Robert B. Smith

Multilevel Modeling of Social Problems

A Causal Perspective

Springer

Multilevel Modeling of Social Problems

Robert B. Smith

Multilevel Modeling of Social Problems

A Causal Perspective

Springer

Robert B. Smith
Social Structural Research Inc.
3 Newport Road
Cambridge, MA
02140 USA
rsmithphd@comcast.net

Please note that additional material for this book (i.e., companion data sets and syntaxes for the analyses) can be downloaded from http://extras.springer.com

ISBN 978-90-481-9854-2 e-ISBN 978-90-481-9855-9
DOI 10.1007/978-90-481-9855-9
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010943038

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To grandchildren Adam, Jay, Eli,
Isaac, and Dev

Preface

Contemporary societal problems are complex, intractable, and costly. Aiming to ameliorate them, social scientists formulate policies and programs; then, to test the efficacy of the planned interventions, they develop study designs and conduct policy research. All too often the results are disappointing because the theories guiding the policies and programs are inappropriate and the study designs are flawed; moreover the empirical databases for answering the research questions are often sparse. This book confronts these difficulties by defining the following five-step process: Analyze the roots of the social problem both theoretically and empirically; formulate a study design that captures the nuances of the problem; gather empirical data providing valid and repeatable measures; model the multilevel data using appropriate multilevel statistical methods to uncover potential causes and any biases to their implied effects; and, finally, use the results to refine theory and to formulate evidence-based policy recommendations for implementation and further testing.

The core chapters apply this process to analyze the following societal problems I have studied: political extremism; global human development; violence against religious minorities; computerizations of work; reform of urban schools; health care utilization and costs; and parental reluctance to vaccinate children. These chapters address the multilevel data structures of the social problems by grouping observations on microunit (level-1) by more macrounits (level-2) (e.g., professors are grouped by their university), and by presenting multilevel statistical modeling in contextual, longitudinal, and meta-analyses. These chapters apply qualitative typologies that may explain the differences between the macrounits, thereby crafting a "mixed-methods" approach that combines qualitative attributes with quantitative measures.

Rather than beginning with a novel statistical model bearing on statistical theory and searching for illustrative data, each core chapter begins with a pressing societal problem. Exemplifying the usefulness of multilevel modeling for the quantifications of effects and for causal inference, the chapters serve as vivid exemplars for teaching students. This use of examples reverses the usual procedure for introducing statistical methods. The specific substantive problem motivates theoretical analysis, gathering of relevant data, and application of appropriate statistical procedures. Readers can use the supplementary data sets and syntaxes to replicate, critique, and advance the analyses, thereby developing their independent ability to produce applications of multilevel modeling. These data sets and syntaxes are available for downloading from http://extras.springer.com/

Multilevel models can capture the contextual and longitudinal aspects of social problems and provide evidence for causal inferences. But researchers working on social problems seldom develop multilevel models, even though the problems and the empirical data require such analyses. Facilitating the use of these statistical methods, this book provides examples of hierarchical data structures for which multilevel models are appropriate; namely, colleges and their professors, global regions and their constituent countries, countries and their citizens, organizations and their employees, schools and their students, and hospitals and their patients.

Software such as SAS's Proc Mixed and Proc Glimmix, MLwiN, M+, HLM, GLLAMM, SPSS Mixed Models, Stata, and BUGS have enabled researchers to conduct analyses of multilevel data sets after having mastered the logic of the method and the syntax of the computer programs. But students and researchers who desire to learn and apply these techniques are confronted with technical books that stress the statistical theory that may be too abstract for the general reader. Furthermore, these texts present examples mostly drawn from fields other than the social and behavioral sciences. Recent developments in the assessment of causal relationships in observational studies have stimulated discussions and research studies that differentiate the analysis of correlations from the analysis of causes. This literature can be technical and difficult for the uninitiated to grasp.

Through its substantive and incremental approach, this book meets the need for a basic introduction to the logic of multilevel modeling of social problems, and the assessment of causal relationships produced by such models. Each core chapter tackles a problem bearing directly or indirectly on aspects of social and economic development; each example is drawn from research practice and illustrates new theoretical or methodological principles. The four parts of the book and their component chapters can be read in any order but they are best read in sequence. Each advances the knowledge gained from mastery of the material that earlier chapters present; a reading of this book from chapter to chapter will develop the reader's knowledge and intuition. By replicating the analyses of each chapter using the available data sets and syntax, the learner will be able to solve similar problems and then tackle advanced methods and applications.

This book strikes a working balance among vital substantive problems, statistical theory, and statistical practice. Although it places a heavier emphasis on research practice than on statistical theory, each chapter explicates the theory bearing on its statistical modeling. It complements other books that present the formal, mathematical, or theoretical aspects of these methods. These texts are geared more toward mathematical statisticians, applied statisticians, and students of statistics, whereas this book is directed toward the social sciences (sociology, econometrics, education, government, history) as well as public health (health promotion, policy, and management) for advanced undergraduates, graduate students, and teachers of research methods; professional researchers; and managers of policy research. This book will also interest statisticians because the chapters present substantive analyses using statistical methods, and data sets and computer syntaxes are available for use in replications and reanalyses. The motivated learner will master this material by reading the chapters, thinking about their content, and replicating and advancing the analyses.

Acknowledgements

I thank Philip Gibbs and Jill Tao of the SAS Institute for their patient tutoring about Proc Mixed and Proc Glimmix; Russell Wolfinger, also of SAS, for suggesting how to introduce typologies in multilevel models; and Greta Ljung, a colleague in the Boston Chapter of the American Statistical Association, for clarifying the fixed- and random-effects paradigm of meta-analysis. William M. Mason's and Stephen W. Raudenbush's short courses on hierarchical linear modeling introduced these methods to me; the Boston Chapter's numerous workshops and presentations on causality and multilevel modeling advanced my knowledge.

For their interest in this project and their apt suggestions, I thank Myriam Poort, Miranda Dijksman, Stefan van Dijl, Esther Otten, and Berendina Van Straalen at Springer; and for coordinating the production of this book, I thank Sunil Padman and his colleagues at SPi Technologies India. For their comments on various chapter drafts, I thank Kevin Bales, Richard Berk, Sarah Cook, Helen Fein, Bruce Goldberg, Lilian Handlin, Rezaul Khandker, Irina Koltoniuc, Alice Kasakoff, C. Deborah Laughton, Ruth Leeds Love, William A. Lucas, Henry May, Laird McCulloch, Elizabeth Newton, Vivek Pradhan, Misha Salganik, Robert Shelly, Roland Stark, and several anonymous reviewers. For their insightful comments on the complete manuscript, I thank Lawrence E. Hazelrigg and James W. Hardin. For her loving companionship that made this book possible, I especially thank Joanna Handlin Smith.

All of the chapters in this book are timely and contemporary: my access to Harvard University's Widener and Cabot libraries enabled me to advance my sociological and statistical knowledge; teaching a short course at the Massachusetts Institute of Technology helped me clarify the exposition. The chapters in their present form have not been published previously. However, I did derive aspects of some chapters from my prior published work: Chapter 2 is drawn from my explication of *The Academic Mind* that appeared in *Applications of Mathematics in Models, Artificial Neural Networks and Arts*, edited by V. Capecchi et al. (© Springer 2010, 163–193). Portions of Chapter 3, on James Coleman's causal model, appeared in my article on inferential causal models (*Quality & Quantity*, vol. 37, issue 4, 337–361; © Kluwer 2003) and in my Introduction to Coleman's

Mathematics of Collective Action (©Aldine Transaction 2006, *vii-li*). Some of the modeling of electoral voting in Chapter 4 was published online in *Case Studies in Business, Industry, and Government Statistics: An Internet Journal* (vol. 2, no. 2, 127–146; © Robert B. Smith 2009). Key parts of Chapter 7, on global human development, appeared in *Quality & Quantity* (vol. 43, issue 1, 1–34, © Springer 2009), as did key parts of Chapter 8 on contemporary violence against European Jewish people (vol. 42, issue 2, 135–180; 2008, © Springer 2006). Major portions of Chapter 9, on how an automated fraud detector influenced employee moral, appeared in the *Evaluation Review* (vol. 26, no. 1, 3–39; © Sage Publications 2002), as did major portions of Chapter 14, on meta-analytic consolidations of findings about health care utilization and costs (vol. 25, no. 3, 288–330; © Sage Publications 2001). I thank these journals for allowing me to reuse my writings in these new chapters.

Cambridge, Massachusetts Robert B. Smith

Contents

List of Figures

List of Tables

Acronyms

ADL	Anti-Defamation League, a Jewish defense organization
AIC	Akaike Information Criteria
AMOS	Analysis of Moment Structures, a computer program.
ANOVA	Analysis of Variance
AR(1)	Auto Regressive covariance structure
ASD	Autism-spectrum disorders
BIC	Bayesian Information Criterion
BIC	Bayesian Information Criteria
BUGS	Bayesian inference Under Gibbs Sampling, a computer program
CATE	Average causal effect conditional on the covariates in the sample
CCT	controlled clinical trial
CDC	Centers for Disease Control and Prevention
CDSR	Cochrane Database of Systematic Reviews
CFRD	Computerized Fraud Detector
CS	compound symmetry covariance structure
DARE	Database of Abstracts of Reviews of Effects
DF or *df*	Degrees of Freedom
DID	Difference-in-Differences
DTaP	Diphtheria and tetanus toxoids and acellular pertussis (a vaccine)
DTP	diphtheria-tetanus and pertussis vaccine (a vaccine)
DTwP	diphtheria-tetanus and whole-cell pertussis (a vaccine) vaccine
EEs	Enrollees in an insurance plan
EMBASE	A biomedical and pharmacological bibliographic database
G20	Group of Twenty, see OECD.

GLLAMM	Generalized Linear Latent and Mixed Models, a computer program
H_0	Null Hypothesis
HDI	Human Development Index, a measure of social and economic development
HLM	Hierarchical Linear and Non Linear Modeling, a computer program
iidN	independent identically distributed Normal distribution
IOM	Institute of Medicine
LL	log likelihood
MEDLINE	An online database of 11 million citations and abstracts from health and medical journals and other news sources
-2*LL*	-2 times the log likelihood
ML	Maximum Likelihood, a method of estimation of the parameters of models
MLM	Multilevel Linear Models, a synonym for Hierarchical Linear Models
MLwiN	MultiLevel modeling using wiNdows, a computer program.
MMR	Vaccination for measles, mumps, and rubella
MSPAP	Maryland School
NDD	Neurodevelopmental disorders
NHS	National Health Service
OECD	Organisation for Economic Co-operation and Development
OLS	ordinary least squares
PATE	Population average treatment effect
PDD-NOS	Pervasive Developmental Disorder—Not Otherwise Specified
PROC GLIMMIX	Generalized Linear Multilevel Mixed Modeling procedure in SAS
Proc Means	The means procedure in SAS
Proc Mixed	The Mixed modeling procedure in SAS
RCTs	randomized controlled trials
REML	Restricted Maximum Likelihood, the default method of estimation in SAS mixed models
SAS	Statistical Analysis System, a computer program
SEMS	Structural Equation Models
SFA	Success For All, a highly structured curriculum for underachieving schools
SPSS	Statistical Package for the Social Sciences, a computer program
STATA	Statistical Analysis, a computer program

SY	School year
TAAS	Texas Assessment of Academic Skills
TCV	Thimerosal-containing vaccines
TEA	Texas Education Agency
Toep(2)	Two-banded Toeplitz covariance structure
TYPE	a typology variable
UN(1)	Banded Main Diagonal covariance structure
UNDP	United Nations Development Program
VAERS	Vaccine Adverse Event Reporting System
χ^2	chi square

Overview

This book endorses the premises of the human development index: a long, healthy life is better than one that is short and racked by illness; literacy and knowledge are better than illiteracy and ignorance; economic well-being is better than poverty; and freedom to pursue one's life goals is better than having one's life chances stunted by restricted opportunity and violence. Human development is better than human stagnation and regression. Accordingly, this book assumes that social problems are present when people are short-lived, illiterate, ignorant, poverty stricken, or unfree. Some concerned people aiming to alleviate social problems may choose a life of activism and social service; others may choose to conduct research about how to nurture individual freedom and reduce impoverishment.

One purpose of this book is to provide program managers and researchers with tools that can sharpen both their research on social problems and the inferences they draw from the findings. It does so by introducing conceptions of causality and multilevel modeling; and then elucidating the practical methods of contextual analysis, evaluative research, and research summaries. It probes the intersections of social problems, multilevel modeling, and causality, by focusing each core chapter on a substantive problem and by assessing relationships among the macro and micro variables that form the system being analyzed. These core chapters use typologies that group (i.e., classify or nest) the random macrolevel factors. Ideally, the resulting subgroups will be more homogeneous than the ungrouped macrolevel factors, and the variance of the associated random effects will approach zero. Such random effects clarify whether the categories of the typology account for the variance between the macrolevel units. Supplementing the core chapters are chapters on research summaries, one applies meta-analytic techniques to consolidate findings about the effectiveness of nurses aiming to contain health care costs; another investigates how experts have evaluated evidence concerning the alleged adverse effects of childhood vaccinations on autism. The concluding chapter then assesses the validities and causal level of the results of each core chapter.

Part 1 Introductory Essays

Chapter 1 introduces fundamental concepts of contextual analysis, spurious association, and interpretive chains of relationships and describes the chapters composing Part 1. Chapter 2 explicates the logic of *The Academic Mind*, a classic contextual analysis of the effects of McCarthyism on college teachers, in order to prepare the reader for the subsequent discussions of causal notions and multilevel modeling. Chapter 3, the first of two explicating notions of causality, discusses *stable association*, which includes classic causality (i.e., how the relationship between two variables is affected by a test factor) and robust dependence (i.e., how the relationship between two variables is affected by numerous test factors). It also discusses *potential outcomes*, which includes causality as an effect of an intervention and causal models for the effects of attributes. Chapter 4 continues this explication of notions of causality by elucidating *dependency networks*, which include graphical models, association graphs for loglinear models, generating processes, and structural models in policy research. By applying these conceptions of causality, the reader will be better able to evaluate and improve the causal inferences in their own work, and to critique the causal claims of other research studies, especially those of this book. Chapter 5 introduces relevant vocabulary, concepts, and notational conventions and draws on examples from the other chapters to clarify eleven uses for multilevel models.

Part 2 Contextual Studies

The chapters in Part 2 focus on determining the various *causes of an effect*. Chapter 6 serves as an introduction; for each of the next three chapters it shows how a particular social context and its covariates influence the response variable. By clarifying the study design, measures, results, and policy implications, this introduction further explicates how each of the three substantive chapters addresses the social problem inspiring the research. These chapters illustrate how researchers can apply multilevel models to analyze the implied causal forces of various micro- and macrounits at different levels of analysis, which; in descending order are: countries grouped by regions of the world; observations at different time points grouped by countries; and employees grouped by their organizational unit.

By studying factors that account for regional disparities in global human development, Chapter 7 highlights this book's focus on social problems and social theory (Merton 1982). Each subsequent core chapter explores a social problem stemming in part from gaps in human development. As defined by the United Nations Development Program (UNDP), the measure of human development, or, the human development index (HDI), combines economic, educational, and health measures. The response variables in Chapter 7 are a county's rank and scores on the HDI; the explanatory variables are the country's predominant culture and its measures on such potentially manipulable factors as national debt, corruption, civil disorder, democracy, and slavery. The countries of the world are the level-1 units and the regions of the world are the level-2 units.

