used to identify a set of calibration weights that are as similar as possible to the pre-weights while also meeting the calibration constraints. This is achieved by minimising a distance function between the pre-weights and the calibration weights. There is a range of software available to calculate calibration weights.

Combining Methods

It is common, and often desirable, to combine more than one weighting method on a single survey. For example, one could use sample-based class weighting methods for nonresponse, using design weights as pre-weights for the nonresponse analysis, and subsequently carry out post-stratification, using the (design + nonresponse) weights as pre-weights (see Fig. 1). This can be an appropriate way of making the best use of available auxiliary information if different information is available at the level of the selected sample and the population. The calibration approach is often viewed as a natural extension of this idea, combining both design weights and calibration constraints in a standard way. In fact, it is not necessary that the pre-weights in a calibration approach are the design weights. Some researchers have developed nonresponse weights (which themselves may use design weights as pre-weights) and then used those as the calibration pre-weights. There may be advantages in this approach if there are auxiliary data available for the selected sample that are not available at population level to form the calibration constraints. Typically, for many surveys, for example in public-use survey data files, the final weight is the product of three weights: design, nonresponse adjustment, and post-stratification (or calibration) weights.

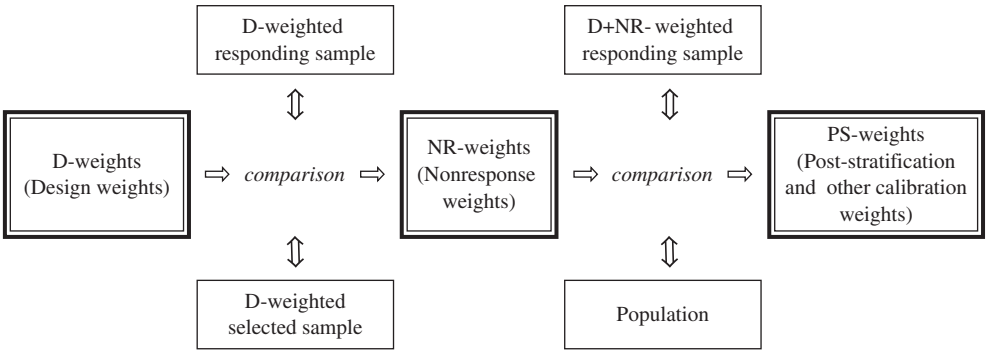


Figure 1 A typical weighting process.

Design-Based and Model-Assisted Methods

The "design-based" and "model-assisted" approaches to inference from survey data are often thought of as two rival schools of thought. In reality, pure design-based inference is only possible in the complete absence of any survey errors other than random sampling variation. In other words, for example, there must either be no nonresponse or nonresponse must be completely random. In any realistic situation, inference relies upon some model(s), though these are often not stated explicitly. Estimation using only design weights makes implicit assumptions (models) about the nature of nonresponse error and coverage error, for example. The development of weights to deal with these sources of error merely makes the models explicit.

Using Weights in Analysis

Impact of Weights

In the case of nonresponse weights particularly (but also design weights), there is a trade-off to be made between the variance and bias of estimates. Variability in the weights will tend to increase the variance of estimates, so the weighting will only be successful at reducing mean square error if this is outweighed by the reduction in bias. The impact of weights on the variance of estimates can be measured by the design effect due to weighting. The squared coefficient of variation of the weights, plus one, provides an approximation to this design effect, under the simplifying assumption that population variance (of the survey measure) does not vary across weighting classes.

It is common for researchers to assess variation in weights and to consider whether the consequent increase in variance is acceptable relative to the expected bias reduction. If it does not appear acceptable, large weights may be trimmed by some more or less arbitrary method, or some weighting classes combined. An important consideration is that bias, unlike variance, acts independently of sample size, so controlling the range of weights tends to be a more important consideration for surveys with small samples than for surveys with large samples.

Types of Weights

Sometimes, survey data sets will contain more than one weight variable, corresponding to multiple purposes or types of analysis. For example, in longitudinal surveys, respondents to wave t may be assigned a cross-sectional weight (for estimates relating to the cross-sectional population at time t) and one or more longitudinal weights (for estimates relating to longitudinal populations at time t and one or more previous time points). On other occasions, data sets may contain component weights (e.g., a design weight, a nonresponse weight, a post-stratification weight, and/or other calibration weight), which must be used in combination. It is not unusual for confusion about the status of weight variables to result in incorrect analysis. On the other hand, provision of multiple weights can often improve the quality of analysis and provide important information for more sophisticated users.

Documentation of Weights

To avoid confusion of the sort mentioned above, it is important that the origin, meaning, and purpose of each weight added to a data set is fully documented. This is an important element of the task of survey documentation and directly affects the usability of survey data. Quality of documentation varies greatly and there are no generally agreed standards.

Applying Weights

Historically, much descriptive estimation based on survey data has used unweighted data. Objections to the use of weights were both practical (it takes time and effort to

develop weights) and ideological (the calculation of weights requires reliance on assumptions that may not be justifiable). Such objections were particularly common in situations where design weights were equal (or almost equal) for all sample units. Of course, ignoring nonresponse involves the stronger assumption that data are missing completely at random (MCAR). The arrival of increased computing power and flexible software addressed many of the practical objections, while the ideological objections were addressed by an increased understanding of the role of weighting and the broader context of statistical inference. By the late 1980s, weighting was usual practice for most public sector social surveys. For analytical estimation, practice remains divided in terms of using weights that are typically developed primarily for descriptive purposes. Alternatives to weighted analysis are to incorporate weighting and calibration variables as covariates in the analysis (not always possible for all analysts) or to model substantive outcomes and missing data outcomes simultaneously.

To use weights in any statistical analysis, the contribution of each responding unit to an estimate should be weighted by the appropriate weight. This is conceptually equivalent to counting each responding unit w_i times when constructing the estimate. In most statistical software packages this is easily achieved by specifying the weight variable prior to specifying the estimation to be carried out. There are, however, differences between packages in the way that weights are treated. Also, some packages allow user specification of how the weights should be treated. The differences do not usually affect point estimates (of means, proportions, correlation coefficients, regression coefficients, etc.) but can have a large impact on estimates of standard errors (and consequently hypothesis testing and model fitting). It is therefore important that weights are specified correctly by users. When analysis is carried out using weighted data it is common practice to report both weighted and unweighted sample sizes. This provides the reader with some information on both the relative precision of estimates (e.g., for subgroups) and the relative contribution of subgroups to estimates for larger domains, including the total study population.

See Also the Following Articles

Population vs. Sample • Stratified Sampling Types

Further Reading

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Am. Statist. Assoc.* **87,** 376–382.

Holt, D., and Smith, T. M. F. (1979). Post stratification. *J. Roy. Statist. Soc. Ser. A* **142,** 33–46.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodol.* **12,** 1–16.

Kish, L. (1992). Weighting for unequal P_i. *J. Official Statist.* **8,** 183–200.

Lundström, S., and Särndal, C.-E. (2001). *Estimation in the Presence of Nonresponse and Frame Imperfections.* Statistics Sweden, Örebro.

World Health Organization Instruments for Quality of Life Measurement in Health Settings

Shekhar Saxena

Mental Health: Evidence and Research, World Health Organization, Geneva, Switzerland

Mark van Ommeren

Mental Health: Evidence and Research, World Health Organization, Geneva, Switzerland

Glossary

health A state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.

health care setting Setting in which services are provided to a population to maintain health and prevent and cure diseases.

human immunodeficiency virus (HIV) The virus that causes acquired immune deficiency syndrome (AIDS).

quality of life Individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns.

World Health Organization United Nations technical agency responsible for health.

The World Health Organization Quality of Life (WHOQOL) instruments assess respondents' perception and subjective evaluation of various aspects of the quality of life. They are designed to measure quality of life related to health and health care. The instruments have been developed within cross-cultural multicenter projects. Experts and lay health care users in each center have been involved in this process to produce reliable and valid instruments applicable across cultures. There are four WHOQOL instruments: WHOQOL-100, a 100-item generic quality of life assessment instrument; WHOQOL-BREF, a 26-item brief version of the generic WHOQOL-100; WHOQOL-HIV, a specific module to assess quality of life in persons who are HIV-positive; and WHOQOL-SRPB, a specific module to assess the spiritual, religious, and personal belief component in quality of life.

Rationale for the Development of Quality of Life Instruments

The World Health Organization (WHO) initiated the development of international quality of life assessment instruments in the 1990s for several reasons. First, because health, according to WHO's constitution, is "a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity," there is a need to go beyond morbidity and mortality and to assess disability, functioning, perceived health, and quality of life. Second, most measures of health status have been developed in a single cultural setting, and the adaptation of these measures for use in other settings is time-consuming and often unsatisfactory. Third, the increasingly mechanistic model of medicine, concerned largely with the eradication of disease and symptoms, reinforces the need to introduce broader, humanistic concerns into

health care. WHO's initiative to develop a quality of life instrument therefore arises from a need for a genuinely international measure of quality of life—taking into account sociocultural diversities—and from commitment to the promotion of a holistic approach to health and health care.

Quality of life: WHO's Concept and Definition

Due to the lack of a universally agreed upon definition of quality of life, the first step in the development of the WHO Quality of Life (WHOQOL) instruments was to define the concept. In 1995, WHO defined quality of life as "individuals' perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards, and concerns." It is a broad-ranging concept, incorporating the person's physical health, psychological state, level of independence, social relationships, personal beliefs, and relationship to salient features of the environment. This definition highlights the view that quality of life refers to a subjective evaluation, which includes both positive and negative dimensions and which is embedded in a cultural, social, and environmental context.

Given that quality of life is a multidimensional construct to be amenable to measurement, it is necessary to identify various components of quality of life. Consultations and investigators from field centers proposed several broad domains assumed to contribute to an individual's quality of life. Each domain was further divided into a series of specific areas (facets and subdomains; Table I). This hierarchical structure of the WHOQOL allows for a quality of life profile of respondents with domain and facet scores.

Table I Composition of the Generic WHOQOL-100: Six Domains and 25 Facets

Domain I: Physical	1. Pain and discomfort
	2. Energy and fatigue
	3. Sleep and rest
Domain II: Psychological	4. Positive feelings
	5. Thinking, memory, learning, and concentration
	6. Self-esteem
	7. Bodily image and appearance
	8. Negative feelings
Domain III: Level of independence	9. Mobility
	10. Activities of daily living
	11. Dependence on medication and treatment
	12. Work capacity
Domain IV: Social	13. Personal relationships
	14. Practical social support
	15. Sexual activity
Domain V: Environmental	16. Physical safety and security
	17. Home environment
	18. Financial resources
	19. Health and social care: availability and quality
	20. Opportunities for acquiring new information and skills
	21. Participation in and new opportunities for recreation/leisure
	22. Physical environment
	23. Transport
Domain VI: Spiritual, religious, and personal beliefs	24. Spiritual, religiousness, and personal beliefs
General facet	25. Overall quality of life and general health perceptions

WHOQOL-100: Generic Quality of Life Assessment

The first WHOQOL instrument to be developed was the WHOQOL-100. The procedure for the development was as follows. First, focus groups were run in several field centers (Bangkok, Bath, Madras, Melbourne, Panama, St. Petersburg, Seattle, Tilburg, and Zagreb) to (i) examine the meaning, variation, and perceptual experience of the quality of life construct in different locations of the world and (ii) test the face validity and comprehensiveness of initial WHOQOL domains and facets, which had been proposed by an international group of experts. Participants in these focus groups were mostly individuals from the general population in contact with health care providers. The data from the focus groups revealed that within each cultural settings quality of life cannot be easily described in terms of one or several words or phrases; rather, it is the breadth and content of quality of life that characterize it. Moreover, it was found that the issues raised by participants in focus groups mostly reflected the issues covered in the WHOQOL domain and facet structure proposed by experts in an initial meeting, but modifications to the facets nevertheless proved necessary.

The next step was to provide a definition for each facet. This consisted of a conceptual definition, a description of various dimensions along which rating can be made for a facet, and a listing of some example situations or conditions that might specify various levels of intensity of a facet. Definitions of each facet were translated into the language of the field centers following a standardized WHOQOL translation method, which is an iterative process of forward and backward translation complemented by review by monolingual focus groups and bilingual experts.

The translated facet definitions were discussed in further focus group work in the field centers. Focus group participants included inpatients, outpatients, informal caregivers, health personnel, and persons from the general population. The focus groups involved (i) discussion on how different facets affect quality of life, (ii) consideration of how to ask about the facets, (iii) rating the "importance" of each facet on a 5-point scale, and (iv) other issues considered important by participants to quality of life. On the basis of the focus groups in different countries, the WHOQOL facet structure was revised.

A question-writing panel was assembled in each of the field centers. Questions formulated in the local language were translated in English to develop a global item pool of approximately 1800 questions. After identifying semantically equivalent and poorly formulated questions, the pool was reduced to approximately 1000 questions. The principal investigator in each of the field centers rank ordered the questions for each facet according to "how much it tells you about a respondent's quality of life in your culture." From the combined rankings for all centers, 235 questions were selected for the WHOQOL pilot instrument. Five-point semantic differential response scales were derived for each of the instrument's language versions according to standardized methodology.

Pilot testing involved the administration of the 235 questions to 250 health care users and 50 healthy respondents in each of the 15 culturally diverse field centers ($N = 4500$). A series of frequency, reliability, inter-item correlation, interfacet, and discriminant validity analyses were run on the pilot data (i) at the level of individual centers, (ii) summarized across individual centers, and (iii) on the pooled global data. The final selection of items took into account a number of desirable features of the facets and items, including the goal of having four items per facet (the minimum number required for scale reliability analyses), the degree of conceptual overlap between items (which was minimized insofar as possible), and the findings of aforementioned psychometric analyses. The resulting 100-item instrument, called the WHOQOL-100, has since been shown to be reliable and valid for use in diverse cultures. The instrument has 24 specific facets (distributed in six domains) as well as 1 general quality of life facet (Table I). The WHOQOL-100 is available in more than 40 languages and available at the following Web site: http://www.who.int/evidence/assessment-instruments/qol.

WHOQOL BREF: Brief Generic Quality of Life Assessment

Although the WHOQOL-100 is a detailed assessment instrument of individual facets relating to quality of life, it may be too lengthy for use in large epidemiological studies or clinical trials in which quality of life is only one variable of interest. In such instances, an assessment tool will be more likely incorporated into studies if it is brief. The WHOQOL-BREF was therefore developed, which is much shorter but still provides summary scores at the domain level. To maintain comprehensiveness in the abbreviated version, one item from each of the 24 WHOQOL-100 facets was included. On the basis of multivariate, psychometric analyses of new and existing data sets, a 26-item scale was developed consisting of one item from each of the 24 facets plus two general items covering overall quality of life and general health. The items were regrouped into four domains: physical health, psychological aspects, social relationships, and environment. Domain scores produced by the WHOQOL-BREF correlate very highly with domain scores of the WHOQOL-100. The WHOQOL-BREF thus performs as a brief quality of life measure.

WHOQOL-HIV: Quality of Life Assessment for Persons with HIV

The WHOQOL-100 and WHOQOL-BREF are generic instruments designed to assess quality of life among people with a variety of health problems. However, people with specific health problems tend to have additional quality of life concerns. For this reason, the WHOQOL Group encouraged the development of disease-specific WHOQOL modules that would be administered with the generic instruments to assess disease-specific quality of life concerns. WHO started by organizing the development of a specific WHOQOL module for use among people with HIV/AIDS. Although increasingly research has been focusing on minority populations in the United States, there had been no instruments to assess the quality of life of people living in developing countries, where the vast majority of people with HIV/AIDS live and where quality of life concerns may be different. HIV-specific instruments developed for use in U.S. and European contexts are difficult to adapt and use in low-income countries.

The WHOQOL-HIV was developed as follows. A consultation of international experts was convened to review the suitability of the generic WHOQOL for assessing people with HIV. Additional facets were proposed to address specific concerns of people living with HIV/AIDS. Focus groups were then conducted at six culturally diverse centers. Participants comprised people with HIV/AIDS, informal caregivers, and health professionals. The aims were to review the adequacy of the WHOQOL-100 for assessing people with HIV/AIDS and to generate additional facets and items for a pilot module. A total of 115 items were proposed, covering 25 new facets for

assessment of quality of life specific to living with HIV/AIDS. The new facets included symptoms of HIV, body image, sexual activities, work, social inclusion, disclosure, death and dying, and forgiveness. In a pilot study, 900 people from six culturally diverse sites completed the WHOQOL-100 along with 115 HIV-specific items. Respondents were HIV asymptomatic, HIV symptomatic, people with AIDS, or asymptomatic without HIV/AIDS. Using standard WHOQOL development methodology, a series of psychometric analyses were conducted—including frequency, interitem correlations, reliability, and multidimensional scaling—to select the best items from the pilot module. This resulted in the selection of 20 items covering 5 new facets for inclusion in a field trial of the WHOQOL HIV module (Table II). The field trial involved administering the WHOQOL HIV field test module to 1334 people with HIV/AIDS from seven culturally diverse centers (Australia, Brazil, Italy, Thailand, Ukraine, and two centers in India—Bangalore and New Delhi). Experience with the module demonstrated good reliability and good discriminant validity, with poorest quality of life found for those who reported being the least healthy. The module provides a means for quality of life assessment for HIV/AIDS in diverse cultural settings. The WHOQOL-HIV is available at http://www.who.int/mental_health/resources/evidence_research/en.

WHOQOL-SRPB: Spirituality, Religion, and Personal Beliefs as Components of Quality of Life Assessment

Traditionally, generic assessments of quality of life do not routinely address specific aspects of religion, spirituality, or existential well-being. However, there is increasing evidence that peoples' beliefs may be important contributors to quality of life. Spirituality may be especially important in particular cultural and ethnic groups. For this reason, WHO developed a WHOQOL module for assessment of spirituality, religion, and personal beliefs (SRPB) in different cultures using standard WHOQOL instrument development methodology. Briefly, an international consultation was conducted to generate potential facets related to SRPB. Experts in the field participated, representing major religions of the world, as well as participants

Table II Additional WHOQOL-HIV Facets and Items[a]

Symptoms
1. How much are you bothered by any unpleasant physical problems related to your HIV that you may have?
2. To what extent do you fear possible future (physical) pain?
3. To what extent do you feel any unpleasant physical problems related to your HIV infection prevent you from doing things that are important to you?
4. To what extent are you bothered by fears of developing any physical problems?
Social inclusion
5. To what extent do you feel accepted by the people you know?
6. How often do you feel that you are discriminated against because of your health condition?
7. To what extent do you feel accepted by your community?
8. How much do you feel removed/alienated/emotionally distant from others/those around you?
Forgiveness
9. How much do you blame yourself for your HIV infection?
10. How bothered are you by people blaming you for HIV status?
11. How guilty do you feel about being HIV positive?
12. To what extent do you feel guilty when you need the help and care of others?
Fear of the future
13. To what extent are you concerned about your HIV status breaking your family line and your future generations?
14. To what extent are you concerned about how people will remember you when you are dead?
15. To what extent does any feeling that you are suffering from fate/destiny bother you?
16. How much do you fear the future?
Death and dying
17. How much do you worry about death?
18. How bothered are you by the thought of not being able to die the way you would want to?
19. How concerned are you about how and where you will die?
20. How preoccupied are you about suffering before dying?

[a] Facet scores are calculated as average scores across four items.

who did not profess any religious faith. Various facets were generated and then reviewed by focus groups in 15 centers (Egypt, Brazil, Uruguay, Argentina, Spain, Italy, the United Kingdom, Lithuania, Turkey, Israel, India, Malaysia, Thailand, China, and Japan). A total of 92 focus groups at these sites involved input from 701 people. The aim of the focus groups was to ensure the applicability of the suggested facets to quality of life and to generate items for inclusion in a questionnaire. Based on the qualitative data and the quantitative importance ratings, the relevance of 15 facets was confirmed. For each of these facets, 7 items were generated. The WHOQOL-SRPB pilot module thus consisted of 105 items, which were added to the WHOQOL-100 during pilot testing. Using standard WHOQOL instrument development methodology, the 15-facet pilot module was reduced to an 8-facet, 32-item SRPB module for field testing (Table III). The facets address spiritual connection, experiences

Table III SRPB Facets and Corresponding Items[a]

Connectedness to a spiritual being or force	
1.	To what extent does any connection to a spiritual being help you to get through hard times?
2.	To what extent does any connection to a spiritual being help you to tolerate stress?
3.	To what extent does any connection to a spiritual being help you to understand others?
4.	To what extent does any connection to a spiritual being provide you with comfort/reassurance?
Meaning of life	
5.	To what extent do you find meaning in life?
6.	To what extent does taking care of other people provide meaning of life for you?
7.	To what extent do you feel your life has a purpose?
8.	To what extent do you feel you are here for a reason?
Awe	
9.	To what extent are you able to experience awe from your surroundings (e.g., nature, art, and music)?
10.	To what extent do you feel spiritually touched by beauty?
11.	To what extent do you have feelings of inspiration/excitement in your life?
12.	To what extent are you grateful for the things in nature that you can enjoy?
Wholeness and integration	
13.	To what extent do you feel any connection between your mind, body, and soul?
14.	How satisfied are you that you have a balance between mind, body, and soul?
15.	To what extent do you feel the way you live is consistent with what you feel and think?
16.	How much do your beliefs help you to create coherence between what you do, think, and feel?
Spiritual strength	
17.	To what extent do you feel inner spiritual strength?
18.	To what extent can you find spiritual strength in difficult times?
19.	How much does spiritual strength help you to live better?
20.	To what extent does your spiritual strength help you to feel happy in life?
Inner peace/serenity/harmony	
21.	To what extent do you feel peaceful within yourself?
22.	To what extent do you have inner peace?
23.	How much are you able to feel peaceful when you need to?
24.	To what extent do you feel a sense of harmony in your life?
Hope and optimism	
25.	How hopeful do you feel?
26.	To what extent are you hopeful about your life?
27.	To what extent does being optimistic improve your quality of life?
28.	How able are you to remain optimistic in times of uncertainty?
Faith	
29.	To what extent does faith contribute to your well-being?
30.	To what extent does faith give you comfort in daily life?
31.	To what extent does faith give you strength in daily life?
32.	To what extent does faith help you to enjoy life?

[a] Facet scores are calculated as a average scores across four items.

of awe and wonder, wholeness and integration, meaning of life, spiritual strength, inner peace/serenity/harmony, hope and optimism, and faith. The SRBB module should be administered together with the WHOQOL-100 or WHOQOL-BREF. The WHOQOL-SPRB is available at http://www.who.int/mental_health/resources/evidence_research/en.

Use of WHOQOL Instruments

The WHOQOL instruments are being applied throughout the world in various ways. The WHOQOL is frequently included in epidemiological surveys and randomized clinical trials. The WHOQOL-BREF is used as a routine outcome measure for health issues ranging from pain management and cancer treatment to severe depression. The WHOQOL is also used to study the quality of life concept in greater depth. Finally, the WHOQOL is used in training and program evaluation efforts to focus attention on health-related issues beyond mortality and morbidity.

See Also the Following Articles

Biomedicine • Nursing • Religious Affiliation and Commitment, Measurement of

Further Reading

Sartorius, N., and Kuyken, W. (1994). Translation of health status instruments. In *Quality of Life Assessment: International Perspectives* (J. Orley and W. Kuyken, eds.). Springer-Verlag, Berlin.

WHOQOL Group (1995). The World Health Organization Quality of Life assessment (WHOQOL): Position paper from the World Health Organization. *Social Sci. Med.* **41,** 1403–1409.