This chapter also introduces the single-equation specification of the multilevel model and the use of qualitative typologies as explanatory factors. It distinguishes between correlations of variables and their implied causal effects by first estimating the effects of the covariates on a country's rank on the HDI. It then searches for implied causal effects by examining whether test factors eliminate the variability between the regions when each is used sequentially to nest the regions, ideally creating homogeneous region test-factor subgroups. When the regions are nested by typologies indicating a country's level of democracy or the absence of slavery (i.e., the absence of bonded servitude, sexual exploitation, and so forth), the resulting subgroups become homogeneous and the variance between the regions disappears. This suggests that, at least for these data, these nesting variables cause the disparities in human development among the regions of the world. These results, along with other findings about civilizations, national debt, corruption, and violence, suggest policy recommendations that would enhance human development.

Chapter 8 explores the causes of contemporary violence against European Jews. As recent events document, countries with high counts of such violence may also be vulnerable to broader attacks by extremists. The count of violent events in a country at a specific year is the level-1 variable, the country is the level-2 variable, and a country is grouped with other countries according to a typology based on the sizes of their Jewish and Muslim populations. Controlling for population and attitudinal characteristics of the country, longitudinal Poisson regression models test three competing theoretical models: mobilization of the perpetrators by their perceptions of events in the Middle East, bystander indifference, and cognitive ambivalence. The events in the Middle East mobilize the perpetrators of the violence; however, in the explanation for the weak response of ordinary Europeans to the violence, cognitive ambivalence proves to be more important than bystander indifference. The organizations providing the opinion surveys and counts of violent events aggregated these data to the level of the country; as such, inferences about the behavior of individual Europeans are problematic.

Putting causal inferences on more solid footing, Chapter 9 analyzes survey data on individual insurance claims workers to explain their discontent about the introduction into their workplace of a computerized detector of automobile insurance fraud. Networks of corrupt lawyers, physicians, and policyholders stage accidents and submit fraudulent claims that increase the costs of insurance. Countering such activities, insurance companies employ claims workers who focus on uncovering and curtailing insurance fraud. This chapter develops a dual macroequation and microequation approach in order to develop multilevel models. These models account for the discontent of the claims workers, many of whom perceived the automated fraud detector as threatening the substantive complexity of their work and job security. The workers (level-1) are grouped according to their claims offices (level-2), which then are classified by a typology of the number of new computer systems (zero, one, or two) being introduced in the offices. Because the simultaneous introduction of two new computer systems in an office reduces to insignificance the variance between claims offices regarding discontent with computerization, this factor is identified as causal (when the workers' ranks in the office and their receptivity to innovation are held constant). The practical research questions asked by top management inspired this study, but this chapter reconceptualizes the key variables theoretically, thereby broadening the initial evaluative research goal.

Part 3 Evaluative Research

These chapters focus on determining the *effects of a cause*, that is, the multiple consequences of an intervention. Serving as an introduction, Chapter 10 underscores a crucial social problem, the underachievement of students in primary and secondary schools in the USA. It first defines four policy orientations aiming to address this problem: fatalism, pragmatic activism, pluralism, and comprehensive school reforms (CSR) implemented by external consultants. It then presents the study designs the next two chapters use to evaluate CSR. These consultants provided consultation, training, professional development, and new instructional strategies to schools with underperforming students. These two chapters evaluate the extent to which the educational achievements of minority children (i.e., the effects) can be improved by the comprehensive school reforms (i.e., the cause). Because the schools were not randomly assigned to the treatment groups, these studies exemplify quasi-experiments, not true experiments. When subjects (in this case, schools) are not randomly assigned to the target and comparison groups, such evaluative studies as these are vulnerable to the effects of selection bias; that is, the observed effects may be due to differences between the subjects that existed prior to treatment.

As Chapter 10 explicates, these chapters compensate for selection bias and strengthen causal inferences by applying difference-in-differences (DID) designs coupled with, respectively, matching and propensity scores. A propensity score is a subject's (e.g., a school's) predicted probability of being in the treatment group. This probability is usually calculated from a logistic regression of treatment group membership (coded 1 or 0) on a set of antecedent predictors. The effects of that set of predictors on the response variable are minimized when differences in the propensity scores are controlled through the procedures of matching, stratification, or regression.

With controls for a set of covariates that may include propensity scores, the basic DID design takes the difference between two differences to estimate the average treatment effect: the first difference is that between the means of the response variable in the target group in time periods before and after it receives the target intervention; the second difference is that between the means of the response variable in the comparison group in those same time periods; and the third difference is the difference between those two differences. DID designs use each subject as its own control, thereby strengthening causal inferences about the average effect of a treatment.

Chapter 11 evaluates comprehensive school reforms in Harford County, Maryland; it aims to minimize the effects of selection bias by assessing matched target and comparison schools, as well as schools that are not matched. The aggregated test scores for a school at a time point are the level-1 variable, the school is the level-2 variable, and the type of school—target, matched, or not-matched—classifies the schools. The analysis compares the longitudinal change in test scores for the target schools receiving the comprehensive school reform treatment with the change in test scores for the matched comparison schools receiving the null treatment. It also compares the change in the not-matched schools receiving the null treatment with the change in the matched schools. The large difference in performance between the target schools and their matched schools, and the small difference in performance between the matched and not-matched schools, substantiate the causal interpretation of the effectiveness of the school reform treatment for this study period.

Evaluating the effects of comprehensive school reforms in Houston, Texas; Chapter 12 studies change across three school years (SY), from SY 1999–2000 through SY 2001–2002. By applying multilevel logistic regression models that specify appropriate covariance structures, this chapter analyzes binary response variables indicative of the performance of the schools. The estimates of the effects on the logit scale are transformed first into odds ratios and then into proportions, differences in proportions, and effect sizes; the tests of significance are based on the logit-scale effects and the odds ratios.

In conjunction with its DID design, this chapter aims to minimize selection bias by controlling for a school's propensity score, the predicted probability that a school receives the reform treatment rather than the comparison (i.e., null) treatment. By matching treatment and comparison schools on the basis of their propensity scores, or by including the scores as a covariate in a multilevel model, selection bias can be reduced because these scores remove the potentially spurious effects of the treatment on the response variable due to the many predictors in the regression equation that produced these scores. To obtain propensity scores, this study estimates logistic regression models that are appropriate for the binary (1, 0) response variable. The covariates in this regression model include a wide range of variables that reflect information prior in time to the assignment of schools to the treatments and the outcomes, thereby enabling the propensity scores to control for potentially spurious effects.

Chapter 12 primarily analyzes the effects of the reform treatment relative to the comparison schools that received the null treatment, but it also probes the effects of extra teachers who were assigned to the low-performing schools in fifth grade. The comprehensive school reforms improved the mathematics, reading, and writing achievement of the students from Time 0 to Time 1; and the extra teachers improved the performance of the students in the comparison schools from Time 1 to Time 2. Because of the logic of the DID design, the improvements in the comparison schools reduced the estimates of the overall effects of the reforms from Time 0 to Time 2. However, the extra teachers did not affect the scores on the fourth grade writing test, and this allowed the target reforms to exhibit significant overall improvements on this test. These gains are neither due to random fluctuation nor are the gains confounded with positive effects of Success for All (SFA), a highly structured innovative curriculum used in some of these schools.

Part 4 Research Summaries

Chapter 13 reviews current efforts to reform health care in the USA; it identifies two salient problems. Many people experience restricted access to appropriate care because they lack health insurance and the associated health care costs are prohibitive. Preventive vaccinations are available to children, but some parents restrict access to this care because of beliefs that vaccinations and autism are causally related. Regardless of which health care reforms are eventually implemented, the spiraling costs of health care will need to be contained and reduced.

Chapter 14 bears on the containment of medical costs by applying meta-analytic procedures to consolidate findings about the effectiveness of nurses who try to reduce hospital admissions, length of stay, and expense by preauthorizing and concurrently reviewing the medical necessity of care. Nurses with such functions could possibly help control the costs of Medicare and Medicaid, especially when these medical services are not constrained by other managed care programs. Meta-analysis enables researchers to consolidate findings from two or more different studies about the effect of a treatment on an outcome, and to determine the average effect size, along with its statistical significance, confidence interval, and appropriate scope of inference. Because this chapter develops the concepts of fixed effects and random effects, readers unfamiliar with these concepts could benefit from reading this chapter early on.

Chapter 15 confronts the second health problem, the misleading beliefs of parents about the efficacy and side effects of preventive childhood vaccinations. Although no credible scientific evidence links childhood vaccinations to increasing rates of autism, many parents, who may be overly conscientious and protective, still are not allowing their children to be vaccinated because of their beliefs that vaccinations cause autism. As the number of un-vaccinated children increases, the increased number of such children increases their own risk of ill health, as well as increasing the risk of contagion and the spread of preventable diseases to others.

This chapter traces some of the origins of the “vaccination-autism” controversy to questionable but highly publicized studies; describes the beliefs of parents of autistic children, a number of whom state that the vaccinations caused their child’s autism; and summarizes the procedures that groups of health scientists in Europe and the United States applied in their separate assessments of the evidence concerning childhood vaccinations and autism-spectrum disorders (i.e., autism, atypical autism, and Asperger’s syndrome); both groups concluded that the scientific evidence does not support a causal linkage. The European group from the Cochrane Collaboration applied a meta-analytic approach. The American group from the Institute of Medicine applied an informal Bayesian approach in which they initially assumed a position of neutrality concerning any linkages between vaccinations and adverse consequences, and then modified their initial position one way or the other on the basis of their scrutiny of the evidence.

Applying aspects of these approaches, the concluding Chapter 16 attempts to tackle the problem of unsubstantiated beliefs indirectly by clarifying notions of causality, defining criteria for judging different aspects of the validity of a study, and applying these criteria to gauge the level of causality of the empirical findings of the multilevel modeling. Hopefully, the reader's participation in this exercise in evaluating evidence will help sharpen critical skills so that he or she will not be misled by any spurious claims of causality in this book, or elsewhere.

Part I
Introductory Essays

Chapter 1
Concepts and Considerations

> *Causal inferences are made from* observational studies, natural experiments, *and* randomized controlled experiments. *When using observational (non-experimental) data to make causal inferences, the key problem is* confounding. *Sometimes this problem is handled by subdividing the study population* (stratification, *also called* cross-tabulation), *and sometimes by modeling. These strategies have various strengths and weaknesses, which need to be explored.*
>
> —David A. Freedman (2005, 1)

Providing a background for the subsequent expositions, this chapter introduces three fundamental concepts of multilevel modeling and causality: contextual analysis, spurious association, and chains of relationships. It then previews topics covered in Chapters 2 through 5.

Contextual Analysis

Human behavior can be conceptualized as being influenced by three factors: (1) a person's prior personal dispositions, which include perceptions, attitudes, values, desires, beliefs, capabilities, and schemas; (2) the impingement of social environments on that person; and (3) the interactions between the predisposing and environmental factors. These factors imply a multilevel analysis of at least two levels, that of the individual (referred to as level-1) and that of the environment (referred to as level-2). A *contextual study* exemplifies a multilevel analysis because it includes variables on the individual and on the environment. *Contextual effects* are the cross-level interactions between the personal and environmental variables, and the study of these interactions defines *contextual analysis*. The latter includes *comparative analysis*, which links the level-2 variable directly to a level-1 response.

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_1, © Springer Science+Business Media B.V. 2011

These hypothetical data of Box 1.1 can illustrate a contextual study, comparative analysis, and contextual analysis (Lazarsfeld et al. 1972, 222–223):

Box 1.1 Data for interpreting two individual characteristics and one collective characteristic

Context	Unexceptional teachers			Exceptional teachers			Marginal table		
Individual Students	Ethnic-majority	Ethnic-minority	Total	Ethnic-majority	Ethnic-minority	Total	Ethnic-majority	Ethnic-minority	Total
Passing test	350	300	650	375	375	750	725	675	1,400
Not passing test	150	200	350	125	125	250	275	325	600
Total	500	500	1,000	500	500	1,000	1,000	1,000	2,000

Putting aside other differences in family background, for ethnic-majority and ethnic-minority students, these data report the number of school children passing an achievement test in two contexts: students in classrooms taught by unexceptional teachers (1,000 students) and students in classrooms taught by exceptional teachers (another 1,000 students). The two individual characteristics are a student's ethnic affiliation and his or her achievement; the contextual characteristic is the quality of the classroom teachers who teach the students. Research findings suggest that ethnic-majority students often outperform ethnic-minority students on standardized tests of academic achievement. This is a level-1 relationship because both variables—ethnicity and achievement—are characteristics of students. The hypothetical data in the right-most marginal table of Box 1.1 echoes this relationship: for the 2,000 students, 72.5% of the ethnic-majority students pass the achievement test compared with 67.5% of the ethnic-minority students; the difference of 5 percentage points indicates a gap in achievement.

The comparative analysis directly links the students' academic performance (level-1) to the quality of their classroom teachers (level-2): Collapsing the distinction between ethnic-majority and ethnic-minority students (i.e., marginalizing over ethnicity), the first two total columns show that 65% of the students in classrooms taught by unexceptional teachers passed this test compared with 75% of the students in classrooms taught by exceptional teachers. Apparently, the quality of the classroom teachers has a positive effect of 10 percentage points.

The contextual analysis examines how the quality of the classroom teachers (level-2) influences the relationship between the two individual characteristics (at level-1). The left-most partial table shows that when the teachers are unexceptional, ethnic-majority students are more likely to pass the test than ethnic-minority students, 70% compared with 60%, a difference of 10 percentage points. However, the right-most partial table shows that when the teachers are exceptional, there is no difference in achievement between the two types of students; 70% of the ethnic-majority students pass the test and 70% of the ethnic-minority students pass the test.

Box 1.2 summarizes these differences, emphasizing the effects of the different contexts:

Box 1.2 Percentage of ethnic-majority and minority students passing a test in different contexts

Context: Students:	Unexceptional classroom teachers (%)	Exceptional classroom teachers (%)	Difference in gain (%)
Ethnic-majority	70	75	5
Ethnic-minority	60	75	15
Minority gap	−10	0	10

Holding constant the individual-level effects of ethnic background and the comparative effect of the quality of the classroom teachers, the contextual effect (synonymously, cross-level interaction effect) suggests that ethnic-minority students in classrooms taught by exceptional teachers will exhibit more improvement in their performance (15 percentage points) than ethnic-majority students in the classrooms taught by exceptional teachers (5 percentage points); an average cross-level interaction effect is 5% (Difference in Gain divided by 2). The evaluative studies of Part 3 of this book also suggest that improving teacher quality, a contextual characteristic of the classroom, is fundamental to educational reform.

More generally, a contextual study requires at least two hierarchically organized units in which the macrounits include (i.e., contain) the microunits. Although this book focuses primarily on data hierarchies of two levels, here is an example of a data hierarchy of four levels: school districts include schools, schools include classrooms, and classrooms include students. A macrounit is a collective and the microunits are its members (Lazarsfeld and Menzel [1961] 1972). The collectives have properties in common and the measures on these properties vary: classrooms are a collective, the quality of the teacher is a property of a classroom, and teacher quality varies from classroom to classroom. The members of these collectives have certain properties; the measured values on these properties vary: ethnic background is a property of the students and its nominal measure may be majority or minority; academic achievement is a property of the students, its measure may be binary (e.g., pass or fail a test) or may vary continuously from low to high. A contextual analysis examines how different collectives with the same key properties, but with different measured quantities on these properties, affect the relationship between two measures on the properties of its members. More concretely, classrooms with varying qualities of their teachers differentially affect the relationship between the type of student and amount of academic achievement. Thus, ethnic-minority students in classrooms taught by exceptional teachers exhibit more improvement in their performance than ethnic-majority students in similar classrooms taught by exceptional teachers. For each ethnicity, improvement is the difference between the average score in academic achievement when the students are in classrooms taught by exceptional teachers, compared with the average score in achievement when the students are in classrooms taught by unexceptional teachers. This difference is larger for the ethnic-minority students than for the ethnic-majority students, and the difference between these differences is a measure of the contextual effect.