WHOQOL Group (1998a). The World Health Organization Quality of Life assessment (WHOQOL): Development and general psychometric properties. *Social Sci. Med.* **46,** 1569–1585.

WHOQOL Group (1998b). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychol. Med.* **28,** 551–558.

WHOQOL HIV Group (2003a). Initial steps to developing the World Health Organization's Quality of Life instrument (WHOQOL) module for international assessment in HIV-AIDS. *AIDS Care* **15,** 347–357.

WHOQOL HIV Group (2003b). Preliminary development of the World Health Organization's Quality of Life HIV instrument (WHOQOL-HIV): Analysis of the pilot version. *Social Sci. Med.* **57,** 1259–1275.

WHOQOL HIV Group (2003c). World Health Organization's instruments in quality of life measurement in health settings. *AIDS Care*.

World Health Organization (2002). *WHOQOL SRPB Users Manual.* World Health Organization, Geneva.

Wundt, Wilhelm

David J. Murray
Queen's University, Kingston, Ontario, Canada

Glossary

culture A word used by Herder, and later widely adopted in the social sciences generally, referring to the corpus of beliefs of a community concerning customs, morals, religions, myths, and the arts; for Herder, the written literature of a (literate) community was a reflection of its culture.

experimental psychology That branch of psychology concerned mainly with the influences of physical stimuli on events taking place within the mind and body of a single individual; experimental psychologists usually believe that data collected in a laboratory can be organized within a scientific framework that includes hypotheses framed in physiological and/or mathematical terms.

social psychology That branch of psychology concerned mainly with the influences of other people on the events taking place within the minds and bodies of one or more individuals; research involving investigative manipulations of the behavior of individuals who are members of small groups is often described nowadays, particularly in North America, as being an aspect of "experimental social psychology."

Volk Herder's word for a group of people united in terms of a common culture.

Völkerpsychologie A term invented by Lazarus and Steinthal, and quickly adopted by Wundt, that referred to the scientific attempt to provide a history of the means whereby a *Volk* attained its culture.

Wilhelm Maximilian Wundt (1832–1920), at Leipzig in 1879, founded the first Institute of Experimental Psychology in a university setting. But, throughout his career, he struggled to be clear himself about the relationship between "experimental psychology" (which for him overlapped extensively with "physiological psychology") and *Völkerpsychologie* (his word for something only approximating the modern expression "social psychology," as will be explained below). The former is concerned with scientific accounts of what transpires within an individual's mind during a particular window of time (which might last only seconds), while the latter is concerned with scientific accounts of what transpires within and between groups composed of individuals. This explains why a distinction has been made between "intra-individual" and "inter-individual" psychology in the headings listed above.

Introduction

Wundt was born on August 16, 1832, in the small town of Neckarau, near Mannheim, Germany; his father was a Lutheran minister who moved from Neckarau when Wundt was one year old to various positions before settling more permanently in Heidelsheim in central Baden from 1836–1844. When Wundt was aged 14, he entered the Gymnasium (a high school for students intending to go to university) in nearby Heidelberg, and in 1851, at the age of 19, he spent a year studying medicine at the University of Tübingen before returning to Heidelberg to take up full-time medical studies at the university there. While still a student, he published an article (on salt concentration in the urine), and his Ph.D. dissertation, on neurophysiology, was defended in 1856 while he was gaining medical experience in Karlsruhe. There he also served briefly as a research assistant to Ewald Hasse before moving to Berlin to study physiology with J. Müller (1801–1858) and E. du Bois-Reymond (1818–1896). He then returned to Heidelberg to study with H. von Helmholtz (1821–1894). There, he wrote a second thesis (*Habilitation*) in 1857 that allowed him to present lectures on physiology at Heidelberg. In the following year, he

entered Helmholtz's service as a research assistant; he worked with Helmholtz from 1858 to 1863.

He was to stay in Heidelberg for a total of 17 eventful years, from 1857 to 1874. Not only did he work with Helmholtz, but he was promoted from *Privatdozent* (a rank allowing him to give lectures but not be on full-time faculty) to *ausserordentlicher Professor* (roughly, associate professor on full-time faculty) teaching anthropology and medical psychology, a rank he held from 1864 to 1874. In addition, he participated actively in the politics of the country then known as Baden, whose legislative offices were in Heidelberg; he was embroiled in the question of German unification, because Baden, in 1864, when Wundt was elected to the Baden legislature, was undecided as to whether to confederate with Saxony, Bavaria, Prussia, and other German-speaking countries.

However, he left politics after four years, partly because he found that political bickering was so time-consuming and enervating that he preferred to make his life's goal one of academic, rather than political, achievement. In 1872, at the age of 40, he married Sophie Mau (born January 23, 1844, in Kiel; died April 15, 1912, in Leipzig). In 1874, he was called to be *ordentlicher Professor* (roughly, full professor and chairman of a department) at the University of Zürich in Switzerland, but this turned out to be a temporary position prior to Wundt's receiving the call to be *ordentlicher Professor* at the University of Leipzig. At Zürich, his title had been that of Professor of Inductive Philosophy; at Leipzig, his title was that of Professor of Philosophy. His academic career at Heidelberg had included the teaching of anthropology and medicine. The possession of such an unusually well-rounded background of knowledge was exceptionally useful when it came to his appealing to the authorities at Leipzig for the foundation of his Institute of Experimental Psychology.

During his years at Heidelberg, he may have formulated, independently of Helmholtz, the idea that most of our perceptions are based on unconscious inferences. But Wundt's major achievement was to have written two books on psychology that were unusually far-reaching in scope. The first was his *Beiträge zur Theorie der Sinneswahrnehmung* (Contributions to the Theory of Sense Perception) of 1862, a work whose Introduction delineates a plan for a future science of psychology, one that would include *Völkerpsychologie*. The second was his *Vorlesungen über die Menschen und Tierseele* (Lectures on Human and Animal Psychology) of 1863, a work in two volumes, the second of which included the longest discussion of *Völkerpsychologie* in his early writings. It is important to note that neither book has been translated into English, although in 1961, Shipley translated the Introduction to the *Beiträge*. The *Vorlesungen* did come out in a second edition almost 30 years later and was translated by Creighton and Titchener, but the second edition was radically different from the first edition and omitted almost all reference to *Völkerpsychologie*.

Wundt's Early Views on Intra-Individual and Inter-Individual Psychology

Wundt's early views on the nature of psychology cannot be properly evaluated without some knowledge of the historical background with respect both to intra-individual and inter-individual psychology against which he himself evaluated his own contributions. With respect to intra-individual psychology, there are many secondary sources, including the work of Boring and Murray, that describe how research on sensation and perception by E. H. Weber (1795–1878), G. T. Fechner (1801–1887), F. C. Donders (1818–1889), and, of course, Helmholtz provided the catalytic impulse that transformed armchair psychology, with its tradition going back through the associationists to Descartes and Aristotle, into the kind of experimental psychology that Wundt himself would propagate once installed at Leipzig. Wundt also acknowledged the importance of the well-intentioned mathematical psychology of J. F. Herbart (1776–1841); Herbart's system was an analogy in mental science to Newton's system in physical science. But when Wundt described Herbart's intra-individual psychology in the first edition (1874) of his great textbook, the *Grundzüge der physiologischen Psychologie* (Principles of physiological psychology), he criticized it, on mathematical grounds, so harshly that Wundt probably helped to make Herbart's mathematical psychology unfashionable for the whole of the 20th century.

Wundt's views on inter-individual psychology, as crystallized in his changing conceptions of the role of *Völkerpsychologie* vis-à-vis psychology as a whole, were initially based on some conceptions of group psychology put forward by T. Waitz (1821–1864), H. Steinthal (1823–1899), and M. Lazarus (1824–1903), who in turn professed themselves to have been favorably influenced by Herbart. The details of Herbart's influence on Waitz, Steinthal, and Lazarus have been provided by Ribot and by Danziger; it should be noted that Lazarus and Steinthal invented the word *Völkerpsychologie* and used it in the title of a new journal, the *Zeitschrift für Völkerpsychologie und Sprachwissenschaft* (Journal of *Völkerpsychologie* and Language Theory), that they founded in 1859. It is not well known that Herbart tried to extend the mathematical psychology that he had worked out for intra-individual psychology to the psychology of groups.

Herbart called the science concerned with people behaving in groups *Staatswissenschaft*, that is, a science

concerned with the state. Nowadays, such a title might be found in a book of readings concerning political science rather than social psychology. Nevertheless, it was Herbart's belief that his intra-individual psychology could be carried over, *mutatis mutandis*, from the individual to the group, and, for Herbart, it was difficult to disentangle group behavior from political behavior. The subject matter of *Staatswissenschaft* concerned groups that ran the gamut from being hierarchically organized, with extreme power being given to the people at the top of the hierarchy, to being disorganized, with nobody having any real power.

Herbart's mathematical psychology, as applied to the behavior of political groups, extended the "statics" and "mechanics" of the individual mind to the state itself. Any state is composed of many individuals who are mainly competing rather than cooperating; at any moment in historical time, a condition of equilibrium will be arrived at in which the majority of individuals are able to combine their competitive and altruistic inclinations. But disequilibrium can easily be introduced into the system by political unrest, outside invasion, or new ideologies, and peace will only be re-achieved when a new equilibrium condition is arrived at, one that will be attained when some unruly individuals are temporarily or permanently deprived of power within the group. Moreover, just as individual ideas within an individual's mind can either remain independent of each other or fuse with each other, so individuals within a group can remain independent of each other or fuse into subgroups ("pressure groups" or "break-away groups," as we might now say).

Wundt, in his 1862 Introduction to the *Beiträge*, began his discussion of the psychology of groups (calling it, in agreement with others, "sociology") by saying that the best way of collecting the relevant data was to use a broadened method of observation. But the "broadened method," he maintained, would not include the Newton-like equations proposed by Herbart. Wundt claimed that sociology had been created "by an extension of the results of the psychological observation of the individual, to the lives of the nations. Now, however, this science has gradually begun to free itself from the basis on which it rests, and to establish its own foundation. This foundation consists in the determination of a great number of facts through *statistics*."

By "statistics" in this sentence, Wundt meant the collection of data that provided evidence for the relative frequency of physiological characteristics (e.g., measurements of chest girth) or major life events (e.g., divorce) associated with individuals within a group. For example, by 1862, data had been collected on the frequency of suicides, not only within a given nation, but also within given age ranges, genders, seasons of the year, and so on. The person most responsible for collecting such data was the mathematician L. A. Quetelet (1796–1874), who obtained much of his data from census information collected by the Belgian government. Quetelet argued that many data sets concerning individuals showed a symmetric variability around a common arithmetic mean and were probably distributed, therefore, in a Gaussian manner. Quetelet thus provided a methodology that served as a background for the introduction not only of descriptive, but also of inferential, statistics into the social sciences. Wundt, along with Lazarus and Steinthal, was indirectly responsible for the relegation of Herbart's views on the social sciences into near oblivion. Wundt also played a small but important part in ensuring that Quetelet's views about the importance of arithmetic means led to those means' being adopted as the most widely used measures of central tendency in reports concerning groups.

It might be noted that Fechner in 1897 would later maintain, contrary to Quetelet, that the measure of central tendency that best described many collectives was the mode, rather than the mean. It might also be noted that, as will be described below, Wundt's moving away from conceiving groups as political entities led him to a pre-Herbartian conception of groups as cultural entities.

Wundt's Career at Leipzig

When Wundt arrived in Leipzig in 1875, he was riding the crest of a newly acquired fame resulting from the success of his 1874 textbook of physiological psychology. The first edition of this book was concerned not only with describing the latest knowledge about the nervous system and the brain but also with describing new trends in experimental psychology, including psychophysics, studies of sensory perception, and research on reaction time. Wundt's textbook went through a total of six editions between 1874 and 1908–1911, but only the first volume of the fifth edition was made available in English. In the third edition, published in 1887, at the height of the Institute's fame, there is no reference to *Völkerpsychologie* in the index.