Spurious Correlations

A contextual analysis is a special case of a more general procedure in which the assumed causal effect of variable x on variable y is examined under controls for different antecedent test factors t. If the hypothesized causal relationship $x \rightarrow \mathrm{y}$ disappears when a test factor t that is prior to both x and y is controlled ($yx.t = 0$), then the initially observed relationship is spurious (i.e., x and y are conditionally independent, given a control for a prior t), as the following five examples illustrate.

Fire Engines and damage to property (Lazarsfeld 1955a, 123)

It has been found that the more fire engines that come to a fire, the larger is the damage. Because fire engines are used to reduce damage, the relationship is startling and requires elaboration. As a test factor, the size of the fire is introduced. The partials then become zero and the original result appears as the product of two marginal relationships; the larger the fire, the more engines—and also the more damage.

Marital status and candy consumption (Zeisel 1985, 151–152, 158)

Again we begin with a correlation in which being married is associated with a difference in behavior, in this case with the amount of candy eating. Fewer married women eat candy than do single women. . . . The relation between being married and eating less candy is fully explained by the fact that married people are, on the average, older than single people, and because older people eat candy less frequently. If married and single people of equal age are compared, the association between marital status and candy eating disappears. . . . In symbols:

getting married ⟵ getting older ⟶ eating less candy

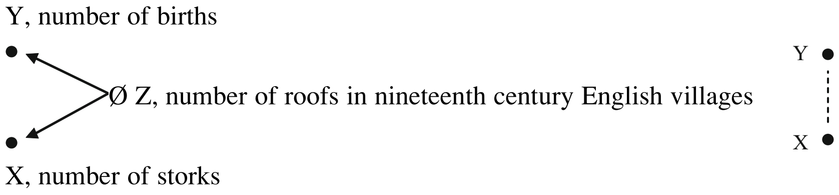
Note the reversed position of the first arrow: Getting older is not only the cause of eating less candy but also the cause—not the effect of—getting married.

Marital status and candy consumption (Simon [1954] 1957, 38–39)

The data consist of measurements of three variables in a number of groups of people: x is the percentage of members of the group that is married, y is the average number of pounds of candy consumed per month per member, z is the average age of members of the group. A high (negative) correlation, r_{xy}, was observed between marital status and amount of candy consumed. But there was also a high (positive) correlation, r_{yz}, between marital status and age. However, when age was held constant, correlation $r_{xy.z}$, between marital status and candy consumption was nearly zero. ... The correlation between marital status and candy consumption is spurious, being a joint effect caused by the variation in age.

Births and storks (Wermuth 2003, 50)

[This] example for spurious association was used by G. Yule more than 100 years ago to argue that correlation is not causation. Ignoring the common explanatory variable, i.e., marginalizing over Z, leaves Y and X associated:



Fixing levels of the common explanatory variable, i.e., conditioning on Z, shows Y independent of X.

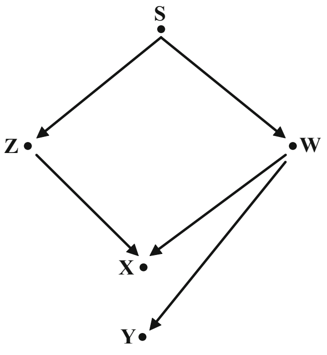
Spurious association with temporal information (Definition 2.7.5, Pearl [2000] 2009, second edition, 56–57)

Two Variables X and Y are spuriously associated *if they are dependent in some context S, if X precedes Y, and if there exists a variable Z satisfying*:

1. $(Z \perp\!\!\!\perp Y \mid S)$. [$Z$ is conditionally independent of Y in context S]
2. (Z not $\perp\!\!\!\perp X \mid S$). [$Z$ is not conditionally independent of X in context S]

Figure 2.5(b) [reproduced below] illustrates the intuition behind Definition 2.7.5. Here, the dependence between X and Y cannot be attributed to causal connection between the two because such a connection would imply dependence between Z and Y, which is ruled out by condition 1.

Pearl's Fig. 2.5(b): Illustration of how temporal information permits the inference of spurious associations (between X and Y) from the conditional independencies.



As the above five examples show, when a prior relevant test factor eliminates the $x \rightarrow$ y relationship, then the initial correlation is spurious and not causal. However, if the test factor t intervenes between x and y, a chain of relationships may result.

Chains of Relationships

A three-variable causal chain results when the initial $x \rightarrow$ y relationship is eliminated by a control for an intervening test variable t, then $x \rightarrow t \rightarrow$ y, as the following five examples illustrate.

Marriage and absenteeism (Lazarsfeld 1955a, 124)

> We, here, shall use the term "interpretation" for Type MI. The difference between "explanation" and "interpretation" in this context is related to the time sequence between '*x*' and '*t*.' In an interpretation, the '*t*' is an intervening variable situated between '*x*' and '*y*' in the time sequence. … It was found during the war [i.e., World War II] that married women working in factories had a higher rate of absence from work than single women. … Test factor: more responsibilities at home. This is an intervening variable. If it is introduced and the two partial relationships—between marital status and absenteeism—disappear, we have an elaboration of type MI.

Marriage and absenteeism (Zeisel 1985, 157–158)

> Why do married women have a higher rate of absenteeism than do single women? Because married women have more housework and housework results in greater absenteeism. … More housework is the result of being married and is, in turn, the cause of higher absenteeism. In symbols—the arrows point in each case from the cause to the effect—the relation would read as follows:
>
> $$\text{getting married} \longrightarrow \text{more housework} \longrightarrow \text{more absenteeism}$$
>
> The important point is that the relation between more housework and getting married cannot be reversed. To have more housework will *not* increase the likelihood of getting married.

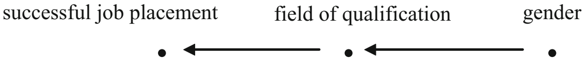
Marriage and absenteeism (Simon [1954] 1957, 39)

> The data consist again of measurements of three variables in a number of groups of people: x is the percentage of female employees who are married, y is the average number of absences per week per employee, z is the average number of hours of housework performed per week per employee. A high (positive) correlation, r_{xy}, was observed between marriage and absenteeism. However, when the amount of housework, z, was held constant, the correlation $r_{xy.z}$ was virtually zero. In this case, by applying again some common sense notions about the direction of causation, we reach the conclusion that z is an intervening variable between x and y: that marriage results in a higher average amount of housework performed, and this, in turn, more absenteeism. … There was a true causal relationship [between x and y], mediated by the intervening variable z. Clearly, it was not the statistical evidence, but the "common sense" assumptions added afterward, that permitted us to draw these distinct conclusions.

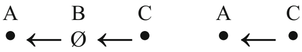
Spurious dependence (Wermuth 2003, 49)

> The first example for spurious dependence concerns the question of discrimination against women and data from the German labor market for academics,

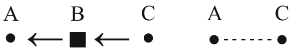
whose field of qualification was either mechanical engineering or home economics. The well-fitting independence graph is



Ignoring the intermediate variable, i.e., marginalizing over *B* (Ø) leaves A dependent on *C*: the data appear to indicate discrimination, since men have a more than five times higher chance for successful job placement.



However, including the information of the field of qualification by fixing levels of the intermediate variable, i.e., conditioning on B (■), shows A independent of C.



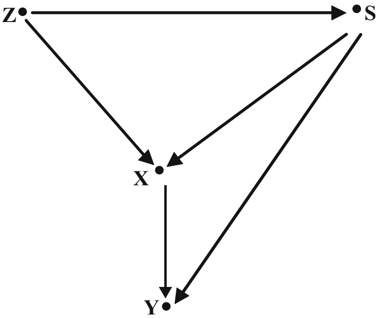
Genuine causation with temporal information (Definition 2.7.4, Pearl [2000] 2009, second edition, 56)

A variable X has a causal influence on Y if there is a third variable Z and a context S, both occurring before X, such that:

1. (Z not ⫫ Y ∣ S) [Z is not conditionally independent of Y given context S]
2. (Z ⫫ Y ∣ S ∪ X). [Z is conditionally independent of Y given context S and X]

Temporal precedence is now used to establish Z as a potential cause of X. This is illustrated by Fig. 2.5(a): If conditioning on X can turn Z and Y from dependent to independent (in context S), it must be that the dependence between Z and Y was mediated by X; given that Z precedes X, such mediation implies that X has a causal influence on Y.

Pearl's Fig. 2.5(a): Illustration of how temporal information permits the inference of genuine causation (between *X* and *Y*) from the conditional independencies.



The authors of the above five examples all conceptualize the intervening variable as having a causal effect on the response variable; however, they interpret the causal effect of the antecedent variable differently (e.g., Simon and Pearl compared with Wermuth).[1] The subsequent chapters consider interpretations of such empirical relationships in detail.

Contextual Analysis and Multilevel Models

Chapter 2 reviews the logic and measures of a classic contextual analysis, *The Academic Mind: Social Scientists in a Time of Crisis*, by Lazarsfeld and Thielens (1958). This study probes how political extremism affected the job security and academic freedom of teachers of social science soon after the difficult period of McCarthyism, circa 1955; its logic could inform present-day studies of "political correctness" in contemporary academia. Forming a social mechanism, three of its pivotal variables are the professor's occupational apprehension concerning job security, a variable at level-1; the professor's underlying permissiveness (i.e., openness to new ideas) relative to conservatism (i.e., prudence), also a variable at level-1; and the severity of the incidents about academic freedom at the academic institution, a variable at level-2. Institutional incidents interacted with the professor's permissiveness to produce increases in apprehension. The strength of this contextual effect varied with the initial degree of permissiveness, the severity of the incidents, the protectiveness of the administration, and the quality of the college.

In many ways, this classic study is still state of the art: Lazarsfeld and Thielens focused on an important problem, developed appropriate concepts and measures, and uncovered the nuances of the relationships. However, they did not fully explicate their conceptions of causality and they could not apply multilevel statistical modeling, which had not yet been invented. Consequently, the investigators could not simultaneously analyze the effects of all of their key variables; their a-statistical research technology, which was based on the close inspection of cross-tabulations among a few variables, limited the range of their analysis. Toward advancing Lazarsfeld and Thielens's statistical methods and notions of causality, the next chapters address some limitations of their fine, classical analysis. Chapters 3 and 4 develop notions of causality and Chapter 5 sketches uses for multilevel models drawing on examples from this present book.

Notions of Causality

Lazarsfeld and Thielens were appropriately very cautious about asserting causal relationships among their variables, even though their explanatory structure linked contexts, mechanisms, and outcomes (Pawson 1989, 2006, 73–104; Pawson and Tilley 1997). Searching for social mechanisms, their notion of a causal relationship was based on an elaboration procedure in which the investigator examined the stability of the $x \rightarrow y$ relationships at level-1 by controlling for the effects of various contextual test factors t at level-2, usually one at a time. In his earlier writings, Lazarsfeld (1955a, 125) explicitly defined a causal relationship between two attributes: "If we have a relationship between 'x' and 'y'; and if for any *antecedent* test factor, the partial relationships between x and y do not disappear, then the original relationship should be called a causal one. It makes no difference here whether the necessary operations are actually carried through or made plausible by general reasoning."[2]

Table 1.1 outlines the topics covered by the two chapters on notions of causality, noting the roots of the conceptions and the illustrative examples. These types

Table 1.1 Three notions of causality

Notions of causality	Sources	Examples discussed in Chapter 3 and Chapter 4
Three causal notions	Cox and Wermuth (1996, 2001, 2004)	General discussion
Stable association (Chapter 3)	Simon's formalization of spurious correlation ([1954] 1957), Hill (1965)	Citation
Causality in classic social research	Lazarsfeld's (1955a) elaboration procedure (with Kendall 1950), Hyman (1955), (Glock 1967)	General discussion
Causality in time series	Granger (1986)	Brenner's (1973, 2005) studies of economic stresses and mental health
Meta-analysis	DerSimonian and Laird's (1986) fixed- and random-effects paradigm	Consolidation of effects of precertification of medical care, data from Chapter 14
Potential outcomes (Chapter 3)	Rubin (1974) and Holland (1986)	General discussion
Causality as an effect of an intervention	Rubin's (1974, 2006) causal model, Gelman and Meng (2004), Imbens and Wooldridge (2009)	Data on computerization of work from Chapter 9
Causal effects of attributes	Coleman's (1964, 1981) causal model	Data on computerization of work from Chapter 9
Dependency networks (Chapter 4)	Simon's ([1953] 1957) formalization of causal ordering and identifiability	Citation
Graphical models	Cox and Wermuth (1996), Pearl ([2000] 2009, second edition), Morgan and Winship (2007), and many others	General discussion and analysis of data on an election campaign
Association graphs for loglinear models	Goodman (1972a, b), Agresti (1996), Christensen (1997)	Interaction effects among the issues in an election campaign
Path-analytic generative processes	Jöreskog and Sörbom (1979), Duncan (1975), Goldthorpe (1998)	Survey of Cambridge and MIT exchange students
Causality in policy research. "All Causes" Models	Heckman (2005a, b), Skrondal and Rabe-Hesketh (2004)	General discussion

of causal notions—stable association, potential outcomes, and dependency networks—are based on Cox and Wermuth's distinctions (1996, 58, 219–227; 2001, 65–69; 2004). These notions call to mind the zone of causality achieved by the best studies in, respectively, survey research, experimentation, and structural equation modeling.

Stable Association and Potential Outcomes

Chapter 3 builds on Lazarsfeld's definition of a causal relationship by defining two of the three notions of causality, stable association and potential outcomes. It provides answers to this question: Is the x → y relationship spurious or potentially causal.

Stable Association

Causality as stable association (Cox and Wermuth 2001, 65–68) aims to determine if a relationship between two ordered variables $x \rightarrow y$ is robust, when the investigator tests the relationship by controlling for potential explanatory variables that could determine both x and y. Subject-matter knowledge and considerations of temporal order are crucial in establishing which of the two variables is prior to the other, and the priority of the test variables. If the $x \rightarrow y$ relationship is robust, and a full range of allowable test variables do not destroy their association, then there may be a causal relationship in the sense of Herbert Simon's formalization ([1954] 1957, 38–39). The heuristic criteria for causality of medical researcher Hill (1965) are consistent with Simon's analysis; here is Cox's succinct summary (1992, 292) of Hill's criteria:

> An effect obtained in an observational study is relatively likely to have a causal interpretation [if the effect]
>
> (a) Is large
> (b) Is reproduced in independent studies
> (c) Shows a monotone relation with "dose"
> (d) Corresponds to a "natural experiment"
> (e) Behaves appropriately when the potential cause is applied, removed, and then reinstated
> (f) Is consistent with subject-matter knowledge
> (g) Is, for example, predicted by reasonably well-established theory

Guided by these considerations, the section in Chapter 3 on stable association discusses causal notions in classic social research, time series analysis, and meta-analysis. After this discussion, the chapter considers causality as an effect of an intervention—Rubin's potential outcomes model and Coleman's stochastic causal model for attributes.