It was probably his graduate students who suggested to Wundt that the classroom he had been given in one of the buildings on the Leipzig campus serve as the foundation of an expanded teaching laboratory. This provided the seed that led Wundt to apply for an Institute (German, *Seminar*) to be devoted to experimental psychology in particular. Such an arrangement would allow more graduate students to study with Wundt, and Wundt himself then also receive financial support from the University to reimburse teaching and research personnel, as well as to purchase apparatus and supplies. All his earlier research had been self-funded. The impetus that an Institute would give to the furtherance of experimental research would also provide material for a journal for the dissemination of this research; the first issue of his new journal *Philosophische Studien* appeared in 1881, and the third

issue included a report of the first Ph.D. dissertation defended at the Institute, namely, that of M. Friedrich in 1883.

Wundt's subsequent career was devoted to the expansion of the Institute (by the early 1880s, it had grown from one classroom to about seven, and in 1896, the Institute was moved to occupy the whole floor of a new building); to the development of the journal (in 1901, its name was changed to *Psychologische Studien*); to the teaching of practical classes in experimental psychology (following a trend in the science departments of many German universities, a trend quickly imitated in other countries); to the cordial reception of foreign visitors (who would study the "new psychology," as it came to be called); and to the continuation of his writings, both on philosophy and on psychology. The main works that occupied his final two decades at Leipzig included the sixth edition of his textbook, the writing of his autobiography (published in 1920), and the first of the many editions of his multivolume work entitled *Völkerpsychologie* that appeared between 1900 and 1920.

Following his retirement from teaching in 1917, Wundt died in his home on August 31, 1920. He was preceded in death by his wife by eight years; she had borne three children, Eleonora (born 1876), Max (born 1879), and Lilli (born 1880, who died at age four). Wundt received many honorary degrees; these are listed by Meischner and Eschler.

Wundt's Original Research Contributions: Intra-Individual Psychology

This section will be short, partly because there are many secondary sources available that discuss Wundt's contributions to experimental psychology, and partly because the focus of the present article is on his contributions to social measurement. Nearly all of the research performed at Leipzig was carried out in the context of practical classes or dissertations. With respect to practical classes, Wundt would think of several research projects and allow small groups of graduate students, led by a more experienced student or research assistant, to carry out these projects, making use of the equipment Wundt had acquired. With respect to dissertations, these could be experimental or philosophical, and dissertations of both kinds appeared, in condensed form, in *Philosophische Studien* or, after 1901, in *Psychologische Studien*. The research contributions of Wundt himself that are best remembered today are his tridimensional theory of feeling and his analyses of the psychological processes involved in the perception of very briefly presented visual information. Murray has described Wundt's attempt to integrate the natural sciences (*Naturwissenschaften*) with the human sciences (*Geisteswissenschaften*) and has summarized Wundt's views on the dangers of introspection as a scientific method, his views on "psychic causality," his so-called "structuralism," and his contributions to "voluntarism" as an approach to psychology. His attempt to integrate the sciences included the incorporation of both physiological psychology and *Völkerpsychologie* into a general science of mind and behavior.

Wundt's Original Research Contributions: Inter-Individual Psychology

Wundt wrote little on *Völkerpsychologie* during the first half of his 41-year residence at Leipzig (from 1879 to 1900). There was, however, one article in *Philosophische Studien* in which Wundt (1886) stressed that the psychological laws determining the behavior of groups were not necessarily of the mechanical kind associated with Herbart, but were laws that were relatively independent of the physiological events taking place in the brains of individuals and more dependent on physical events that determined the environment, including the linguistic environment in which an individual found himself. But, in the course of writing the various editions of his books on logic and on ethics, Wundt kept up with new findings reported by anthropologists and others exploring as yet unknown corners of the world, and in particular, he kept track of data concerning the languages, customs, morals, myths, laws, and religious beliefs of the peoples indigenous to those regions. All this information would find its appropriate place in the major work of the last half of his stay at Leipzig, namely, the *Völkerpsychologie*.

The story of the various parts and editions of this work is extremely complicated; we cut through the tangle to say that in the final publication of the book in 10 volumes, Volumes 1 and 2 were concerned with Language and were translated into English in 1973. Volume 3 was concerned with Art, Volumes 4, 5, and 6 with Myth and Religion, Volumes 7 and 8 with Society, Volume 9 with Law, and Volume 10 with Culture and History. However, Volumes 1 to 6 represented third or second editions of earlier parts of the *Völkerpsychologie*, and Volumes 7 to 10 appear to have been first editions. To add to the confusion, Wundt published a separate one-volume work in 1912 entitled *Elemente der Völkerpsychologie: Grundlinien einer psychologischen Entwicklungsgeschichte der Menschheit* (Elements of *Völkerpsychologie*: Outlines of a psychological history of the development of mankind).

This book was translated into English by Schaub in 1916, with the main title rendered as *Elements of Folk*

Psychology, a translation that was acceptable in Schaub's time, but is now seriously misleading. This is because the expression "folk psychology" has taken on a new meaning in the final decades of the 20th century. It is found in books on the philosophy of mind and refers to the preconceptions that ordinary people have about how the mind works. Nowadays, a reasonable translation of *Völkerpsychologie*, as used by Wundt, might be "the psychology of peoples." It is not to be translated as social psychology; some of the most widely read books of the early 20th century that contained the words social psychology in the title, such as the textbooks of Allport, McDougall, and Thouless, referred nowhere to Wundt's *Völkerpsychologie*. Nor is it to be translated as "cross-cultural psychology," which usually deals with topics such as adaptations by individual groups to emigration, or the effects on indigenous peoples of new technologies.

One of Wundt's own definitions of *Völkerpsychologie*, as given in the Introduction to the first edition of Volume 1, was as follows:

> *By this term, we can denote that mixture of actual observations, received theories, and putative facts of which the representatives of individual scientific disciplines avail themselves when they are unable to offer an appropriate psychological interpretation [for their data].*

This definition almost moves *Völkerpsychologie* out of the realm of psychology and into the realms of archeology, anthropology, and history itself.

Three main comments can be made about the multivolumed book entitled *Völkerpsychologie*. First, in his resistance to a Herbartian approach, Wundt actually moved to a position that had been expressed by writers prior to Herbart who had been concerned with the evolution of human thought. Danziger has explained how the term *Volk* was given life in Germany in the writings of J. G. Herder (1744–1803), for whom a *Volk* was a group united, not so much in terms of a political power hierarchy, or even in terms of a legislated system of moral rules, but in terms of what Herder called a "culture." In Herder's view, the language common to the group provided a medium that bonded the separate elements of that culture. Many answers to traditional philosophical questions concerning morality, truth, and art were thereby embedded in the culture of a community, and Herder had himself believed that those answers in turn depended on the physiological characteristics of, and environmental influences on, the individuals in that community. Wundt's emphasis on the linguistic, artistic, mythic, religious, and moral determinants of the evolution of the mentality of a community at any period in its historical development is coherent with Herder's definition of a culture as being identifiable with the mentality of a *Volk*. Wundt's return to Herder's notion of a cultural community had also been anticipated by azarus and Steinthal, according to Danziger.

Second, Wundt's *Völkerpsychologie* differed both from Herbart's system and from modern psychological systems, such as behaviorism and Gestalt psychology, insofar as each of these last three systems is mainly concerned with the scientific description and prediction of behavioral and mental events in terms of measurements, such as the strength or duration of stimuli, the degree to which wholes subsume parts, and the latency or vigor of vocal or motor responses. These schools are not particularly interested in the contents of thoughts or in the qualitative (as opposed to quantitative) nature of responses. It is in fact possible, as Karl Bühler suggested, to divide systems of psychology into two kinds, one concerned with the contents of thoughts of individuals and one that is relatively unconcerned with contents and more concerned with providing a general scientific context appropriate to a mental science. In modern times, psychoanalysis is the school most representative of the first kind of system, and behaviorism is the school most representative of the second kind of system.

Wundt's *Völkerpsychologie* clearly can be classified as an example of the first kind of system, and Herbart's as an example of the second kind. But modern social psychology and cross-cultural psychology, along with modern sociology, contain investigations of both kinds; for example, the degree of successful adaptation by immigrants to new cultures varies not only with quantitative determinants such as the length of stay since their arrival in the new culture, but also with qualitative determinants such as the country of origin of the immigrants, their first language, and the generation (first or second) of immigrant. Research investigations like these are difficult to classify neatly under either of Bühler's headings. The fact that Wundt's *Völkerpsychologie* is so relentlessly content-bound as to be difficult to fit into any system that attempts quantification of its data has pushed it to the sidelines of the history of the social sciences.

Third, one can flip through the pages of any of the volumes of the *Völkerpsychologie*, or of the single-volumed *Elements of Folk Psychology*, and find very few numbers and certainly no equations or formulas. Wundt's early enthusiasm for the collection of statistical data *à la Quetelet* appears to have been replaced by a principled concern with qualitative matters when he came to write the *Völkerpsychologie*. As a consequence, he still wished to determine the general causes ("scientific laws") that would explain his data. In fact, he developed the theory that the evolution of a culture went through several separate stages from the primitive to the sophisticated; for example, he saw the history of a religious system as developing from fear-inspired nature worship, through a stage of demon worship followed by a stage of hero worship, to its final appearance in a polytheistic or monotheistic form of god or goddess worship. He also developed the theory that the evolution of language went

through a mimetic phase prior to its speech phase. The data on which he based these theories are assembled in the multi-volumed *Völkerpsychologie*; his schematic outline of the history of the evolution of religious, moral, family, and legal systems within cultures was provided in the one-volume work entitled, in English, *Elements of Folk Psychology*.

Wundt's Influence on Social Measurement

Although Wundt played a part in encouraging the collection of statistical data as one of the responsibilities of a social scientist, his own contribution to social psychology was to have emphasized the importance of cultural beliefs in determining the behavior and thinking patterns of an individual member of that culture. As a consequence, Wundt played little part in the advancement of experimentation as a method in social psychology, or in the advancement of the use of correlational techniques as adjuncts both to descriptive and to inferential statistics, or in the advancement of any approach to sociology or social psychology based on an ideology. In the laboratory, he encouraged his students to analyze their data in a quantitative manner, including the calculation of measures of variability, as well as of central tendency, and he admired Fechner's use of Gaussian assumptions in the calculation of absolute or differential thresholds in tasks involving sensory discriminations. But even though, in his Introduction to the *Beiträge*, Wundt had forcefully advocated his view that the data of experimental psychology should include the data of child psychology, animal psychology, and *Völkerpsychologie*, he himself only investigated the last of these in detail and then in such a way that he made little use of the descriptive statistics that he himself had recommended in that Introduction. Wundt's contribution to social measurement was that he laid a solid grounding for the institutionalization, within academia, of the mental sciences alongside the physical and natural sciences. Once this had been achieved, the way had been paved for the institutionalization, within academia, of social psychology, cross-cultural psychology, the psychology of religion, and other branches of psychology concerned with the behavior of individuals within groups.

See Also the Following Articles

Experiments, Overview • Laboratory Experiments in Social Science • Social Psychology

Further Reading

Boring, E. G. (1950). *A History of Experimental Psychology.* 2nd Ed. Prentice-Hall, Englewood Cliffs, NJ.

Boudewijse, G.-J., Murray, D. J., and Bandomir, C. A. (2001). The fate of Herbart's mathematical psychology. *Hist. Psychol.* **4,** 107–132.

Bringmann, W. G., and Tweney, R. D. (1980). *Wundt Studies.* C.J. Hogrefe, Toronto.

Danziger, K. (1983). Origins and basic principles of Wundt's *Völkerpsychologie*. *Brit. J. Soc. Psychol.* **22,** 303–313.

Kwak, K., and Berry, J. W. (2001). Generated differences in acculturation among families in Canada; A comparison of Vietnamese, Korean, and East-Indian groups. *Int. J. Psychol.* **36,** 152–162.