Potential Outcomes

Randomized clinical trials exemplify the application of experimental designs to the testing of the effects of prototypical drugs. Ideally, the design calls for the random assignment of subjects to treatment and control groups, with the random assignment ensuring that predisposing factors appear equally in both groups of subjects. The treatment group is exposed to the experimental treatment and the control group is exposed to the control treatment, which may be a placebo, a benchmark standard, or an established drug.

Rubin's (1974) potential outcomes model of causality is rooted in this experimental perspective. The causal effect is defined as the synchronic difference between the outcome for a person when assigned to the experimental treatment compared with the counterfactual outcome for that person if that person had been assigned to the control treatment (and vice versa). Because the person cannot be exposed simultaneously to both the experimental and control treatments, an exact causal effect is difficult if not impossible to quantify; this conundrum is referred to as the fundamental problem of causal inference. But, if the experimental and control groups are composed of closely equivalent subjects due to random assignment, matching, or strong statistical controls, then an average causal effect can be estimated.

In observational studies of interventions in which random assignment is absent, the target and control groups will be vulnerable to preexisting differences between the subjects exposed to the target treatment and those exposed to the control treatment. Rubin (1974, 2006) and his many colleagues (in Gelman and Meng 2004) have applied his potential outcomes perspective to observational studies geared toward assessing causality as an effect of an intervention (Cox and Wermuth 2001, 68–69). They have developed procedures for reducing bias via matching, matching combined with regression analysis, and propensity scores, all of which can ameliorate the shortcomings due to the lack of random assignment. The section on potential outcomes discusses these procedures.

Causal Effects of Attributes

Using a common data set, this chapter also compares Rubin's notion of causality that probes the various *effects of a cause* to Coleman's (1964, 1981) causal model for basic attributes that probes the various *causes of an effect*. By centering the covariates by their overall means, the estimate of a treatment effect in Coleman's model becomes mathematically identical to the estimate of an average treatment effect in Rubin's causal model. This explication reintroduces Coleman's once path-breaking but now neglected causal conceptions to contemporary discussions of causality in the social sciences. Because Coleman's model readily quantifies the effects of intrinsic characteristics of a person on a response variable, it complements Rubin's causal model that readily quantifies the effects of a manipulated intervention on the response variable.

Dependency Networks

Chapter 4 focuses on the third notion of causality, dependency networks, providing answers to this question: "How do the variables form a system, a network of hypothesized relationships?" Causality based on stable association and causality based on potential outcomes require that the test variables be antecedent to the cause x or at least on equal footing with x (i.e., simultaneity of x and t); these controls cannot intervene between x and its effect y, x → y. Contrariwise, this chapter develops the idea that a test factor may intervene between x and y producing a dependency network. A minimal dependency network includes an intervening variable t mediating the effect of x on y: $x \rightarrow t \rightarrow y$. More elaborate networks include blocks of prior antecedent variables, blocks of intervening variables that have different priority, and blocks of response variables. The ordering of the variables in the system is crucial; most often the investigators order the variables by applying subject-matter knowledge, theory, temporal order (e.g., earlier → later), ordering by other heuristic principles (e.g., general → specific), and various statistical tests (e.g., the much larger of two reciprocal effects is the correct direction of effect; or, with the direction of effect reversed, the model that fits the data best justifies the direction of effect). Chapter 4 shows how graphical models, association graphs, generative process models, and "all causes" structural models formalize such assumptions.

Graphical Models

In their expositions of graphical models, Cox and Wermuth (1996, 135–142, 172–177; 2001, 70–74; 2004) sketch this illustrative recursive procedure: There are four blocks of variables; the variables in each block are defined. After exploring and cleaning the data, the analysis begins by regressing y, a variable in block 4, on variables in block 3, block 2, and block 1. Then variables in block 3 are regressed on variables in block 2 and block 1. Then block 2 variables are regressed on block 1 variables. The effect sizes and their significance are reported in tables. Finally, the variables that are conditionally dependent are depicted by linkages in a graph and the findings are interpreted. In the graph, the blocks are depicted as boxes, the variables (i.e., the nodes) in a box are depicted as small circles, and the edges (i.e., linkages) connecting the variables are depicted as straight lines with or without arrow heads for, respectively, asymmetric and symmetric relationships. Conditionally independent relationships between two variables are not depicted by lines. Such graphs as these differ from diagrams depicting path analytic models and structural equation models (SEMs). In these models, missing edges do not necessarily correspond to statements about conditional or unconditional independencies, error terms are included as nodes, and effects cannot always be interpreted as regression coefficients (Cox and Wermuth 1996, 194–196). This section of Chapter 4 on graphical models illustrates this recursive model-building strategy by analyzing the voting choice in the 1992 presidential election in the United States; the recent 2008 election had similar issues and social determinants of vote.

Association Graphs

Goodman (1972a, 1972b, 1978) developed graphs similar to those of path analysis to depict causal relations based on findings from analyses that apply his loglinear and logit models. This section of Chapter 4 illustrates Goodman's approach by developing association graphs that depict the findings from an empirical analysis of the interactions among four issues and the voting choice in the 1992 presidential election. The loglinear models identify the crucial interactions and logistic regression models quantify the effects of the issues and their interactions on the voting choice.

Generative Processes

The mechanism of such processes link a putative cause x to a response y through a chain of intervening interpretive variables: $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow y$. Because such mechanisms are rare in contemporary social science, this section of Chapter 4 develops an example in some detail. It examines how the different pedagogies of academic institutions (x) affect the research experience their exchange students report (t_1), which affects the students' self-reported confidence in their basic research skills (t_2), which affects their self-reported innovative ability (t_3), which directly effects their indicators of consideration of novel uses for technology (y). The models are estimated using AMOS (Analysis of Moment Structures), a computer program for estimating path-analytic and structural equation models (SEMs). The example tests various alternative orderings of the variables by assuming directions of effect that differ from those in the postulated model, and then eliminating the alternative models that do not fit the data closely.

Econometric Models in Policy Research

This section of Chapter 4 briefly discusses aspects of Heckman's rich and nuanced notions of causality in policy research, as he explicated his ideas for sociologists (2005a, 2005b). His scientific model of causality begins with a theory of the substantive process being studied. It then focuses on the definition, identification, and estimation of effects: (1) *define* a model of the phenomena based on a scientific theory, the model is a set of hypotheticals or counterfactuals; (2) *identify* relevant parameters of the model using hypothetical population distributions and mathematical analysis (i.e., explore the model's implications using hypothetical data); and (3) *estimate* the parameters empirically using samples of real data to test the model and the underlying theory (i.e., ground the model empirically). Heckman's modeling of educational reforms illustrates this three-step process, which can also be applied to computer simulations and agent-based models (see Smith 2010b), mathematical models, and SEMs.[3]

Uses for Multilevel Models

Substantiating the general points by drawing upon examples from the subsequent chapters, Chapter 5 clarifies 11 uses for multilevel modeling that this book develops. In doing so, it also discusses criteria for assessing the goodness of fit of multilevel models; differences between maximum likelihood (ML) and restricted maximum likelihood (REML) estimation; and relevant vocabulary, concepts, and notational conventions. Table 1.2 organizes the 11 uses under three main topics—contextual studies, evaluative research, and research summaries—which are the core of this book. Understanding these uses facilitates the learning of these methods.

Contextual Studies

Chapter 7 on human development illustrates three uses. When, as in the example for use (1), *analysis of clustered macrounits*, observations on such macrounits as countries are clustered because the countries are contained within their geographic region, the multilevel model can quantify the variance between the regions and the variance of the countries within the regions, thus correcting the standard errors for the clustering. Continuing this example, this leads naturally to use (2), *testing the significance of the level-2 variance*. If the variance between the regions is not statistically significant, then, because the clustering is minimal, regular regression procedures can be used and multilevel statistical models are not necessary. If, on the other hand, the variance between the regions is statistically significant, then the analysis can focus on introducing factors that account for this variance. These factors can be at the regional level, or they can classify the levels of the random

Table 1.2 Eleven uses for multilevel models

Salient examples of use	The chapters providing the illustrative examples
Contextual studies	
1. Analysis of clustered macro units	Chapter 7, Global Human Development
2. Testing the significance of a level-2 variance	Chapter 7, Global Human Development
3. Nesting the level-2 random variable by a classification typology	Chapter 7, Global Human Development
4. Modeling counts of events	Chapter 8, A Globalized Conflict
5. Modeling clustered microlevel units	Chapter 9, Will Claims Workers Dislike a Fraud Detector?
6. Applying macro and micro explanatory variables	Chapter 9, Will Claims Workers Dislike a Fraud Detector?
Evaluative research	
7. Estimating the treatment effects in differences-in-differences (DID) designs	Chapters 10 through 12 on evaluative research
8. Borrowing strength	Chapter 11, Target, Matched, and Not-Matched Schools
9. Using propensity scores to reduce bias	Chapter 12, Using Propensity Scores
Research summaries	
10. Meta-analysis	Chapter 15, Gatekeepers and Sentinels
11. Evidence for causal inferences	Chapters 3 and 4 on three zones of causality; Chapter 15, Childhood Vaccinations and Autism, and Chapter 16, Gauging Causality in Multilevel Models

effect of region into groups; this *nesting of the random variable by a typology* exemplifies use (3). Nesting creates subgroups based on the intersection of the random macrovariable and the nesting factor; these subgroups usually will be more homogeneous than the unnested random macrovariable. If the random effects in these subgroups are minimal, then the variance between the regions will approach zero, and the nesting factor may be a cause of the regional differences.

Response variables can be dichotomous, ordinal, counts, or continuous variables. Chapter 8 exemplifies use (4), *modeling counts of events*. It shows how generalized linear mixed models can be applied to a specific case; namely, the change over time in counts of violent incidents against Jewish people living in contemporary Europe.

By reporting the effects of a new computer system on employee morale, Chapter 9 illustrates two uses. The investigators administered survey questionnaires to all of the claims workers in six insurance claims offices, and not to a random sample of workers stratified by their office. Because these observations are clustered and not independent, a multilevel model was developed that lessens the troublesome consequences of clustered data, this mitigation exemplifies use (5), *modeling clustered microlevel units*. Continuing this example, the multilevel model combines macrolevel and microlevel equations. A macrolevel typology that classifies the claims offices was introduced into the equation; it explained the variance between the claims offices, this explanation exemplifies use (6), *applying macro and micro explanatory variables.*

Evaluative Research

The three chapters on evaluations of comprehensive school reforms illustrate how multilevel models can estimate the average treatment effects defined by difference-in-differences (DID) study designs; this is use (7). To quantify the average treatment effect, the basic DID design takes the difference between two differences: The first difference is that between the means in the target group in time periods before and after the intervention. A second difference is that between the means in the comparison group in the same time periods as those of the target group. Quantifying the average treatment effect, the third difference is that between the first two differences. Chapter 10 introduces this basic DID design.

The importance of the DID design stems in part because of its consistency with the potential outcomes causal perspective: Prior to assignment to either the target treatment or the comparison treatment, a unit has potential outcomes under either of these treatments. After assignment to one treatment, the unit has realized outcomes due to that treatment, and counterfactual outcomes due to the other treatment. The causal effect of the received treatment cannot be quantified because information about the counterfactual outcome is missing. However, if the treatment and comparison groups are closely matched, then an average causal effect can be quantified following the logic of the difference-in-differences design.

The chapters presenting evaluative research elaborate this basic DID design. By applying multilevel modeling to three treatment groups of elementary schools—target, matched, and not matched—Chapter 11 illustrates use (8), *borrowing strength.* The use of three treatment groups rather than two strengthens causal inferences about the effectiveness of the school reforms. This chapter borrows strength by using information on all of the schools, rather than on some of the schools, and by developing one parsimonious equation for each response variable that applies to all of the schools, with each school having a unique pattern of measures on the covariates.

Chapter 12 exemplifies use (9), *using propensity scores to reduce bias.* These scores are the predicted probability of a unit being assigned to the treatment group. Here, they are based on the logistic regression of membership in the treatment group coded (1) versus the comparison group coded (0) on 16 covariates that are assumed to represent information collected prior to assignment to the treatment and comparison schools and to the outcomes. The use of these scores as a mean-centered covariate in the multilevel logistic regression model controls for the effects of the variables composing the scores; it improves the precision of the estimates of the effects of the educational reforms.

Research Summaries

Findings from studies on the same topic may accumulate, but they do not become cumulative results unless they are consolidated. The fixed- and random-effects paradigm of Chapter 14 exemplifies use (10), *meta-analysis* of scattered findings to create cumulative results. The chapter applies meta-analytic techniques, which are implemented by a spreadsheet program, to consolidate findings about the effects of utilization review nurses on medical expense. The spreadsheet program first calculates the homogeneity of the fixed-effects estimates across the studies, and, if the estimates diverge, it calculates the random-effects estimates. But are such effects causal?

Chapter 15 on the null effect of childhood vaccinations on autism, and the concluding Chapter 16 answer such questions by first applying criteria for valid research to the studies at hand. Then, based on the veracity of the findings, the zone of causality achieved by a study can be classified according to the definitions of Chapters 3 and 4 as no causality, stable association, potential outcomes, or dependency networks. Because of limitations in design and implementation, and the absence of replications, most studies in the social sciences do not meet the most stringent criteria for causal inference. As Cox and Wermuth (1996, 219) state: "Our reason for caution is that it is rare that firm conclusions about causality can be drawn from one study, however carefully designed and executed, especially when that study is observational." Consequently, this book introduces the concept of *zones of causality*. This notion rewards investigators and their studies by giving some credit for aiming to meet the most stringent criteria; replications of the studies may enable the research to reach stricter standards.

Multilevel models can thus establish *evidence for causal inferences*, use (11). Such evidence enables the investigator to ascertain whether the effects are correlational or causal. The answer depends on the rigor of the study and on the governing notion of causality that is applied in the assessment of the findings.

Conclusion

In sum, the subsequent introductory essays provide a general introduction to contextual studies, notions of causality, and multilevel modeling. By explicating an exemplary classic study, Chapter 2 interrelates contextual analysis and multilevel modeling; the application of contemporary statistical methods and a sharpening of its notions of causality would improve this fine study. Addressing these issues, Chapters 3 and 4 clarify contemporary notions of causality, and Chapter 5 discusses 11 uses for multilevel models. With this introductory knowledge at hand, the previously uninitiated reader will be better prepared to understand, evaluate, and advance multilevel modeling, especially of social problems.

Endnotes

[1] Pearl's d-separation criterion suggests that in this chain of relationships (Pearl [2000] 2009, second edition, 16–17, and personal communication of 4/2/2010):

marital status (i) → amount of housework (m) → more absenteeism (j)

1. The amount of housework blocks the effect of marital status on more absenteeism;
2. Marital status has a causal effect on absenteeism via more housework.

Definition 1.2.3 (d-Separation)

A path p is said to be d-separated (or blocked) by a set of nodes Z if and only if

1. p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in Z, or
2. p contains an inverted fork (or collider) $i \rightarrow m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

"Marital status" and "more absenteeism" are marginally associated but become conditionally independent when "the amount of housework" is held constant. That is, by conditioning on "the amount of housework," "getting married" has no direct effect on "more absenteeism." Figuratively, the intervening variable blocks the flow of information along the path (Pearl [2000] 2009, second edition, 17), knowing the amount of housework renders knowing marital status irrelevant to absenteeism.

Wermuth labels the following chain as exemplifying spurious dependence:

successful job placement ← field of qualification ← gender

Here, "field of qualification" blocks every path from "gender" to "successful job placement." She interprets this model as implying that "field of qualification" causes "successful job placement," and that "gender" does not cause "successful job placement" via its effect on "field of qualification." Although Pearl interprets this model as showing that gender causes successful job placement indirectly through field of qualification (personal communication, 4/2/2010), he implies that legal cases support Wermuth's interpretation ([2000] 2009, second edition, 127):

> Another class of examples involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

Whether such antecedent x variables indirectly cause the response y via their effects on intervening variables t requires further study. For example, focus on $t \rightarrow y$ and this relationship is found not to be spurious when antecedent propensity scores or test factors x_i are controlled. Then, the effect of an extra x that affects y and t is controlled, and $x_{(\text{extra})} \rightarrow t \rightarrow y$. What is gained? The $t \rightarrow y$ relationship is still not spurious, t is a candidate cause of y. Additionally, t interprets the effect of $x_{(\text{extra})}$ on y.

[2] Lazarsfeld's statement suggests the following path-analytic definitions of causal effects: Let x be prior to y and t be either prior to or on equal footing with x. Then, if $\beta_{yx.t}$ (the path coefficient) is not equal to zero and t has no direct effect on y, then by marginalizing over t, the bivariate relationship r_{yx} would express the total effect of x on y. However, the path-analytic decomposition $r_{yx} = \beta_{yx.t} + r_{xt}\,\beta_{yt.x}$ (i.e., direct effect of x + shared effect with t) suggests that the estimated causal effect of $x = r_{yx} - r_{xt}\,\beta_{yt.x} = \beta_{yx.t}$, which is the direct effect of x on y. However, if x is prior to t and $x \rightarrow t$, then the putative causal effect of x on y is the correlation: $r_{yx} = \beta_{yx.t} + \beta_{tx}\,\beta_{yt.x}$ (i.e., direct effect + indirect effect of x through t). The size of the path-analytic causal effect is affected by the type of arrow linking x and t in the path diagram, even though the size of $r_{xt} = \beta_{tx}$. As Pearl

stresses (Pearl 2002, 208) and exemplifies (Pearl [2000] 2009, second edition, 154–157), a graph of the assumed relationships paired with parameter estimates can clarify notions of causality.

[3] Pearl (2002, 208) defines a probabilistic cause in the context of a probability distribution and a directed acyclic graph (i.e., a model) as follows: "X is a probabilistic cause of variable Y if $P\ (y \mid do\ (x)) \neq \mathrm{P}\ (y)$ for some values x and y." The agent-based model of Smith (2010b) can be used to clarify this notion of causality in the framework of the model. Let the response variable Y be a measure of the nastiness of the agents. A number of prior variables such as cost-benefit distributions, prejudice, sociometric ties, and so forth are set in advance as parameters of a Monte Carlo run of 20 iterations through the model. These parameters define a context S (for set in advance). Additionally, each agent can have his own unique measure of legitimacy that he attributes to authority; these may range from low to high, say from 0.6 to 1.2. Then, the "naturally occurring" distribution of Y, the output distribution, is referred to as $P(Y \mid S)$; that is, the naturally occurring probability distribution of Y given the parameters S. Next, a second run of the model is implemented, holding constant the parameters composing S. However, in this run the legitimacy of authority parameter is not allowed to vary naturally; it is set by the experimenter so that each agent now has the value $x = 1.3$. The model now creates a new output probability distribution. This distribution is referred to as $\mathrm{P(y \mid do(x) \cup S)}$. The causal effect of the intervention that changed the amount of legitimacy of authority is the difference between the results of the otherwise identical two runs of the model: the causal effect of the intervention is $\delta = P(y \mid do\ (x) \cup S) - P(y \mid S)$.

Given this model and the settings of the runs, this expression is totally consistent with Rubin's potential outcomes view of causality. The agents in both runs are initially identical. Prior to assignment to one run or the other, the agents have potential outcomes if assigned to either run. After assignment and the running of the Monte Carlo trials, the agents have realized outcomes under the intervention (i.e., the treatment) and under the "naturally occurring" null treatment. The fundamental problem of causal inference is not applicable here because the agents in the trials are identical and fully observable; the only difference between them is random variation due to the random number generator that assigns benefits and costs, and the experimental intervention. Thus, the average causal effect of the intervention is $\delta = P(y \mid do\ (x) \cup S) - P(y \mid S)$. The model produces scores for the variables for each agent after each iteration. Consequently, at the level of the individual agent, the agent's nastiness at a point in time in one trial could be compared to his nastiness at the same point in time in the other trial. The individual-level causal effect of the intervention at a point in time for agent i can be defined as $\delta_i = y_i(1) - y_i(0)$, with y indicating the agent's response, 1 indicating the intervention group, and 0 the "naturally occurring" null group (Morgan and Winship 2007, 33). Average causal effects would result by averaging such differences for individuals across the agents at various points in time.

Chapter 2
Contextual Analysis and Multilevel Models

> *Keeping still to the over-all statistical picture, the causes of apprehension were then considered. To organize our material we followed the well-established idea that all human experiences are determined by two broad groups of elements: the characteristics of the people themselves and those of the environment in which they live and work.*
>
> —Paul F. Lazarsfeld and Wagner Thielens, Jr. (1958, 159–160)

The political sociology of higher education is enjoying a renaissance in the United States, but there are few if any contextual analyses. The findings reported by Gross and Simmons (2007) in their comprehensive review of contemporary research may be paraphrased roughly as follows: Numerous recent studies focus on political differences among professors and on how the attitudes of academics differ from those of the public at large. Some on the right believe that the left has captured academia and that instruction is consequently biased. Others question these assertions, believing that academics exhibit a variety of political beliefs and that many professors do not express their personal politics in the classroom. Given the research aim of either documenting or debunking such assertions, the typical study follows the logic of a public opinion survey: It samples individual professors and compares their responses to those of the general public; the effects of institutional and departmental contexts on political beliefs are not examined in depth. Some studies exhibit methodological flaws of sampling design, questionnaire construction, measurement, data analysis, and interpretation; most often, the investigators do not examine closely the influence processes and social mechanisms that shape faculty opinion.

Blumer (1956) critiqued such survey research studies as these for ignoring the effects of context and social networks and for merely describing relationships among psychological variables that link one attribute of a respondent to other attributes. Taking up Blumer's challenge, Lazarsfeld and Thielens (1958) developed a paradigm for contextual analysis that culminated in *The Academic Mind.* Probing the pressing question of the effects of McCarthyism on academia, the investigators asked: How did the climate of fear—generated globally by the cold war against communism and manifested locally on college campuses by attacks on the character of individual teachers because of their alleged political beliefs—affect colleges and universities

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_2, © Springer Science+Business Media B.V. 2011

and their social scientists? The following explication of the logic of their study aims to underscore the importance of contextual analysis for contemporary sociological studies, tighten the linkage between contextual analysis and multilevel statistical modeling, and guide future research on "political correctness" in contemporary academia.

The Academic Mind

Lazarsfeld and Thielens hypothesized that variables at the level of the individual teacher (level-1) and at the level of the academic institution (level-2), along with their cross-level interactions, affected the key outcome variable, namely, a teacher's apprehension about being singled-out and punished for expressing his or her political beliefs. Because they thought that apprehension was jointly determined by variables at the level of the academic institution and by variables at the level of the teacher, they randomly sampled 165 educational institutions and 2,451 social science teachers within these institutions (1958, 371–377). These data, which are available in data archives for secondary analysis, characterize the population of social scientists and academic institutions during April and May 1955, a period when the effects of McCarthyism were still evident. The investigators used the empirical findings to draw inferences about the population of American academic social scientists during that time period. Because apprehension was the key outcome, they referred to this survey informally as the "teachers' apprehension study."

Apprehension and Its Correlates

The investigators developed measures of apprehension by first conducting detailed exploratory interviews. They asked the interviewees to describe any experiences as a teacher that made them feel uneasy about their academic freedom, induced worry about how their political views could affect their professional advancement, or made them cautious about expressing potentially controversial thoughts. On the basis of these detailed interviews, the investigators then created questionnaire items that assessed worry and caution, the two dimensions they had defined for apprehension. Below, marked with "(A)" (for apprehension), are the statements that formed the final six-item apprehension index.

To assess worry, the questions asked (1958, 76), Have you ever worried or wondered that: (A) students could pass on warped views of what you have said leading to false ideas about your political beliefs; (A) potential employers might ask others about your political biases in teaching; (A) there may be gossip in the community about you because of your politics; (A) expression of your political opinions could jeopardize your promotion or job security; your political beliefs could make you unpopular with the alumni; and the college administration may keep a political file or dossier on you and on other faculty members.

Whereas the questions about worry directly tapped the teacher's state of mind, the questions about caution ascertained whether or not the teacher had acted in ways that would prevent potential controversies about his or her political beliefs. These questions essentially asked (1958, 78), Have you at least occasionally: made statements or told anecdotes that made it very clear that you are neither an extreme leftist or rightist; refrained from participating in some political activity so as not to embarrass the trustees or the administration; refrained from discussing political topics with colleagues in order not to embarrass them; (A) not recommended readings that could lead to criticism that these were too controversial; and (A) toned down your writing because you worried that they might cause controversy.

The worry and caution indexes are strongly correlated—the ordinal measure of association gamma $(\gamma) = 0.70$.[1] Nonetheless, the investigators chose the six items that most directly indicate the underlying sentiment of occupational apprehension, the dimension along which the respondents could best be classified. By simply summing the dichotomized replies to the six items, the investigators created an index that ranges from zero (no apprehension) to six (considerable apprehension). They grouped the scores as Low (0, 1) = 54%, Medium (2, 3) = 33%, and High (4–6) = 13%, and, most often, because only countersorters and not computers were then generally available, to simplify the data processing, they dichotomized this variable, analyzing the determinants and consequences of high plus medium = 1 versus low = 0 apprehension.[2]

Having built the index of apprehension, the investigators then turned to validating it by clarifying its relationships with other variables. They found that "vulnerable" teachers, those who had been involved in a personal incident or who were members of a controversial organization, were more likely to exhibit a high degree of apprehension, with personal incidents having the stronger average effect.[3] Other correlations with apprehension were as follows: the higher the teachers' levels of apprehension (at least through scores 0 through 4), the more likely the teachers were to protest bans on controversial speakers and on debates about admission of communist China to the United Nations; to read left-of-center journals like *The Nation, The New Republic*, and the now defunct *The Reporter*; and to be alert to issues of academic freedom and civil liberties (1958, 92–112). At each level of vulnerability, teachers who were more concerned about civil liberties were also more likely to exhibit high apprehension (1958, Fig. 4-8, 109).

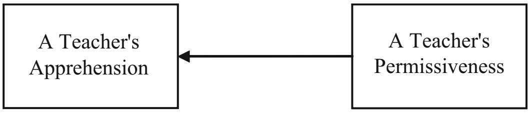
The investigators related a teacher's apprehension to the college context by forming a typology of academic institutions based on size (the number of students) and type of organization (private, public, teachers, colleges, Protestant, and Catholic). The nine types are private (large and small), public (very large, large, small), teachers' colleges, Protestant, Catholic (large and small). The investigators found that teachers at small Catholic institutions and small public institutions were less apprehensive than those at the other types of institutions (see their Fig. 3-8, 90). Attacks were more frequent at institutions of higher quality, but a protective administration could reduce its faculty's amount of apprehension induced by such attacks (1958, 167–174). At the same time, a teacher's breadth of permissiveness (rather than conservatism) increased apprehension.

Assessing Permissiveness

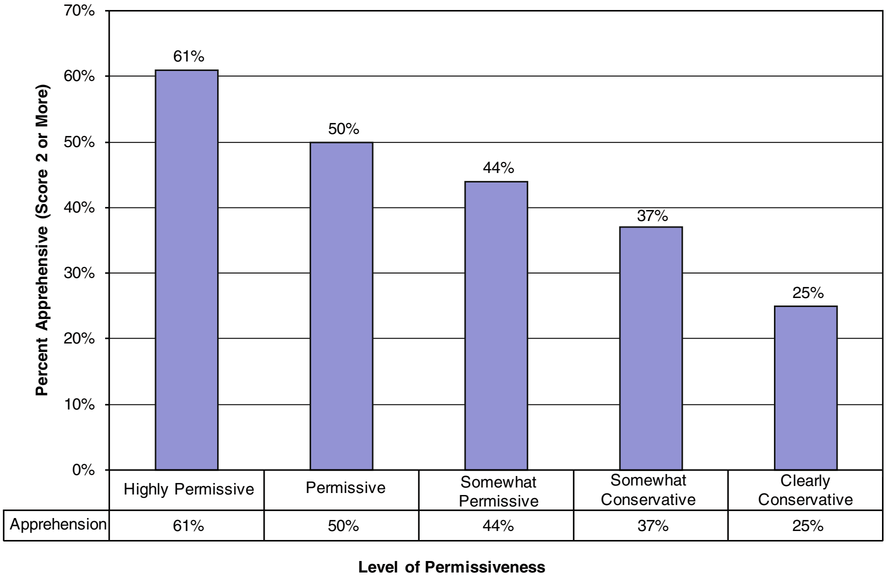
A professor's apprehension was influenced by the scope of his or her permissiveness. Permissive professors were more likely to permit freedom of expression for leftist political views on campus, whereas the more conservative professors were less likely to do so. Other indicators of a permissive outlook were that a professor would not fire a faculty member who admittedly is a communist and would allow the young Communist League on campus. Indicators of a conservative outlook were that a professor: Would not allow Owen Lattimore to speak on campus (Lattimore was an expert on China whom Senator Joseph McCarthy accused of being a spy for the Soviet Union); would fire a store clerk who admitted to being a Communist; would not allow the Young Socialist League on campus; and considered it a luxury to have a radical teacher on the faculty.

From these indicators, the investigators created a bipolar index of permissiveness versus conservatism (see their pages 125–127 for the details). They classified teachers as highly permissive if they gave two permissive replies and no conservative replies. At the other extreme, they classified teachers as highly conservative if they gave two or more conservative replies and no permissive replies. The other teachers exhibited different combinations so that scores on the permissiveness index ranged from: 0 = clearly conservative (14%); 1 = somewhat conservative (14%); 2 = somewhat permissive (29%); 3 = quite permissive (21%); and 4 = highly permissive (22%).[4] Creating a trichotomous index, the investigators combined the two categories at either extreme.

The investigators clarified the meaning of permissiveness by relating its index to a number of indicators that distinguished the political left from the political right; these measures were similar to those they used to validate their apprehension index. Compared with their conservative counterparts, permissive teachers were more likely to vote Democratic, read liberal magazines, belong to professional and controversial organizations, favor classroom discussion of political topics, support academic freedom, be professionally productive, say their own academic freedom had been threatened, say they had been reported to higher authorities, and acknowledge they felt pressures to conform politically (1958: 132–156). Not surprisingly, they also were more likely to exhibit apprehension, which was in part a consequence of their permissiveness. Because the investigators conceptualized permissiveness versus conservatism as a basic attitudinal predisposition, they viewed permissiveness as a predetermining variable that induced apprehension—worry and caution—rather than assuming the opposite.[5] Thus:



The investigators depicted these relationships in a bar chart (Fig. 6-15) similar to that in Fig. 2.1; both charts clearly show that apprehension increases with increases



	Highly Permissive	Permissive	Somewhat Permissive	Somewhat Conservative	Clearly Conservative
Apprehension	61%	50%	44%	37%	25%

Fig. 2.1 Permissive teachers are more apprehensive than conservative teachers

in permissiveness. These variables are measured at level-1, but *The Academic Mind* also reports the effects of incidents, a level-2 contextual variable that describes the professors by a property of their academic institutions.