Meischner, W., and Eschler, E. (1979). *Wilhelm Wundt.* Urania–Verlag, Leipzig.

Murray, D. J. (1988). *A History of Western Psychology.* 2nd Ed. Prentice-Hall, Englewood Cliffs, NJ.

Ribot, T. A. (1886). *German Psychology of Today, The Empirical School.* 2nd Ed. (J. M. Baldwin transl.). Scribner, New York.

Rieber, R. W. (ed.) (1980). *Wilhelm Wundt and the Making of a Scientific Psychology*. Plenum Press, New York.

Wundt, W. (1886). Über Ziele und Wege der Völkerpsychologie. [On the aims and methods of Völkerpsychologie.] *Philosoph. Stud.* **4,** 1–27.

Wundt, W. (1916). *Elements of Folk Psychology: Outlines of a Psychological History of the Development of Mankind.* (E. L. Schaub transl.). Macmillan, New York. [Original work published 1912.]

Wundt, W. (1961). Introduction to the Beiträge zur Theorie der Sinneswahrnehmung. In *Classics in Psychology.* (T. Shipley, transl.). Philosophical Library, New York. [Original work published 1862.]

Wundt, W. (1973). *The Language of Gestures* (A. Blumenthal ed., J. S. Thayer, C. M. Greenleaf, and M. D. Silberman, transl.). Mouton, The Hague, the Netherlands. [Original work published 1921.]

Yule, George Udny

Leslie Hepple
The University of Bristol, Bristol, UK

Glossary

autoregressive model A time-series model in which the value of a series at time t depends on its past values, together with a random disturbance. A second-order autoregressive model is thus $y_t = a_1 y_{t-1} + a_2 y_{t-1} + e_t$, $t = 1, \ldots, T$.

correlation coefficient A statistical measure of the degree of association or covariation between two series. A value of $+1$ indicates perfect positive association, -1 indicates perfect negative association, and 0 indicates complete independence.

first-differencing The construction of a time series based on changes (or differences) in level from one period to the next, in the form $z_t = y_t - y_{t-1}$. Higher order differencing may also be employed.

multiple regression A statistical model in which the values for a dependent variable (y) are explained as a function (usually linear) of several exogenous or causal variables (x_1, x_2, x_3, etc.) and a random disturbance term: $y_t = b_0 + b_1 x_{t1} + b_2 x_{t2} + b_3 x_{t3} + e_t$, $t = 1, \ldots, T$.

partial correlation A correlation between two variables after the roles of other influencing factors have been taken into account, or "partialled out."

George Udny Yule (1871–1951), British statistician, was one of the key innovative figures in the generation of the work of Karl Pearson, William Sealy Gosset (who used the pseudonym "Student"), and Ronald A. Fisher, who constructed the basis of statistical theory in the early 20th century. Yule worked closely with Pearson on correlation and regression in the 1890s, but soon established an independent role. He contributed seminal ideas in several arenas related to the growth of quantitative social science. His work on multiple regression and correlation was applied to social policy questions and he formulated the first use of conditioning on exogenous variables to assess policy instruments. His work was also formative in the development of time-series analysis, in autoregressive and differenced series, as a precursor of modern cointegration analysis. His work on contingency table analysis broke free of Pearson's biometric framework, and again laid key foundations for later work.

Biography and Major Contributions

George Udny Yule was born in Scotland in 1871, into a family with deep Scottish roots and a scholarly tradition. Yule is a Scottish surname and Udny was a family name ultimately deriving from a place in Aberdeenshire. George Yule's grandfather had been a Persian and Arabic scholar, his father produced a definitive edition of Marco Polo's travels, and both his father and uncle (an administrator in India) received knighthoods. Yule was educated at Winchester, which placed strong emphasis on classical scholarship, and Yule's later facility at writing Latin verses about the theory of small samples must derive from these years. But it was science and engineering that appealed to the young man, and he went on to study engineering at University College, London University, graduating in 1890 at age 19. He then spent 2 years in the engineering workshops before a spending a year doing research on the physics of electrical waves with Professor Heinrich Herz in Bonn, Germany. Yule's first published papers (at the age of 22) were on this physics research, but in the summer of 1893, he returned to London and changed his research field to statistics.

Yule became a demonstrator (assistant) to Karl Pearson, a Professor of Applied Mathematics at University College, and remained in this position until 1899. Pearson was then at his most innovative, constructing measures of statistical inference to implement Francis Galton's

ideas on correlation and inheritance. During the 1890s, Yule was a close collaborator with Pearson, making advances in net or partial correlation, multiple correlation, and regression. Although Yule had the title of Assistant Professor, his salary was extremely low, and early in 1899, he left for an administrative post with a London examination board. However, from 1902 to 1909, he also held a (part-time) Newmarch Lectureship in Statistics, and also took on other work. The year 1912 was a turning point for Yule; he was appointed a lecturer in statistics at Cambridge University, and he remained there for the rest of his career (apart from civil service for 4 years during World War I). He was elected a Fellow of the Royal Society in 1922, and President of the Royal Statistical Society from 1924 to 1926. He retired at age 60 with heart problems, but continued to do a research until his death in 1951.

Although he was a close associate of Pearson during the 1890s, Yule soon established his own directions, independent of his biometrically oriented mentor. The two men were joint authors on only one paper. Yule became interested primarily in social, economic, and epidemiological applications of statistical methods, and these required assumptions and perspectives different from those of biometrics. Yule developed such methods, and in doing so, clashed with his mentor several times. After he left University College in 1899, he and Pearson followed independent agendas. Yule and Pearson had very different personalities and interests. Yule was an individual research worker: he had no interest (or aptitude) for setting up a laboratory, establishing a journal, gathering acolytes, and running a directed research program, all of which Pearson did. Nor did Yule find Pearson's biometric (and eugenic) program appealing. Yule had diverse interests and sought to develop and apply statistical theory to problems as they came to hand, rather than to be more narrowly focused. Indeed, he once described the academic scientist's role as a "loafer in the world," detached and free from narrow, applied ties. In Yule's case, this was a very fruitful process. Although Yule was an individual researcher, he was also a keen contributor to scientific societies. Most notable was his participation in the Royal Statistical Society in London, which he joined in 1895. This society had a long tradition of using numerical data and quantitative methods to examine social problems, and it brought Yule into contact with civil servants and social scientists. The society's journal became the major outlet for Yule's scientific papers.

Yule made significant contributions in several different areas. The three arenas in which his contributions "made a difference" in the development of quantitative social research were multiple regression modeling, the study of contingency tables, and time-series analysis. The three themes correspond roughly to three periods in his research career: multiple regression, in the last decade of Victorian England; contingency tables, in the first decade of the new century; and time-series analysis, during the 1920s. In each arena, Yule made progress by recognizing the need to relax one or more of the assumptions that underlay the biometric approach to statistical inference, and to construct more appropriate probability tools. (Pearson's seminal work had been closely based on assumptions of multinormal frequency distributions together with independence among interval-scale observations; such assumptions were reasonably appropriate for his applications, i.e., measuring physical characteristics of plants, animals, or people, but they did not work for many socioeconomic applications or for other nonexperimental situations, such as time-series observations of physical phenomena.) Yule's major papers on contingency tables and time-series analysis have been reprinted, but his early regression papers have not.

Yule's intellectual contributions were recognized at the time of his work; through his publications and his direct influence in bodies such as the Royal Statistical Society and the International Statistical Institute, his contributions diffused throughout the world of early quantitative social science in the decades from 1910 to 1940. Though some of the methods have been overtaken by later work, some of his work still has a fresh, "modern" feel to it, and is still relevant.

Multiple Regression

Historians of statistics agree that Yule played a very significant role in the development of statistical methods in the 1890s. John Aldrich writes that "Karl Pearson and G. Udny Yule developed the main interpretations of correlation used by statisticians for the last century or so" and "Between them they made correlation analysis." For Stephen M. Stigler, it is Yule who puts the final keystone, the reintegration of correlation and regression into least-squares theory, into place: "Only one important step remained to be taken in 1895, a step crucial to the completion of the larger program that Galton had launched. Before two years had passed G. Udny Yule was to take that step."

The empirical context for Yule's statistical applications was English Poor Law policy. The 19th-century Poor Law was a tough policy dealing with those unable to support themselves through illness, unemployment, or old age: the workhouse or, for those unable to work, the closed asylum awaited these unfortunate Englishmen. This fate was what was called "in-relief." By the latter 19th century, considerable assistance was being provided by some Poor Law Unions in the form of "out-relief," which was assistance in the home, mainly for the elderly and ill, and there was a very active political debate on "pauperism," as poverty was then termed, a debate with echoes in recent

welfare policies. One strand argued that out-relief encouraged pauperism, whereas others argued there was no relationship. In a paper in the *Economic Journal*, Yule employed data for over 500 spatial units (Poor Law Unions) in England in 1871 and 1891 to test the relationship, using the (brand new) correlation coefficients: for 1871, the correlation between out-relief and pauperism was 0.262 (probable error, 0.025), and for 1891, the correlation was 0.388 (error, 0.022).

Yule did not argue that there was a causal relationship, only an association, and, in response to criticisms from Charles Booth, a wealthy businessman-*cum*-sociologist and devotee of Auguste Comte, Yule recognized the need to take account of other factors, giving him the perfect opportunity to deploy his new tools of multiple and partial correlation. In 1896, Yule wrote that "the proper method to be employed is, it seems to me, that of 'multiple correlation.' This method enables us to deal with facility with three variables, and if need be with more, and to form coefficients of correlation between any two of the variables while eliminating the effect of variations in the third (or others). Such 'net [partial] coefficients' will probably play an important part in future statistical researches." Here his work introduced the concept of controlling for the effects of other influential variables or effects when assessing the association between two variables, a concept that has proved vital to statistical analysis in general and to the nonexperimental social sciences in particular. If such effects were not taken into account, there was a real risk of spurious correlations (or "illusory correlations," as Yule termed them) appearing. Yule began his 1897 paper *On the Theory of Correlation* by stating that "The investigation of causal relations between economic phenomena presents many problems of peculiar difficulty, and offers many opportunities for fallacious conclusions. Since the statistician can seldom or never make experiments for himself, he has to accept the data of daily experience, and discuss as best he can the relation of a whole group of changes; he cannot, like the physicist, narrow down the issue to the effect of one variation at a time. The problems of statistics are in this sense far more complex than the problems of physics."

In tackling the issues of multivariate relationships, Yule broke away from Pearson's assumption of multivariate normality. Instead, Yule framed his analysis within the method of least squares, long established in astronomy and elsewhere as a theory of errors. In *The History of Statistics: The Measurement of Uncertainty before 1900*, Stigler captured the importance of Yule's move: "The paper had a new and broader outlook that at once put the developing theory of correlation in a perspective from which it could deal with the problems of the social sciences and reconciled it formally with the traditional method of least squares from the theory of errors. As such Yule's work marked the completion of a final stage in the development of what could be called Galton's program and formed a cap on nineteenth-century work on statistics for the social sciences."

A further paper in 1899 took Yule's work forward into an explicitly causal framework, presenting the multiple regression model and applying it to the social problem of pauperism. In the latter decades of Victorian England, the levels of pauperism had been falling markedly. Now Yule saw that the real test of explanation was whether changes in the explanatory factors could explain such changes in pauperism. The paper jumps straight into a multicausal approach: "The various causes that one may conceive to effect changes in the rate of pauperism may for clearness be classified under some five heads," which were (1) policy or administration; (2) economic conditions (wages, trade levels, employment), (3) social or industrial character (including density), (4) moral character (such as crime, education, illegitimacy), and (5) the age distribution. In practice, Yule only had data available to take some of these factors into account. For rural regions for the decade 1871–1881, Yule's multiple regression equation was as follows: percentage change in pauperism = $-27.07 + 0.299$ (percentage change in out-relief ratio) + 0.271 (percentage change in proportion old) + 0.064 (percentage change in population). Yule was then able to argue that the coefficient for the out-relief ratio ($+0.299$) "gives the change due to this factor when all the others are kept constant." The model thus assesses the role of the policy instrument, conditioning on what today would be called the other exogenous variables. Yule recognized that his analysis was still susceptible to the omitted factors that he was unable to measure: "there is still a certain chance of error depending on the number of factors correlated both with pauperism and with the proportion of out-relief which have been omitted, but obviously this chance of error will be much smaller than before." In hindsight, with the benefits of a century of subsequent statistical research, it is possible to identify further gaps and limitations in Yule's statistical analysis (such as outliers, spatial autocorrelation, and possible endogeneity of the measure of policy). But it is the remarkable quality and depth of the work that stands out. As Stigler commented, "the paper was in its way a masterpiece, a careful, full-scale, applied regression analysis of social science data."