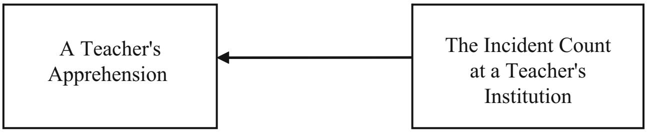
Institutional Incidents

The investigators clarified their use of the word "incident" as follows (44):

> it describes an episode, long or short, in which an attack, accusation, or criticism was made against a teacher, a group of teachers, or a school as a whole. ... This [overt] act might be a listing of the names of supposedly "pink" professors in the gossip column of a local newspaper, a student going to a dean with a charge against a teacher, or a teacher reporting that another man had been passed over for promotion because of his politics.

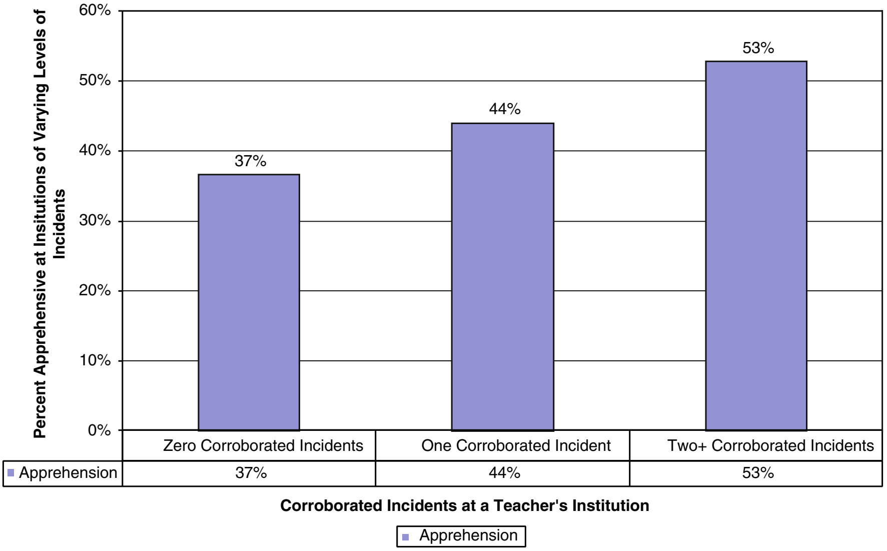
To interpret how the institutions affected their faculties' apprehension, the investigators applied an "attack and defense" model—a strong defense can mitigate the effects of an attack. At the time of the study, right-wing attacks on teachers and their institutions induced apprehension, but if the institution's administration defended its faculty from these incidents, this defense alleviated the apprehension. The investigators measured such incidents by relying on reports from their interviewees. They distinguished corroborated from uncorroborated reports and deleted from their contextual analysis those teachers who had personally experienced an attack or a similar incident. They then characterized the institutions by their count of corroborated incidents (1958, 259).

The count of academic freedom incidents that characterized a teacher's institution was crucial in engendering apprehension, as this diagram depicts:



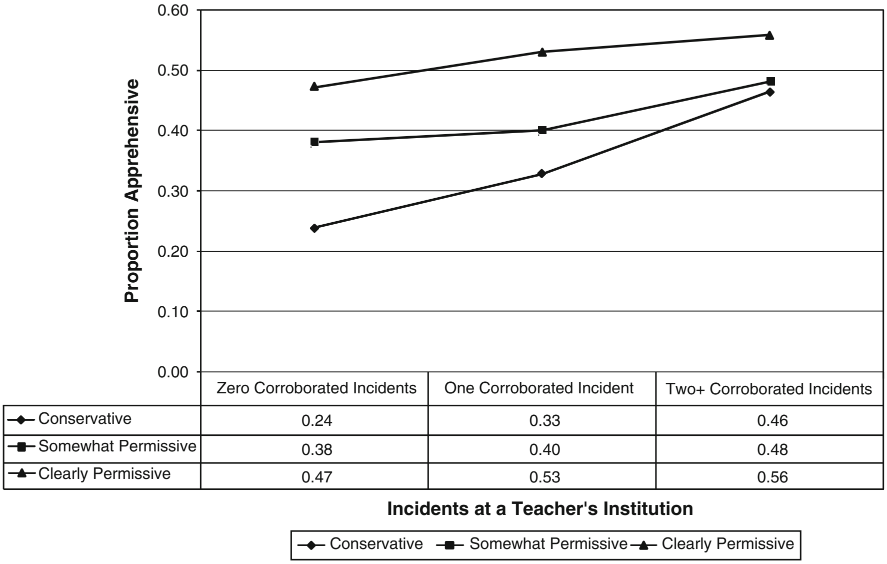
Using data from Fig. 10-9 of *The Academic Mind,* Fig. 2.2 compares the effects of incidents on the apprehension of 1,878 teachers for three types of institutions—those with 0, 1, or 2 or more (hereafter, 2+) corroborated incidents. The bar chart illustrates a *comparative analysis* because it shows how the categories of a level-2 variable shape the extent of a level-1 variable. The teachers included in this analysis had not experienced a personal attack or other political difficulties themselves. Even so, as the number of corroborated incidents at a teacher's institution increases, the intensity of apprehension at that institution also increases. When there are zero corroborated incidents, the baseline (intercept) level of apprehension is 37%. When there is one incident, then the mean level of apprehension increases by 7 percentage points to 44%. When there are 2+ incidents, then the mean level of apprehension increases from the baseline value by 16 percentage points to 53%. Moreover, when the incidents increase from 1 to 2+, the increase in apprehension is 9 percentage points. All of these differences are statistically significant.

Lazarsfeld and Thielens synthesized the microlevel relationship depicted in Fig. 2.1 and the macrocomparative, cross-level relationship of Fig. 2.2 by first



	Zero Corroborated Incidents	One Corroborated Incident	Two+ Corroborated Incidents
■ Apprehension	37%	44%	53%

Fig. 2.2 The greater the number of corroborated incidents at an academic institution, the higher the teachers' apprehension at that institution



	Zero Corroborated Incidents	One Corroborated Incident	Two+ Corroborated Incidents
◆ Conservative	0.24	0.33	0.46
■ Somewhat Permissive	0.38	0.40	0.48
▲ Clearly Permissive	0.47	0.53	0.56

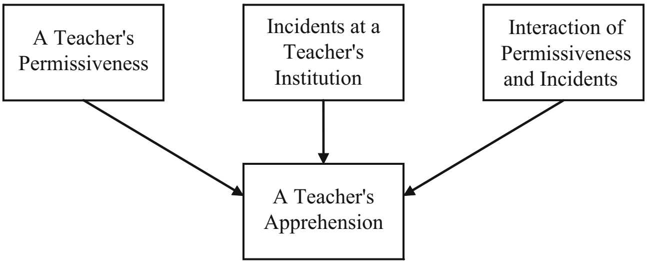
Fig. 2.3 At each level of permissiveness, the greater the number of corroborated incidents at an academic institution, the higher the teachers' apprehension at that institution

sorting the data on the teachers into three groups: clearly permissive (highly and quite permissive), somewhat permissive, and conservative (somewhat and clearly conservative). Then, for each of these groups, they cross-tabulated a teacher's level of apprehension with the number of corroborated incidents at that teacher's institution and they looked for interaction effects.

Figure 2.3 depicts these relationships quantitatively (it uses the data of the investigators' Fig. 10-9 and estimates from SAS's Proc Means, see endnote 6).[6] Holding constant an institution's count of incidents, one set of relationships links the two microlevel variables: the higher the professor's permissiveness, the higher the apprehension. Holding permissiveness constant, the cross-level set of relationships compares the amount of apprehension at institutions with varying incident counts: for each amount of permissiveness, the greater the number of corroborated incidents at a teacher's institution, the greater the amount of apprehension at that institution, even though the teachers included in this figure did not experience attacks or political problems themselves. The institutional context induced apprehension, controlling for the effects of permissiveness.

The effects of context are as follows: the different institutions had different impacts on apprehension, which depended upon the extent of the teachers' permissiveness or conservatism. The apprehension of conservative teachers increased linearly with the increased number of incidents; the increases for the other teachers were less steep. When there were 2+ incidents, then there was very little difference (1 percentage point) between somewhat permissive and conservative teachers.

When there were no incidents, then the difference was much larger (14 percentage points). When the teachers were not predisposed toward apprehension by their basic conservative political orientation, then the institutional context of incidents induced the greater change in apprehension: for conservative teachers, the change was 22 percentage points compared with 10 percentage points for somewhat permissive teachers and 9 percentage points for clearly permissive teachers—context matters! Diagrammatically, assuming that the key determining variables are on equal footing:


A Teacher's Permissiveness
Incidents at a Teacher's Institution
Interaction of Permissiveness and Incidents
A Teacher's Apprehension

Conclusion and Implications

Lazarsfeld and Thielens studied a crucial social problem: how the cold war against communism and radical right-wing extremism engendered a climate of fear at academic institutions; this fear constrained academic freedom and freedom of expression. These investigators identified a causal mechanism for the interplay of stimulus, predisposition, and response: the count of corroborated incidents at an institution (the stimulus) interacted with the teachers' political attitudes, assessed by their leaning toward permissiveness or conservatism (the predisposition), to produce the amounts of apprehension, assessed by worry and caution, about infringements to one's academic and political freedoms (the response). Their pivotal contextual analysis combined (1) a level-1 relationship between permissiveness and apprehension, (2) a comparative relationship between an institution's incident count and its teachers' apprehension, and (3) a contextual interaction effect between permissiveness and incidents as these variables jointly determined apprehension. Their nuanced analysis uncovered the varying effects on apprehension among teachers with different amounts of permissiveness at institutions with different counts of incidents. For conservative teachers, apprehension increased linearly from low values to fairly high values as the number of incidents increased. For the more permissive teachers, the amounts of apprehension were high from the start, even when there were no incidents and, as incidents increased, their apprehension increased but not as much as for conservative teachers.

Notions of Causality

Further developing Lazarsfeld and Thielens's social mechanism view of causality, the next two chapters explicate and provide examples of three basic causal notions; namely, stable association, potential outcomes, and dependency networks. *The Academic Mind* provides elementary examples of all three, as follows: The authors tested the stability of the association between the productivity of a professor (x) and voting Democratic (y) by controlling for three age categories (1958, Fig. 1-7, page 17). Because their control for differences in age (t) did not weaken the $x \rightarrow y$ relationship, this association was thought to be stable.

The potential outcomes perspective can be roughly illustrated by the relationships among apprehension, and incidents (1958, Fig. 10-9, 259; or Fig. 2.3). Prior to assignment to an academic institution that may have 0, 1, or 2+ numbers of incidents (i.e., these are the alternative treatment conditions), the professor has potential outcomes (i.e., different amounts of apprehensions) under each treatment condition. After random assignment to one of these treatments, the professor has a realized outcome under that treatment and two counterfactual outcomes, one for each of the other treatments he has not received. The causal effect of the incidents for this person is the difference between the amount of apprehension under the treatment that he received and the counterfactual amount of apprehension that he would have, had he received one of the other treatments. Since the person can receive only one of these treatments, the individual-level causal effect of the treatment cannot be quantified. This illustrates the fundamental problem of causal inference; after assignment to a specific treatment, only the effects of that one treatment can be observed, the effects of the other treatments cannot be observed and are counterfactuals. However, the average causal effects of the treatments can be estimated by the differences between the average amounts of apprehension under each treatment condition. Referring to Fig. 2.3, for professors actually receiving the zero-incidents treatment, the average proportion apprehensive is 0.37. Then, the average proportion apprehensive if assigned to an institution with one incident would be 0.44. Therefore, the average causal effect of one incident is $0.44 - 0.37 = 0.07$. The average causal effect if assigned to 2+ incidents relative to assignment to 0 incidents would be $0.53 - 0.37 = 0.16$. Of course, this model assumes that the professors are randomly assigned to the treatment conditions; that is, assignments to the treatments are not confounded with prior amounts of permissiveness and apprehension, or other variables. The actual data do not meet these criteria.[7]

Dependency networks include a generative mechanism that links a response y to an x via intervening variables t: $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow y$. Illustrating a dependency network, Lazarsfeld and Thielens's summarizing schematic (1958, Fig. 7-13, 188) has this form: the quality of an academic institution (x) influences the permissiveness of the faculty (t_1), which influences the external pressures (t_2), which influences the performance of the administration (t_3), which in turn influences the apprehension of the professors (y).[8] This qualitative theory linking context, mechanism, and outcome challenges analysts to quantify these relationships by applying multilevel modeling to the data of *The Academic Mind* and to the data of contemporary studies of higher education.[9]

Multilevel Models

Lazarsfeld and Thielens synthesized the level-1 relationship between permissiveness and apprehension and the cross-level relationships by cross-tabulating the three variables. Because of the limitations of their research technology, they could not simultaneously analyze the interrelations among the variables in their summarizing qualitative theory. Using similar cross-tabular procedures, Ladd and Lipset (1976) probed the political polarization of academia induced by the Viet Nam war.

Cross-tabulations may limit the number of variables investigators can consider simultaneously. The development of statistical methods and computing power allowed such contextual analysts as Sewell and Armer (1966) to study comprehensive systems of variables. However, when observations are clustered within units, as they are in multilevel data structures, standard regression procedures are likely to produce erroneous standard errors and confidence intervals; the observations are not independent and the error terms are correlated, thus violating key assumptions of regression analysis. Addressing this problem, Mason, Wong, and Entwistle (1983) developed multilevel modeling and explained their approach to sociological methodologists. The computer programs and explications of multilevel modeling done by Bryk and Raudenbush (1992; Raudenbush and Bryk 2002), Goldstein (1987, 1995, 2003), Littell, Milliken, Stroup, Wolfinger, and Schabenberger ([1996] 2006), and Gelman and Hill (2007), among many others, have made this useful procedure accessible to social and behavioral scientists.

Chapter 5 aims to advance the contextual analysis methods of Lazarsfeld and Thielens by putting forward 11 pivotal uses of multilevel models. The illustrative examples, which are taken from the core chapters of this book, show how multilevel models can advance the earlier method of cross-tabular analysis and the contemporary method of regression analysis. Multilevel modeling can advance contemporary research on pressing social problems because it enables the investigator to study the effects of contexts at one point in time and across time, in systems composed of variables at different levels of the hierarchy of data, for response variables at different levels of measurement.

Acknowledgements The author derived this chapter from *The Academic Mind* revisited: Contextual analysis via multilevel modeling, published in *Applications of mathematics in models, artificial neural networks and arts: Mathematics and society*, ed. Vittorio Capecchi, et al., 163–193; © Springer 2010.

Endnotes

[1] To calculate gamma (γ), first define A as the number of concordant pairs of observations (++ or $--$) and B as the number of discordant pairs of observations ($+-$ or $-+$). Then, gamma $= (A - B)/(A + B)$. For Table 3-3 (1958, 81): $A = 1{,}184 \times 423 = 500{,}832$; $B = 125 \times 719 = 89{,}875$; $A - B = 410{,}957$; $A + B = 590{,}707$; $\gamma = 410{,}957/590{,}707 = 0.70$.

[2] Lazarsfeld and Menzel ([1961] 1972, 229–230) distinguish properties of members in contexts where collectives have been defined as *absolute*, *relational, comparative*, or *contextual*. Apprehension is an absolute property of the professors because its measure was obtained without using information about the academic institution and without taking into account information about social relationships among the professors.

[3] Rough estimates of the average effects of these variables can be calculated by simply taking a simple average of the two conditional relationships. For the data of their Fig. 7-13, the roughly estimated average effect of involvement in an incident on apprehension, controlling for membership in a controversial organization $= ((75\% - 56\%) + (71\% - 36\%))/2 = (19\% + 35\%)/2 = 27\%$. The roughly estimated average effect on apprehension of membership in a controversial organization, controlling for involvement in an incident $= ((75\% - 71\%) + (56\% - 36\%))/2 = (4\% + 20\%)/2 = 12\%$. The roughly estimated interaction effect on apprehension of not being involved in an incident but belonging to a controversial organization is the difference between these differences divided by 2. It equals $(20\% - 4\%)/2 = (35\% - 19\%)/2 = 8\%$. These rough estimates do not take into consideration the different sample sizes and the limitations of the linear probability model.

[4] Similar to apprehension, permissiveness is an absolute property of the professors.

[5] The direction of the effect between permissiveness and apprehension has perplexed some scholars (personal communication). Lazarsfeld and Thielens assumed that permissiveness led to apprehension (see their Fig. 7-13). Implicitly, they may have conceptualized permissiveness-conservatism as roughly analogous to the cluster of variables indicative of authoritarianism. The conservative pole of permissiveness is somewhat analogous to political and economic conservatism, whereas the openness-to-ideas aspect of permissiveness suggests stronger commitments to anti-authoritarian democratic values. If authoritarianism is a variable of personality, then, given this analogy, it is logical to assume that apprehension is in part a manifestation of permissiveness, rather than the opposite. However, both variables could mutually influence each other.

[6] Here is the SAS data set and syntax that relates permissiveness, apprehension, and academic freedom incidents:

```
Data Chapter2;
input incidents permissive apprehension count;
DATALINES;
0  2  1  113
0  2  0  127
0  1  1  108
0  1  0  176
0  0  1  55
0  0  0  175
1  2  1  148
1  2  0  131
1  1  1  84
1  1  0  126
1  0  1  46
1  0  0  94
2  2  1  170
2  2  0  134
2  1  1  65
```

```
2  1  0  70
2  0  1  26
2  0  0  30
;
data chapter2; set chapter2;
Title 'Grouped Data Similar to Figure 2.1.';
proc means N mean;
var apprehension;
class permissive;
freq count;
Run;
Title 'Data for Figure 2.2.';
proc means N mean;
var apprehension;
class incidents;
freq count;
Run;
Title 'Data for Figure 2.3.';
proc means N mean;
var apprehension;
class permissive incidents;
freq count;
Run;
```

[7] Smith (1985a, 68–79) formalizes the key findings of *The Academic Mind* in terms of a generative computer simulation model and a mathematical analysis of some of its implications. The programmed model can be used to conduct Monte Carlo experiments that conform to the potential outcomes causal perspective.

[8] Pawson (1989, 2006) and Pawson and Tilley (1997) advocate a realistic social science that link contexts, mechanisms (i.e., generative processes), and outcomes. Angus Deaton (2009, 4) agrees with this perspective, writing:

> I concur with the general message in Pawson and Tilley (1997), who argue that 30 years of project evaluation in sociology, education and criminology was largely unsuccessful because it focused on *whether* projects work instead of on *why* they work. In economics, warnings along the same lines have been repeatedly given by James Heckman. . . .

[9] Smith (2010a) applies generalized linear mixed models to quantify further the data above in endnote 6; that is, the data from Lazarsfeld and Thielens's Fig. 10-9.

Chapter 3
Stable Association and Potential Outcomes

> *Causality is a very intuitive notion that is difficult to make precise without lapsing into tautology. Two ingredients are central to any definition: (1) a set of possible outcomes (counterfactuals [i.e., hypotheticals]) generated by a function of a set of "factors" or "determinants" and (2) a manipulation where one (or more) of the "factors" or "determinants" is changed. . . . Holding all factors save one at a constant level, the change in the outcome associated with manipulation of the varied factor is called a causal effect of the manipulated factor.*
>
> —James J. Heckman (2005a, 1)

The literature on causality is wide-ranging, difficult at times, and controversial.[1] Rather than delving into all of the nuances of this material, this chapter and the next have a much more modest goal. Through the analysis of examples, these chapters aim to sensitize the reader to various conceptions of causality discussed by social and statistical scientists. With this overview in mind, the reader of this book will be better able to critique assertions about causal effects in the subsequent chapters and in other research reports. By reading the cited material, the reader can gain a more detailed understanding of causality.

These chapters build on Cox and Wermuth's notions of level-zero, level-one, and level-two causality in quantitative research (1996, 219–228; 2001, 65–70; 2004, 287; Goldthorpe 1998; 2007, 191); slightly reconceptualizing their notions and presenting new examples.[2] Although these authors crisply define their notions of causality, the analyses in this present book do not crisply conform to their criteria; at best there are areas of ambiguity about the causal aspects of the findings. Consequently, for critiquing the results of these chapters and of the social sciences in general, because of the ambiguity due to their imprecision, these chapters distinguish these three broad zones of causal notions: *Stable association* (i.e., level-zero causality) focuses on testing a relationship between two variables for spuriousness; it includes causality in classic survey research and robust dependence. *Potential outcomes* (i.e., level-one causality) focuses primarily on causality in experiments and observational studies (Rubin 2003); here it includes causality as an effect of an intervention and causal models for basic attributes. *Dependency networks* (i.e., level-two causality) focuses on systems of interrelated variables; here it includes graphical models, association

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_3, © Springer Science+Business Media B.V. 2011

graphs for loglinear models, causality as a generative process, and causality in policy research (i.e., “all causes” structural economic models). This chapter focuses on the first two zones; the next chapter, on dependency networks. These zones and their illustrative cases can provide a useful perspective and common vocabulary for discussing contemporary meanings of causality in social research and for locating the zone of causality of the findings in the various chapters.

Stable Association

Analysts of stable association aim to determine if a relationship between a response variable and an explanatory variable is robust, when a number of test variables are controlled. If a control for a test factor that is prior to both the explanatory variable and the response variable severely weakens their relationship, then that association is spurious and the test factor is thought to determine the explanatory and response variables. However, if a wide range of test factors on equal footing with x or prior to x and y fail to severely weaken the original relationship, then the explanatory variable may cause the response and the original relationship is thought to be genuine rather than spurious; (Simon 1957, 37–49) codified notions of spurious and genuine correlation.[3] Causality in classic survey research and robust dependence (e.g., time series analysis and meta-analysis) exhibit this pattern of testing relationships for stable association.

Classic Causality

Armed with Hollerith punched cards and counter-sorters, Lazarsfeld (1955a, 1955b) and his colleagues at the Bureau of Applied Social Research (Kendall and Lazarsfeld 1950), along with researchers like Samuel Stouffer at other centers for survey research, closely inspected qualitatively a relationship between a response variable (y) and a stimulus variable (x), and then controlled for the effects of a range of test factors (t) usually one at a time. If a test factor was prior in time or structure to x, and if the $x \rightarrow y$ relationship disappeared in the partial tables when t was controlled (i.e., conditioned on t), then the original relationship was thought to be spurious: $t \rightarrow x$ and $t \rightarrow y$.[4] If t intervened between x and y and the $yx.t$ relationship disappeared in the partial tables, then t interpreted that relationship: $x \rightarrow t \rightarrow y$. This heuristic procedure of elaboration also enabled analysts to assess interaction effects in tables in which one value of t increased the $yx.t$ relationship in one partial table and another value of t reduced it in another partial table, and to develop complex chains like this one: $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow y$ (Hyman 1955, 287, 325–326; Glock 1967, 34–37; Morgan and Winship 2007, 224–227).[5]

Lazarsfeld and Thielens’s (1958) contextual analysis, which the previous chapter reviewed, exemplifies this general approach. In their sample of professors and colleges, the investigators tested the original observed relationship between two pivotal level-1 variables—apprehension (y) and permissiveness (x)—by controlling for the effects of academic freedom incidents (t), a variable at level-2. Because the $x \rightarrow y$ relationship still held when t was controlled (i.e., the $yx.t$ relationship did not disappear), the incidents at a professor's college did not explain the relationship between his permissiveness and his apprehension. Rather, the three variables formed this generative mechanism: incidents (a stimulus), permissiveness (a predisposition), and their cross-level interaction (the combination of stimulus and predisposition) all influenced the level of apprehension (a response). Even so, the investigators' qualitative analysis by close inspection of a few cross-tabulated variables limited their ability to make more appropriate causal inferences. They were unable to control simultaneously for a range of predetermining test variables, and they did not quantify the effects of the variables. Subsequently, such empirically minded causal modelers as Simon (1957), Blalock (1964), Coleman (1964), Rubin (1974), Goodman (1978), Davis (1971, 1975, 1980), Duncan (1975), Jöreskog and Sörbom (1979), Mason et al. (1983), and other social statisticians advanced Lazarsfeld and Thielens's methods.

Causality as Robust Dependence

By quantifying the asymmetric $x \rightarrow y$ relationship while controlling simultaneously for the effects of the test factors, the logic of causality as robust dependence (i.e., level-zero causality) advances that of classic causality. Temporal ordering is a crucial prerequisite; the candidate causal variable x must precede the effect y (Kenny 1979, 2–4; Suppes 1970; Granger 1986, 967–968). Ideally, many replicated studies apply statistical methods to quantify the $x \rightarrow y$ relationship, controlling simultaneously for the variables in a set T that includes a wide range of variables on equal footing with, or antecedent to, x: "This is a statistical association, i.e., nonindependence, with clearly established ordering from cause to response and which cannot be removed by conditioning on *allowable* alternative features" (Cox and Wermuth 2004, 287). The *allowable* controls do not intervene between x and y, are not determined by y, and are sufficient to remove any bias in the relationship between x and y. As explicated by Berk (1988, 158), Suppes's notion of causality is very similar to that of robust dependence. "[Suppes requires that] causes precede effects in time. The prima facie cause C is a genuine cause of E if it is not a spurious cause." Statistical controls for appropriate variables do not make the $C \rightarrow E$ relationship disappear. Granger's causality in time series analysis has a similar structure.

Time series analysis

For inferring causality in time series, Holland (1986, 957) shows that Granger's causality is similar to Suppes's causality:

> In Granger's time series setting, the value of Y is determined at some time point s, and the values of X and Z are determined at or prior to some other time point $\mathrm{r} < s$. I will say that X is not a Granger cause of Y (relative to the information in Z) if X and Y are conditionally independent given Z. Thus X is a Granger cause of Y if different values of X lead to different predictive distributions of Y given both X and the information in Z, that is, if X helps predict Y even when Z is taken into consideration. … Granger noncausality is very much like Suppes's notion of a spurious cause. Both involve the inability of the spurious cause to predict a future event or value given certain other information.

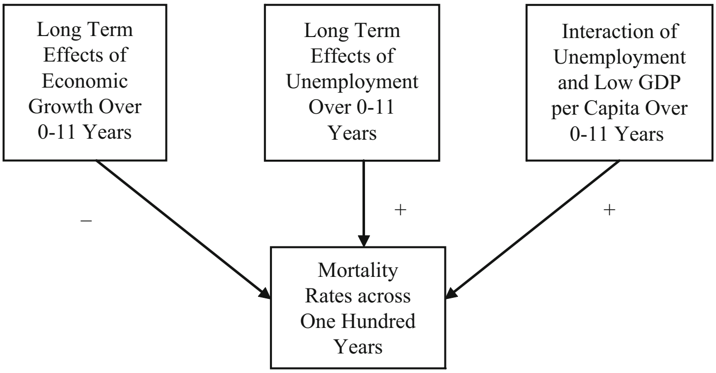
Granger causality is exemplified by Brenner's many fine macrolevel time series analyses of the effects of fluctuations in the economy on indicators of mental health, mortality, and other health-related social variables (1971, 1973, 1989a, 1989b, 2005). Because of contractions in the economy due to recessions and depressions, after a period of time greater than a year or so, the measures on aggregate indicators of malaise and mortality increase, and these relationships hold most strongly for vulnerable people, i.e., those who lose the most in terms of economic and social status and are not protected by social safety nets.

Brenner's recent analysis (2005) of the relationship between economic growth and mortality decline in the United States from 1901 to 2000 exemplifies his research paradigm, which he has developed across his 40-years' of cumulative research on how changes in the economy affect well-being and other social and epidemiological variables. Based on his extensive subject-matter knowledge, and after reviewing the literature on this topic, in his article Brenner plots the inverse relationship between age-adjusted death rates (logarithm rate per 100,000) and real gross domestic product (GDP) per capita (2005, Fig. 3.1, 1215). He finds that sustained economic growth reduces the mortality rate, perhaps because of improvements in nutrition, sanitary engineering, and better housing. Moreover, for industrialized countries at the national level, numerous studies report that unemployment is a significant predictor of higher mortality rates over periods of at least a decade (2005, 1215).

Brenner hypothesizes that short-term and long-term economic growth have beneficial consequences for life expectancy, but the very early phases of growth, that is, within the first year, may have negative consequences. He and other researchers (Neumayer 2005) have offered various reasons for this negative effect, but the absence of microlevel data limits their interpretations, which can be summarized as follows: because of anomie, rapid economic growth and recession may produce mental distress; during the period of rapid growth, innovations at work may induce malaise (e.g., Chapter 9 studies the effects of computer innovations on worker malaise); expansion of the economy may increase the intensity of work-related stress; and prosperity may lead to harmful life-styles—elevated consumption of alcohol, tobacco, calories, and so forth.

Putting these interpretations aside, Brenner focuses primarily on intermediate and long-term consequences, positing that economic growth reduces mortality rates, unemployment for greater than six months increases mortality, and the joint effects of low GDP and high unemployment exacerbate the mortality rates.

Diagrammatically:


Long Term Effects of Economic Growth Over 0-11 Years
Long Term Effects of Unemployment Over 0-11 Years
Interaction of Unemployment and Low GDP per Capita Over 0-11 Years
–
+
+
Mortality Rates across One Hundred Years

By controlling for the effects of wars, Brenner tested whether the effect of the economy on mortality was spurious; he included in his model indicator variables for 1918 and 1945, the years with the largest number of military personnel on active duty. He estimated his model using appropriate econometric techniques and compared the model-based results with the actual data; the fit was very close (Fig. 3.3 and Fig. 3.4, 1219). He found that long-term growth in GDP per capita decreased mortality rates whereas increased unemployment increased mortality rates, as did the interaction of low GDP and high unemployment. However, in the very short term, within the first few months of a decade of data, rapid economic growth is associated with increased mortality, perhaps due to the need for employees to adapt to innovative technologies in combination with the increased intensity of work. Similarly, in the very short run, increased employment is associated with increased mortality, perhaps due to the stresses or reemployment and new jobs. But these short-term effects soon disappear.

Thus, the cumulative effects of increased economic growth and reduced unemployment cause reductions in the mortality rates across 100 years of data for the United States. Because the effect of x on y, controlling for a set of control variables T (i.e., $yx.T$), is best assessed in numerous studies in different settings, a meta-analysis of such findings could enhance estimates of the robustness of the dependence and the stability of the association.