Contingency Tables

The second arena in which Yule made a contribution to quantitative social research was the analysis of categorical variables and contingency tables. The statistical advances of the 1890s, including Yule's own work, had focused on interval or continuous-scale variables, but by the turn of the century, attention was also being given to attribute or

categorical variables. Yule gave the example of mortality from some disease with and without the administration of a new antitoxin: an individual died or lived, had the antitoxin or did not, but there were no continuous scales of variation involved. In other cases, as Yule noted, there might be gradation possible, but the actual data were categorical (blind or not blind, deaf or not deaf). For the two-category, two-variable case, a simple contingency table (as in Table I) can represent the data: thus a is the number of people who had the antitoxin and survived, c is those who had antitoxin but still died, and $a+c$ is the total number of people who had the antitoxin; b and d are defined similarly for the "no antitoxin" groups, and N $(=a+b+c+d)$ gives the total number of people involved. How were correlations or other measures of association to be constructed for such cases?

Yule and Pearson came up with different perspectives on this situation, leading to serious, and sometimes intense, controversy, described in the literature as "the politics of the contingency table." Pearson wanted to treat all such categorical data as simplifications from an underlying normal frequency distribution, so accommodating them within his biometric correlation frame. This is plausible for some cases, but not for others, such as Yule's mortality example. By contrast, Yule took the categories as given, and attempted to construct measures of association relevant to the actual case. A coefficient, calculated directly from the table, should be zero if the two attributes were independent of each other, +1 or −1 if, and only if, they were completely associated in a positive or negative sense. Yule's Q-coefficient, defined as $Q=(ad-bc)/(ad+bc)$, did this, but it had no special justification, and other measures, equally valid, could be constructed but did not always give identical inferences. The controversy between Yule and Pearson, and that between their allies, was seen as exemplary of a deeper division in terms of attitudes to the biometric program and eugenics. This perhaps overinterprets the differences, but certainly it reflected Yule's more pragmatic, flexible approach to inference and Pearson's allegiance to the model of normal frequency distribution.

Both perspectives had their supporters. Pearson's measures were adopted within the emerging field of psychometrics whereas Yule's measures became popular in sociology. The field of contingency table and categorical data analysis has moved on a long way since this early work, and overall has built on Yule's insights rather than on Pearson's more rigid assumptions.

Table I A Simple Two-by-Two Contingency Table

Outcome	*Antitoxin*	*No antitoxin*	*Total*
Survived	a	b	$a+b$
Died	c	d	$c+d$
Total	$a+c$	$b+d$	N

Time-Series Analysis

Yule's third major contribution was to the development of time-series analysis. By the first years of the 20th century, there was recognition that correlating two sets of time-series observations could give rise to apparently significant, but spurious, results. If the two series both followed similar trending or oscillatory paths, then, inevitably, they would appear correlated. Several ways around this were suggested, such as first-differencing the series or removing the trend by linear or polynomial regressions and then correlating the residuals. As David Hendry and Mary Morgan have noted, "This trial and error approach was not matched by any deep understanding of statistical theory." It also drew implicitly on the model of spurious correlation being generated by an omitted "third" variable, in this case, "time": partial out the time factor and the genuine correlation could be retrieved. Yule reviewed the various approaches in a paper in 1921, finding the field confused in aims and methods and demonstrating that random variations could give rise to the apparent correlations.

These issues occupied Yule during the 1920s, and he produced seminal papers rejecting the earlier perspective. Two papers in particular, in 1926 and 1927, were influential, and these were both reprinted in Hendry and Morgan's *The Foundations of Econometric Analysis*, in which they noted that, in contrast to previous students of this problem, Yule's work was "informed by an understanding of statistical theory which was amongst the best of his generation." Yule rejected the argument that the source of the problem was that the variables were correlated with time. Instead, he argued that the problem arose because the observations were correlated with previous values in the same series, and that the changes in the observations were also correlated with previous values of changes. In other words, it was the internal dynamics of the time series that generated what Yule termed "nonsense correlations." Yule's 1926 paper on nonsense correlations begins with an example: for England in the period 1866–1911, the proportion of Church of England marriages to all marriages and the standardized death rate have a correlation coefficient of +0.9512. Yule makes the following comments:

> *Now I suppose it is possible, given a little ingenuity and goodwill, to rationalize very nearly anything. And I imagine some enthusiast arguing that the fall in the proportion of Church of England marriages is simply due to the Spread of Scientific Thinking since 1866, and the fall in mortality is also clearly to be ascribed to the Progress of*

Science; hence both variables are largely or mainly influenced by a common factor and consequently ought to be highly correlated. . . . But most people would, I think, agree with me that the correlation is simply sheer nonsense; that it has no meaning whatsoever; that it is absurd to suppose that the two variables in question are in any sort of way, however indirect, causally related to one another.

Yule investigated correlations between two unconnected time series under three different assumptions: when the individual series are (a) random, (b) random in first differences (which Yule termed "conjunct," to denote series for which all serial correlations were positive), and (c) random in second differences. In modern econometric parlance, these are series integrated of orders 0, 1, and 2, respectively. Classical sampling theory for correlation is valid for the random case a, and Yule's simulations demonstrated this, but not for the other models. Model b generated a much more dispersed distribution (so that "no correlation" would commonly be rejected), and model c generated a U-shaped distribution of correlations, with most values close to either $+1$ or -1. Series of type c Yule labeled "dangerous." To produce these results, Yule engaged in substantial Monte Carlo simulation of the distributions. With the very limited calculator assistance available in the 1920s, the work must have taken Yule many weeks, locked away in his study in Cambridge. Hendry and Morgan were able to replicate his results closely, but using modern computing equipment.

In his paper of the following year, as part of an investigation of Alfred Wolfer's sunspot cycle series, Yule introduced the autoregressive model (though Yule did not used the this term), in second-order form: $y_t = a_1 y_{t-1} + a_2 y_{t-1} + e_t$, where y_t is the series, a_1 and a_2 are autoregressive parameters, and e_t is a random disturbance. He demonstrated how such autoregressive models could generate oscillatory sequences, and also gave the random component what has been called "an active role as a disturbance term," showing how the random disturbances played a central role in the dynamics. They were not simply fluctuations around the oscillations, but could alter the phase and amplitude. Yule neatly used the analogy of a pendulum in a room: "Unfortunately boys get into the room and start pelting the pendulum with peas, sometimes from one side and sometimes from the other. The motion is now affected, not by superimposed fluctuations, but by true disturbances, and the effect on the graph will be of an entirely different kind. The graph will remain surprisingly smooth, but the amplitude and phase will vary continually." Judy Klein, in *Statistical Visions in Time. A History of Time Series Analysis, 1662–1938*, stated that "This work became the foundation for modern time series analysis. It was the first complete formulation of what was later to be called a stationary stochastic process of the autoregressive type. It was also the first instance in which an error term (e) was used to signify random disturbance in a relationship, not just errors in measurement." Other statisticians, such as Ragnar Frisch, Helen Walker, and Herman Wold, elaborated on these ideas, which lie at the heart of both time-series analysis and econometrics. Yule never went beyond the model for a single series to propose a model for bivariate series, and he commented "It is quite beyond my abilities, but I hope that some mathematician will take it up." It can be argued that a full analysis had to await Clive Granger's development of cointegrated time series in the 1980s.

Legacy

Perhaps inevitably, as with other pioneers of quantitative social research, Yule's name appears today only rarely in contemporary textbooks, but his work has been receiving increasing recognition in histories of statistics and econometrics. It is especially the classic time-series papers that are reprinted. Yule's multiple regression work received shorter shrift: when Alan Stuart and Maurice Kendall came to assemble a selection of Yule's papers for a retrospective collection in 1971, they did not include any of these papers. Commenting on the problems of selection from Yule's wide range, they noted that "It was with reluctance that we omitted Yule's pioneering work on pauperism, which is still of methodological interest to sociologists." In case this is read as dismissing obscure work with backhanded praise, it is worth quoting Kendall from his 1951 obituary article on Yule: "the value of some of his contributions has been lost to view in sheer virtue of their success; for example, his work on correlation and regression is now such standard practice that only a student of history would consult the original papers." This is very true: the methods developed and applied by Yule were formative in demonstrating how statistical modeling could be relevant to the nonexperimental context of social and economic analysis. Many of his methods are still in use, and elsewhere his insights have been built up and extended. Yule's remarkable achievement was to make these significant contributions to three different arenas of statistical modeling.

See Also the Following Articles

Contingency Tables and Log-Linear Models • Correlations • Time-Series–Cross-Section Data

Further Reading

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statist. Sci.* **10,** 364–376.

Hendry, D. F., and Morgan, M. S. (eds.) (1995). *The Foundations of Econometric Analysis*. Cambridge University Press, Cambridge.

Hepple, L. W. (2001). Multiple regression and spatial policy analysis: George Udny Yule and the origins of statistical social science. *Environ. Plan. D: Soc. Space* **19,** 385–408.

Klein, J. L. (1997). *Statistical Visions in Time. A History of Time Series Analysis, 1662–1938.* Cambridge University Press, Cambridge.

MacKenzie, D. (1981). *Statistics in Britain 1865–1930. The Social Construction of Scientific Knowledge.* Edinburgh University Press, Edinburgh.

Morgan, M. S. (1990). *The History of Econometric Ideas.* Cambridge University Press, Cambridge.

Porter, T. M. (1986). *The Rise of Statistical Thinking, 1820–1900.* Princeton University Press, Princeton, NJ.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard University Press, Cambridge, MA.

Stuart, A., and Kendall, M. G. (eds.) (1971). *The Statistical Papers of George Udny Yule*. Griffin, London.

Yule, G. U. (1895). On the correlation of total pauperism with proportion of out-relief, I: All ages. *Econ. J.* **5,** 603–611.

Yule, G. U. (1896). On the correlation of total pauperism with proportion of out-relief, II: Males over sixty-five. *Econ. J.* **6,** 613–623.

Yule, G. U. (1897). On the theory of correlation. *J. Roy. Statist. Soc.* **60,** 812–854.

Yule, G. U. (1899). An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades, I. *J. Roy. Statist. Soc.* **62,** 249–295.

Yule, G. U. (1911). *An Introduction to the Theory of Statistics,* 1st Ed. Griffin, London.

List of Reviewers

Marien Abreu

John Adamopoulos
Grand Valley State University

Marvin Alkin
University of California, Los Angeles

Gary Allen
University of South Carolina

Carl Amrhein
University of Toronto

Ted Anagnoson
California State University, Los Angeles

David Andrich
Murdoch University

Bernard Andrieu
Université Nancy

Joshua Angrist
Massachusetts Institute of Technology

David Armor
George Mason University

Richard Ashcroft
Imperial College of Science, Technology and Medicine

Erik Austin
Inter-University Consortium for Political and Social Research

Yasumasa Baba
Institute of Statistical Mathematics

Earl Babbie
Chapmann University

Sharman L. Babior
University of California, Los Angeles

Valentina Bali
Michigan State University

David Banks
Duke University

Robert Bannister
Swarthmore College

Azy Barak
University of Haifa

Ralph Bargmann
University of Georgia

Stephen Baron
Queen's University

Nathaniel Beck
University of California, San Diego

Benjamin Beit-Hallahmi
University of Haifa

Clive R. Bellfield
Columbia University

David Bellhouse
University of Western Ontario, London, Canada

Sara Benesh
University of Wisconsin, Milwaukee

J. Henry Bennett
University of Adelaide

Rebecca Bennett
University of Manchester

Peter M. Bentler
University of California, Los Angeles

H. Russell Bernard
University of Florida

Frank J. Bernieri
University of Toledo

Kurt Beron
University of Texas at Dallas

Brian Berry
University of Texas at Dallas

Peter Blanck
University of Iowa

Glenn Blomquist
University of Kentucky

Robert Bogdan
The Maxwell School of Syracuse University

George Bohrnstedt
Stanford University

Kenneth Bollen
University of North Carolina, Chapel Hill

Roger E. Bolton
Williams College

Noel Bonneuil
Institut National de Études Démographiques

Dorret Boomsma
Free University of Amsterdam

Walter Borman
University of South Florida

Mike Bourne
Cranfield University

Peter Bowler
Queen's University

Janet Box-Steffensmeier
Ohio State University

Robert Brame
University of South Carolina

Michael Brannick
University of South Florida

John Brehm
University of Chicago

Robert Brennan
University of Iowa

Devon Brewer
University of Washington

Ron Briggs
University of Texas at Dallas

Paul Brodwin
University of Wisconsin

Charles Brody
University of North Carolina

Steven R. Brown
Kent State University

Alan Brown
Southern Methodist in Dallas

Harry Bruce
University of Washington

Gordon Bruner
Southern Illinois University

Stanely Brunn
University of Kentucky

Alan Bryman
Loughborough University

Jack Buckley
SUNY Stony Brook

David Burbridge

Peter Burke
Washington State University

James E. Carlson
National Assessment Governing Board

Emma Cave
University of Leeds

Ruth Chadwick
Lancaster University

Ray Chambers
University of Southampton

Richard Church
University of California at Santa Barbara

Paul Cilliers
University of Stellenbosch

Robert Clark
Australian Bureau of Statistics

Matthew Clarke
RMIT University

W. David Clinton III
Tulane University

Randall Collins
University of Pennsylvania

Paul Coomes
University of Louisville

Horacio Levy Copello
Universitat Autònoma de Barcelona

Jeannine Coreil
University of South Florida

Barbara Costello
University of Rhode Island

Brad Cousins
University of Ottawa

Tom Cova
University of Utah

Gary Cox
University of California, San Diego

William Craig
University of Minnesota

Chet Creider
University of Western Ontario, New London

James E. Crimmins
University of Western Ontario

John Cromartie
Economic Research Service, USDA

Paul Cromwell
Wichita State University

Blaise Cronin
Indiana University

James F. Crow
University of Wisconsin

Dennis Culhane
University of Pennsylvania

Theodore Curry
University of Texas at El Paso

Roy D'Andrade
University of California, San Diego

Lorraine Daston
Max-Planck-Institut für Wissenschaftsgeschichte

Gary David
Bentley College

Howard Davis
Warwick Business School

John B. Davis
Marquette University

Paul Davis
Aston Business School

Robyn Dawes
Carnegie Mellon

Myrna Dawson
Guelph University

Suzanna De Boef
Pennsylvania State University

Thomas de Graaff
Free University, Amsterdam

Ton de Jong
University of Twente

Victor de Munck
SUNY New Paltz

Scott Decker
University of Missouri, St. Louis

Alfred DeMaris
Bowling Green State University

Maarten Derksen
University of Groningen

Craig Deville
Iowa Tests of Basic Skills

Ilia Dichev
University of Michigan

Paul Diehl
University of Illinois

Erwin Diewert
University of British Columbia

Don Dillman
Washington State University

JoAnn Dionne
University of Michigan

Fritz Drasgow
University of Illinois at Urbana-Champaign

Pieter J. D. Drenth
Free University of Amsterdam

William Dressler
University of Alabama

Sean Duffy
University of Michigan

Thomas G. Dukich
Advisor to Industry, Government, Education

George Duncan
Carnegie Mellon University

Alice Eagly
Northwestern University

A. W. F. Edwards
Cambridge University

Robert Eisinger
Lewis and Clark

Gloria D. Eldridge
University of Alaska

Dov Elizur
Bar Ilan University

Joe Elliott
University of Sunderland

Phoebe Ellsworth
University of Michigan

Jan Elshout
University of Amsterdam

Richard Ely
Boston University

Robert Emerson
University of California, Los Angeles

Paula England
Stanford University

Cem Ertur
Université de Bourgogne

Ken J. Euske
US Navy

Raymond Fancher
York University

George Farkas
The Pennsylvania State University

David Farrington
Cambridge University

Katherine Faust
University of California, Irvine

Lynette Feder
Portland State University

Nisha Fernando
University of Wisconsin-Stevens Point

Nigel Fielding
University of Surrey

Charles Finocchiaro
Michigan State University

Glenn Firebaugh
Pennsylvania State University

Michael Fischer
University of Kent

Patricia Diamond Fletcher
University of Maryland, Baltimore County

Marvella Ford
VA Medical Center—Houston

Steven Forde
University of North Texas

G. J. A. Fox
Twente University

Frank Francois
Former head of AASHTO

Robert J. Franzese
University of Michigan

Doug Frechtling
George Washington University

David A. Freedman
University of California, Berkeley

Linton C. Freeman
University of California, Irvine

Kim Fridkin
Arizona State University

Michael Friendly
York University

Michelle L. Frisco
Iowa State University

Lance D. Fusarelli
Fordham University

Sam Gaertner
University of Delaware

Michael Gaffikin
University of Wollongong

Norman Gall
University of Calgary

Xiaohong Gao
ACT Inc.

Tommy Garling
Göteborg University

Gerald Gates
US Census Bureau

Andrew Gelman
Columbia University

Richard Gelpke
University of Massachusetts at Boston

Arthur Getis
San Diego State University

Gerd Gigerenzer
Max-Planck-Institute für Bildungsforschung

Nigel Gilbert
University of Surrey

Jean Giles-Sims
Texas Christian University

Jeff Gill
University of Florida

David Gillespie
Washington University

Mike Gillespie
University of Alberta

Garret Glasgow
University of California, Santa Barbara

Jean Berko Gleason
Boston University

Jack Goldstone
University of California, Davis

Hongmian Gong
Hunter College, CUNY

Michael Goodchild
University of California, Santa Barbara

Linda Gottfredson
University of Delaware

Ivor Grattan-Guiness
Middlesex University at Enfield

William W. Graves, IV
University of North Carolina at Charlotte

Ann Green
Yale University

Christopher Green
York University

F. Green
University of Kent at Canterbury

Michael Greenacre
University of Pompeu Fabra

William Greene
New York University

Lee-Anne Greer
East Prince Mental Health

Henrich Greve
Norwegian School of Management BI

Daniel Griffith
Syracuse University

Eve Gruntfest
University of Colorado

Feng Gu
Boston University

Peter Guarnaccia
Rutgers University

Guang Guo
University of North Carolina

Peter Guth
US Naval Academy

James Guthrie
Vanderbilt University

Darrene Hackler
George Mason University

Donald Haider-Markel
University of Kansas

Stephen Hall
Imperial College, University of London

Ronald Hambleton
University of Massachusetts

David Hand
Imperial College

John Hand
University of North Carolina at Chapel Hill

Hans-Tore Hansen
University of Bergen

Brad Hanson
CTB/McGraw-Hill

Sandra Lee Harding
Queensland University of Technology

Wynne Harlen
University of Bristol

James W. Harrington
University of Washington

Richard Harris
University of Bristol

Harm't Hart
University of Utrecht

David Harvey
University of Nevada, Reno

Philip Haynes
University of Brighton

David Healy
University of Wales College of Medicine

Don Hedeker
University of Illinois at Chicago

Willem Heiser
Universiteit Leiden

Paul Herrnson
University of Maryland

Miguel Hernan
Harvard School of Public Health

Frederick Hess
University of Virginia

Ben Highton
University of California, Davis

Travis Hirschi
University of Arizona

Dick Hobbs
University of Durham

Ken Hodges
Claritas

John Hodgson
University of Alberta

Nancy Hoffart
Northeastern University

Howard Hogan
Census Bureau

Herbert Hoijtink
University of Utrecht

Darryl J. Holman
University of Washington, Seattle

Robert D. Hoppa
University of Manitoba

Ruth Horowitz
New York University

Ann Howard
University of North Carolina at Chapel Hill

Youqin Huang

Janice Huber
St. Francis Xavier University

Martina Huemann

John Hughes
Lancaster University

Hiddo A. Huitzing
University of Groningen

Gary Hunt
University of Maine at Orono

Mike Hutchinson
Australian National University

Paul Huth
University of Michigan

John Iceland
University of Maryland

Simon Jackman
Stanford University

Bruce A. Jacobs
University of Texas

Paul T. Jaeger
Florida State University

Craig R. Janes
University of Colorado at Denver

Maria Anna Jankowska
University of Idaho

Robin Jarrett
University of Illinois, Urbana

Eric Jensen
University of Idaho

Steve Jex
University of Wisconsin at Oshkosh

Peter Johnstone
Invermay Agricultural Centre

Dean H. Judson
US Census Bureau

William Kalsbeek
Iowa State University

Alan Karr
National Institute of Statistical Sciences

Elihu Katz
University of Pennsylvania

Kimberly Kempf-Leonard
University of Texas at Dallas

David Kenny
University of Connecticut

Shahjahan Khan
University of Southern Queensland

Doug Kiel
University of Texas at Dallas

Philip Kilbride
Bryn Mawr College

Gary King
Harvard University

Jean A. King
University of Minnesota

Maryon F. King
Southern Illinois University

Ann Marie Kinnell
University of Southern Mississippi

Rob Kitchin
National University of Ireland

Kim Kleinman
Missouri Botanical Garden

John Knodel
University of Michigan

Martin Knott
London School of Economics

Thorbjørn Knudsen
University of Southern Denmark

Timothy Kohler
Washington State University

Peter Koltnow
Former head of American Highway Users Alliance

Lyle Konigsberg
University of Tennessee, Knoxville

Aryeh Kosman
Haverford College

J. Morgan Kousser
California Institute of Technology

Herbert Kritzer
University of Wisconsin

David Kronenfeld
University of California, Riverside

Richard Kulka
Research Triangle Institute

Mei-Po Kwan
The Ohio State University

Seppo Laaksonen
Statistics Finland

Peter A. Lachenbruch
Division of Biostatistics/OBE/CBER zHFM-215

Partha Lahiri
University of Maryland

Johanna Laiho
Statistics Finland

Sanford Lakoff
University of California, San Diego

Rajav Lal
Harvard Business School

Kenneth Land
Duke University

Edward L. Lascher
California State University Sacramento

Diane Leach
National Institute of Child Health and Human Development

Amy E. Learmonth
Rutgers University

Michel Lebas
HEC School of Management, Paris

Ned Lebow
Dartmouth

Ann Ledwith
University of Limerick

Robert D. Lee
Pennsylvania State University

F. K. Lehman
University of Illinois

James Lepkowski
University of Michigan

Susan C. Levine
University of Chicago

Michael Lewis-Beck
University of Iowa

Loet Leydesdorff
Amsterdam School of Communications Research

Rebecca Li
College of New Jersey

Donald Lien
University of Texas at San Antonio

J. Robert Lilly
Northern Kentucky University

Mike Linacre
University of Sunshine Coast

John Loehlin
University of Texas at Austin

Enrique Lopez-Bazo
University of Barcelona

J. Dennis Lord
University of North Carolina at Charlotte

Johann Louw
University of Cape Town

Thomas R. Loveland
EROS USGS Data Center

Patricia Lovie
Keele University

Karen Luftey
University of Minnesota

Steven Lukes
New York University

Guanzhong Luo
Murdoch University

Armando Machado
Universidade do Minho

Colleen MacQuarrie
University of Prince Edward Island

M. Eileen Magnello
Wellcome Trust Centre at UCL

Michael Malbin
SUNY-Albany in Washington

Mary Malina
Naval Postgraduate School

George Marcoulides
California State University—Fullerton

Melvin Mark
Pennsylvania State University

Tony Marley
McGill University

Bernard Marr
Cranfield University

Beth Marsh
Duke University

Lawrence Marsh
Notre Dame University

Jim Marston
University of California, Santa Barbara

Luis Martins
Georgia Institute of Technology

Geoffrey Maruyama
University of Minnesota

Richard O. Mason
Southern Methodist University

Robert Mason
Oregon State University

Christina Mauleon
Chalmers University of Technology

Scott E. Maxwell
University of Notre Dame

Lorraine Mazerolle
Griffith University

George McCall
University of Missouri, St. Louis

Charles R. McClure
Florida State University

James H. McDonald
University of Texas at San Antonio

Doris McGartland-Rubio
University of Pittsburgh

W. Scott McGraw
Ohio State University

John McIver
University of Colorado at Boulder

Linda McKie
Glasgow Caledonian University

Craig McLaren
Australian Bureau of Statistics

Scott Meis
Canadian Tourism Commission

Jerry Melican
American Institute of Certified Public Accountants

M. David Merrill
Utah State University

Jane Meza
University of Nebraska Medical Center

Manus Midlarsky
Rutgers University

Harvey J. Miller
University of Utah

Jay Miller
USDA Forest Service Remote Sensing Laboratory

Tim Miller
ACT, Inc.

Michael Mintrom
University of Auckland

Stephen Mockabee
University of Cincinnati

Linda Molm
University of Arizona

Maria Montero
University of Nottingham, United Kingdom

Susan Moreno
Houston Independent School District

Jane Morrice
The Macaulay Institute

Peter Morrison
Rand Corporation

Betty Morrow
Florida International University

Joseph R. Moskal
Northwestern University and Nyxis Neurotherapies

Patricia Moy
University of Washington at Seattle

Robert G. Mrtek
University of Illinois, Chicago

Stanley A. Mulaik
Georgia Institute of Technology

Wendy Myrvold
University of Victoria

Joan Naymark
Target Corporation

Andy Neely
Cranfield University

Marc Nerlove
University of Maryland

Bill Nichols

Roy A. Nielsen
University of Oslo

Bryan Norton
Georgia Institute of Technology

João Arriscado Nunes
University of Coimbra

Edward J. O'Boyle
Louisiana Technical University

Robert O'Brien
University of Oregon

Barbara Ohlund
University of Nebraska-Lincoln

Morton O'Kelly
Ohio State University

J. Keith Ord
Georgetown University

Bruce Owens
Wheaton College

R. Kelley Pace
Louisiana State University

Kavita Pandit
University of Georgia

Cynthia Parshall
University of South Florida

Diane Paul
University of Massachusetts

Lynn Paxson
Iowa State University

Joseph Pear
University of Manitoba

Leonard I. Pearlin
University of Maryland

Eric Pels
Free University, Amsterdam

Andrew Penner
University of California, Berkeley

Charles Perrings
University of York

Richard Perry
San Jose State University

David Peterson
Texas A&M University

Michael P. Peterson
University of Nebraska at Omaha

Trond Peterson
University of California, Berkeley

E. Pitacco
Università degli Studi di Trieste

Barbara Plake
University of Nebraska

David Plane
University of Arizona

Jennifer Platt
University of Sussex

Leonard Plotnicov
University of Pittsburgh

Andrew Plunkett
Manchester Metro University

John Polk
University of Illinois, Urbana-Champaign

Bruce Pollack-Johnson
Villanova University

Henry Pollakowski
Massachusetts Institute of Technology

Timothy Pollock
University of Maryland

Gerald M. Pomper
Rutgers University

Paula Popovich
Ohio University

Ted Porter
University of California, Los Angeles

Emil Posavac
Loyola University

Stanley Presser
University of Maryland

Kenneth Prewitt
Columbia University

Kevin Quinn
University of Washington

Jeff Rachlinski
Cornell Law School

Charles Ragin
Arizona State University

Senta Raizen
National Center for Improving Science Education

Uday Rajan
Carnegie Mellon

Suparna Rajaram
SUNY-Stony Brook

Nambury Raju
Illinois Institute of Technology

James Ramsay
McGill University

Michael R. Rand
US Department of Justice

T. J. Rao
Indian Statistical Institute

Doug Raybeck
Hamilton College

John Rayner
University of Woollongong

Dwight Read
University of California, Los Angeles

Sean Reardon
Pennsylvania State

Michael Regenwetter
University of Illinois, Urbana-Champaign

Charles Reichardt
University of Denver

Charles Reigeluth
Indiana University

Callie Rennison
US Department of Justice

Claus Rerup
University of Western Ontario

Dan S. Rickman
Oklahoma State University

Nicolas Rieucau

David Rigby
University of California, Los Angeles

Jonathan Rodden
Massachusetts Institute of Technology

Jon Rogstad
Institute for Social Research, Oslo

Thomas A. Romberg
University of Wisconsin, Madison

Karl S. Rosengren
University of Illinois at Urbana-Champaign

Peter Rossel
University of Copenhagen

Kenneth Rothman
Boston University

Ronald Rousseau
KHBO, Industrial Sciences and Technology

Mark Runco
University of Hawaii

Frank Rusciano
Rider University

John Ruscio
Elizabethtown College

Andrea Rusnock
University of Rhode Island

Karla Davis Salazar
University of South Florida

Joseph Salvo
New York City Planning Department

Franz Samelson
Kansas State University

Victor Sampedro
University Rey Juan Carlos, Madrid

Clint Sanders
University of Connecticut

Alan R. Sandstrom
Indiana University—Purdue University Fort Wayne

Kathryn G. Sapnas
University of Miami

Steven Sawyer
School of Information Sciences and Technology, Penn State

Andrew Sayer
Lancaster University

Walter Schaeken
University of Leuven

Susan Schechter
US Office of Management and Budget

Matthew Scheider
National Institute of Justice

Fritz Scheuren
National Opinion Research Center

Renato Schibeci
Murdoch University

Burkhard Schipper
University of Bonn, Germany

Warren Schmaus
Illinois Institute of Technology

Lynda Schneekloth
SUNY/Buffalo

Vic Schoenbach
University of North Carolina at Chapel Hill

Edward Schortman
Kenyon College

Howard Schuman
University of Michigan

Jennifer Schwarz
Washington State University

Norbert Schwarz
University of Michigan

Ruth Meyer Schweizer
University of Berne

Thomas Scruggs
George Mason University

Steven Selden
University of Maryland

Daniel Serra
Universitat Pompeu Fabra

Richard Shaffer
Cal Poly State University

Mordechai Shani
Gertner Institute for Epidemiology and Health Policy Research—Israel

Robert Shapiro
Columbia University

Robert Sherman
California Institute of Technology, Pasadena

Thomas H. Short
Indiana University

Steven Shumate
Texas Tech University

Judith Shuval
Hadassa School of Public Health

Jacob S. Siegel
J. Stuart Siegel Demographic Services

Peter Simonson
University of Pittsburgh

Eleanor Singer
University of Michigan

Gideon Sjoberg
University of Texas at Austin

Chris Skinner
University of Southampton

John Skvoretz
University of South Carolina

Martyna Sliwa
Newcastle Business School

Matthew Smallman-Raynor
Nottingham University

Corwin Smidt
Calvin College

Carol Smith
University of Kansas

Dwayne Smith
University of South Florida

Michael Smithson
Australian National University

Paul E. Sniderman
Stanford University

Michael Sokal
Worcester Polytechnic Institute

Mark Souva
Florida State University

J. Spreeuw
City University—London

John Staddon
Duke University

Charles Stagnor
University of Maryland

Allan C. Stam
Dartmouth College

Brian Steensland
Indiana University

Nicholas Steneck
University of Michigan

Walter Stephan
New Mexico State University

Stephen Stigler
University of Chicago

Reinoud Stoel
Free University Amsterdam

Calvin Streeter
University of Texas at Austin

Robert A. Strong
Washington and Lee University

Irene Styles
Murdoch University

Robert W. Sussman
Washington University

R. B. Sutcliffe
University of Bilbao

Simon Szreter
University of Cambridge

Yoshio Takane
McGill University

Toon Taris
University of Nijmegen

Christopher Taylor
University of Alabama, Birmingham

John Taylor
Florida State University

Yves Thibaudeau
US Census Bureau

Robert Thomas
Editor, Philosophia Mathematica

Joel Thompson
Appalachian State University

John Tiefenbacher
Southwest Texas State University

Stef Tijs
University of Tilburg

Rowland Tinline
Queen's University

Caroline Tolbert
Kent State

Charles W. Tolman
University of Victoria

Evangelos Triantaphyllou
Louisiana State University

David Tufte
Southern Utah University

Michael Turner
University of North Carolina, Charlotte

Stephen Turner
University of South Florida

Ryan D. Tweney
Bowling Green State University

Tom Ulen
University of Illinois—Law

Fusun Ulengin
Istanbul Technical University

Dawn M. Upchurch
University of California, Los Angeles

Arnold B. Urken
Stevens Institute of Technology

Steinar Vagstad
University of Bergen

Fons van de Vijver
Tilburg University

Andre Van Dierendonck
University of Ghent

Hedderik van Rijn
Carnegie Mellon University

Bernard P. Veldkamp
University of Twente

Geert Verbeke
University of Leuven

Robert Vernon
School of Social Work at IUPUI

J. Miguel Villas-Boas
University of California Berkeley

Penny S. Visser
University of Chicago

Alexander von Eye
Michigan State University

Daniel Vorgrimler
Universität Hohenheim

Edward J. Vytlacil
Stanford University

Ted Wachs
Purdue University

Ken Wachter
University of California, Berkeley

Brigitte S. Waldorf
University of Arizona

Robert Wallace
McMurry University

Patrick Walsh
University of Manitoba

Roberta Walsh
Florida Gulf Coast University

Tom Warke
Cambridge University

Richard Webber
University College London

Lori Weber
California State University, Chico

Chris Webster
Cardiff University

Stephan Weiler
Colorado State University

David Weiss
University of Minnesota

Gregory L. Weiss
Roanoke College

Mitchell Weiss
University of Basel, Switzerland

Sheila Weiss
Clarkson University

Eben Weitzman
University of Massachusetts, Boston

Susan C. Weller
University of Texas Medical Branch

Asoka M. Wettasinghe
Department of Infrastructure, Government of Victoria, Melbourne, Australia

Mark Wheeler
Western Michigan University

Nancy White
Bucknell University

Paul H. White
University of Utah

Alan Wilcox
University of Nevada

Pascal Wilhelm
University of Twente

Cecil Williams
University of Minho

Rick K. Wilson
Rice University

Jochen Wirtz
National University of Singapore

David Wong
George Mason University

Ken Wong
Vanderbilt University

B. Dan Wood
Texas A&M University

Robert F. Woolson
Medical University of South Carolina

Gerald C. Wright
Indiana University

James Wright
University of Central Florida

Dominic Wring
Loughborough University

Alice M. Wyrwicz
Northwestern University Feinberg School of Medicine

Malcah Yaeger-Dror
University of Arizona

Claire Yang
Duke University

Jingzhen Yang
University of North Carolina at Chapel Hill

Serdar Yilmaz
World Bank

John A. Young
Oregon State University

Alan Zaslavsky
Harvard Medical School

Tom Zelenock
Inter-University Consortium for Political and Social Research

April Zenisky
University of Massachusetts

Gary Zerbe
University of Colorado

Joe Zhu
Worcester Polytechnic Institute

Christian Zlolniski
University of Texas, Arlington

Christopher Zorn
Emory University