Meta-analysis

A meta-analysis is a statistical study that combines information from several other studies. Meta-analyses implicitly aim to determine causality through robust dependence. Examples in the social sciences include summaries of the effects of desegregation of education on the academic achievement of black students (Crain and Mahard 1983), comprehensive school reform on test scores (Borman et al. 2003), and negative political campaigns on voter participation (Lau et al. 1999; Ansolabehere et al. 1999); the dependencies exist but they are not very robust.

Chapter 14 of this book presents several meta-analytic consolidations; one examines the association between precertification of inpatient medical care and subsequent admissions to hospitals, across eight evaluations that controlled for various sets of test factors. These studies were conducted during the period 1986–1990. Because of the changes in the insured populations under this form of managed care, the results should not be generalized beyond that time period, but they do illustrate the method. Seven of the eight studies report reductions in admissions, but four of these include an upper confidence bound greater than zero. When meta-analytic techniques summarized these data, the fixed-effects estimate of the reduction was −7.6 admissions per 1,000 enrollees (EEs) with a confidence interval from −4.9 to –10.3. Because one study had outlying results, the fixed-effects model for these studies did not hold; it was necessary to calculate the random-effects estimates, which produced a smaller effect bounded by a wider confidence interval: −7.3 hospital admissions per 1,000 EEs (−0.8, –13.8).

The medical statistician Austin Bradford Hill (1965, 295–300) systematized considerations for assessing robust dependence, that can be summarized as follows (also see Cox 1992, 292; Cox and Wermuth 2001, 70; Greenland 2004, 9–10, and the section on "Methods for Reviewing Evidence" in Chapter 15 of this book): the $x \rightarrow y$ relationship is consistent with theory or observations; the assumed cause x precedes the response y in time; the $yx.T$ relationship is strong and not spurious—the stronger the association, the more likely the relationship is causal; the conditioning set T includes test factors on equal footing with x and those that potentially cause both x and y; the response y varies with monotonic changes in x (e.g., the greater the number of cigarettes smoked per day, the greater the risk of lung cancer); the $x \rightarrow y$ relationship is found in numerous studies in different settings using different methods; alternative explanations have been eliminated; and experiments or quasi-experiments confirm the $x \rightarrow y$ relationship.[6]

Potential Outcomes

For a specific unit (e.g., a person), Donald Rubin's (1974) notion of causality defines the causal effect on y of the experimental treatment (E) compared with the control treatment (C) as the difference between two potential outcomes, namely: the synchronic difference between the realized outcome when that unit receives the experimental treatment and the counterfactual outcome if that unit would have received the control treatment; that is, $y(\mathrm{E}) - y(\mathrm{C})$. This comparison of the two outcomes for the same unit when exposed to the experimental and control treatments is the focus of interest; the emphasis is on understanding the *effects of a cause* (Holland 1986, 945). The unit may be a person, a plot of land, an organization, a country, and so forth. Empirically, the unit only receives one treatment, say the experimental treatment, and its realized outcome $y(\mathrm{E})$ under this treatment is known; thus, the counterfactual outcome for that unit if it would receive the control treatment y(C) is unknown. Similarly, if that unit received the control treatment, then its realized outcome under this treatment is known; thus, the counterfactual outcome for that unit if it would have received the experimental treatment is unknown.[7] Information about that unit's outcome under the treatment that it did not receive is missing. Consequently, because of these missing data, the causal effect of the experimental treatment for this unit cannot be calculated directly. However, an average causal effect between the experimental and control units can be estimated and this effect may be very similar to the individual-unit causal effect if the units in the two groups are very closely matched and their assignment to either of the two groups is not confounded by prior differences between them. Informed by this incomplete-data, causal perspective of Rubin and his colleagues, this section discusses causal inferences in observational studies. It then explores some interrelationships between Rubin's causal model and that of James Coleman. Rubin's model appears strained when applied to assess the causal effects of an intrinsic characteristic of a unit. For example, with the unit's gender defining the treatment conditions, the response (R_1) when the unit is male ($C = 1$) compared with the response (R_0) when that unit would be female ($C = 0$) is not only counterfactual but difficult to conceptualize.[8] Coleman developed his model to estimate the causal effects of such intrinsic characteristics.

Causality as an Effect of an Intervention

Cox and Wermuth (2004, 289–290) refer to Rubin's (1974, 693–688; 2006) potential outcomes conception as exemplifying level-one causality, which they conceptualize as intermediate between stable association (level-zero) and their (1996, 2001, 2004) dependency modeling approach (level-two). Rubin's model is especially appropriate for guiding the assessment of implied causal effects in observational studies, in which some units receive the experimental treatment and others the control, but random assignment is absent; the evaluations of educational reform in Part 3 exemplify such studies, as do the evaluations of healthcare change agents in Chapter 14.

As explicated by Holland (1986), Rubin's elementary model of causality assumes that a person (u) in the population of universe U cannot simultaneously receive the experimental treatment (t) and the control treatment (c) so that it is impossible to observe the outcomes (Y) of the treatment effect $= Y_t(u) - Y_c(u)$. This fact is referred to as the fundamental problem of causal inference. But, as Holland (1986, 947) states, information bearing on the average causal effect for the universe of observations can be obtained:

> The *average causal effect*, T, of t (relative to c) over U is the expected value of the difference $Y_t(u) - Y_c(u)$ over the u's in U; that is
>
> $$E(Y_t - Y_c) = T. \tag{1}$$

> *T* defined in (1) is the average causal effect. By the usual rules of probability (1) may also be expressed as
>
> $$T = E(Y_t) - E(Y_c). \tag{2}$$
>
> ... (2) reveals that information on different units that can be observed can be used to gain knowledge about *T*. For example, if some units are exposed to *t*, they may be used to give information about *E*(Y_t) (because this is the mean value of Y_t over *U*), and if other units are exposed to *c* they may be used to give information about *E*(Y_c). Formula (2) is thus used to gain knowledge about *T*. ... [This] statistical solution replaces the impossible-to-observe causal effect of *t* on a specific unit with the possible-to-estimate average causal effect of *t* over a population of units.

Cox and Wermuth (2004, 289) summarize this notion of causality succinctly:

> [F]or each individual there are two notional responses R_1 and R_0 depending upon whether [treatment] C_1 or C_0 is used. Only one of these notional responses can be observed and the other thus is in principle not observable and therefore called a counterfactual. This formulation is combined with an assumption that any difference between R_1 and R_0 is systematic, in an extreme for that $R_1 - R_0 = \Delta$, a constant, i.e., the same for all individuals in the study.[9]

Because an individual cannot simultaneously receive both treatments, the causal effect is estimated as the average difference in response between closely matched units in the treatment and control groups, or between units assigned at random to these groups. (Subsequently, the chapters in this book denote conceptually the average causal effect by δ, rather than by *T* or Δ.)

Rubin's notion of causality has received numerous laudatory and detailed explications including those by social scientists Berk (1988, 158–163), Sobel (1995, 1996, 2005), Winship and Sobel (2004), Morgan and Winship (2007), Gelman and Hill (2007, 167–233), and Imbens and Wooldridge (2009). The latter (2009, 9) opine that Rubin's distinction between the realized outcome and the pair of potential outcomes "is the hallmark of modern statistical and econometric analyses of treatment effects."[10]

To ensure that treatment and control groups are composed of equivalent units (e.g., people), thereby reducing bias due to spuriousness, some researchers match the units on the basis of their propensity score, which is the predicted probability that a unit receives the experimental treatment. A unit's predicted probability of being in the treatment group is most often estimated by the logistic or probit regression of a dichotomous indicator of treatment group membership (1 or 0) on a wide range of variables antecedent to assignment to treatment or control. This matching on the estimated probability of treatment group membership essentially controls for those covariates included in the predictive regression equation.

Applications guided by Rubin's model of causality that use propensity scores to reduce spuriousness due to selection bias can be viewed (perhaps oversimplifying) as having a logical structure $yx.T^p$; the superscript *p* denotes the propensity scores. The set T^p includes the propensity scores and covariates that do not intervene between the cause and the response and that are not descendants of *y*. This structure is similar to that of classic causality (*yx.t*) and of robust dependence (*yx.T*), but the inclusion of the propensity scores and, ideally, a manipulated treatment allow stronger causal inferences.

For comprehensive examples of the use of propensity scores see the chapters by Rosenthal and Rubin (in Rubin 2006, 167–231); the nine chapters by different authors on causal inference and observational studies in Gelman and Meng (2004, 1–108); Gelman and Hill's section on propensity scores (2007, 206–212); the chapter on matching in Morgan and Winship (2007, 87–122); and Harding's (2003) application of these methods to survey data in which the treatment is not a manipulated variable.[11]

Rubin's (1974; 1986, 962) causal model is grounded in the logic of experimental studies in which the treatments are manipulated variables; it focuses on measuring the various *effects of a cause*. Coleman's (1964, 1981) causal model and Goodman's log-linear statistics (1978, 1984) for survey data primarily focus on how the various attributes that are not manipulated are the *causes of an effect* (i.e., the response).

Causal Models for Attributes

Contemporary discussions of causality in the potential outcomes tradition ignore Coleman's causal model for the effects on response variables of basic attributes and continuous variables. Addressing this omission, this section reviews his initial model that is based on the linear decomposition of effects, shows how he extended this model using a multiplicative decomposition of effects (i.e., logistic regression), and relates aspects of both models to Rubin's notion of causality. Since Coleman's (1981) logistic regression model and Goodman's logit model (1972a, 1978, 7–25) are very similar if not identical (Magidson 1978, 27–54), Coleman's interpretation of the causal forces also applies to this model of Goodman.

The People's Choice, the seminal study of an election campaign by Lazarsfeld et al. ([1944] 1948, 25), reported that a person's voting choice as Republican or Democratic was strongly predicted by three primary social characteristics, which the investigators conceptualized as being on equal footing: socioeconomic status (higher versus lower), religious affiliation (Protestant versus Catholic), and residence (urban versus rural). Coleman (1964) views such rather fixed characteristics as causes of an event outcome: causes→event outcome. He develops statistical methods for quantifying estimates of the effects of such nonmanipulable causes by following this general approach: "(1) to begin with the idea of a process, (2) to attempt to lay out a mathematical model that mirrors this process, and then (3) given particular kinds of data, to transform the mathematical model into a statistical model for estimating parameters of the process" (Coleman 1981, 5). As explicated next, there are convergences between the causal models of Coleman and Rubin.

Coleman conceptualizes the causal effects of social attributes via the changes they bring about. His model assumes that continuous variables and attributes of a person can causally change the transition rates between the states of a dichotomous response variable. His linear and logistic multivariate models decompose these rates into the effects of the various continuous variables and attributes. His emphasis on the causal effects of attributes differs from Rubin's more experimental, causal perspective. Initially, Rubin (and Holland) stipulated that attributes cannot have causal efficacy: there is no causation without manipulation of the causal variable (Holland 1986, 945–948, 959; Rubin 1974, 689–693; Berk 1988, 166–168).

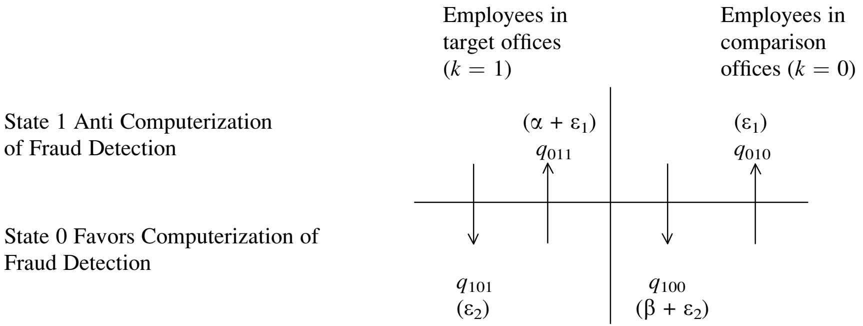
Linear decomposition of effects: one explanatory attribute

These subsections explicate Coleman's model using data from Chapter 9 about the effects of new computer systems on employee attitudes.[12] For an employee, "working in a claims office that has a new computer system" is conceptualized here as a contextual attribute of the employee (Lazarsfeld et al. 1972, 230); that is, the employee is characterized by a property of her office, she works in an office that has a new computer system (coded 1) or she does not (coded 0). Initially, the study design specified that a computerized fraud detector would be installed in three target offices but not in three closely matched control offices; all of the employees would be surveyed. Panel (a) of Table 3.1 presents a cross-tabulation of two attributes of an employee: (1) x, working in a target office ($k = 1$) or working in a control office ($k = 0$); cross-tabulated with (2) y, being against computerized fraud detection (state 1) or favoring computerized fraud detection (state 0), the latter is the attitudinal response measured by the dichotomized "anti-index." The following assumptions govern these data.[13] There are no extraneous variables that:

Table 3.1 The Coleman causal model for the data

(a) The data	Employees in target offices ($k = 1$)	Employees in comparison offices ($k = 0$)
State 1 Anti Computerization of Fraud Detection	$45 = n_{11}$	$19 = n_{10}$
State 0 Not Anti Computerization of Fraud Detection	$32 = n_{01}$	$89 = n_{00}$

(b) The Coleman model


Employees in target offices ($k = 1$)
Employees in comparison offices ($k = 0$)
State 1 Anti Computerization of Fraud Detection
State 0 Favors Computerization of Fraud Detection
$(\alpha + \varepsilon_1)$
q_{011}
(ε_1)
q_{010}
q_{101}
(ε_2)
q_{100}
$(\beta + \varepsilon_2)$

Here $\alpha = 0.41$, $\varepsilon_1 = 0.18$, and $\varepsilon_2 = 0.42$.

(1) jointly determine x and y (the $x \rightarrow y$ relationship is genuine and not spurious); (2) are prior to x and determine the category of x (the propensity score is random); and (3) intervene between x and y (x is the candidate cause). Moreover, (4) allowable extraneous variables that determine y but not x are standardized by their mean (their effects on y appear in the intercept of the regression of y on x). With these assumptions, the data approximate those of a randomized experiment.

Panel (b) of Table 3.1 presents Coleman's linear model for these data; here the transition rate from state i to state j for individuals in category k of the explanatory attribute is denoted as q_{ijk}; the vertical arrows in the diagram portray these causal forces. Each person assigned to a target office would exhibit ambivalence due to the two contradictory forces: q_{011} (up) and q_{101} (down). Similarly, each person assigned to a control office would exhibit ambivalence due to q_{010} (up) and q_{100} (down).[14] The target and control offices are separated by a barrier such that a person cannot work in both sorts of offices at the same time; the fundamental problem of causal inference holds. Thus, this model portrays the potential outcomes for a person if that person who has the attribute "working in a target office" would instead have the attribute "working in a control office," and vice versa. Much like Rubin's causal model, the difference between the potential outcomes defines the causal effect conceptually.[15]

Unlike Rubin's causal model, Coleman's model assumes that an individual's response is not fixed; there is movement from one category of the response variable to the other. Within a category k, the notional meaning of q_{ij} is that in an infinitesimally small period of time, dt, there is a probability $q_{ij}dt$ of moving from state i to j. The rate of reduction in an individual's probability of being in state i (rather than having moved to state j) is the infinitesimal transition probability q_{ij} times the probability that the individual is in state i; specifically, $dp_i = -q_{ij}dtp_i$ (Coleman 1970, 218–219). Within a category k of the explanatory attribute, since there is movement from i to j and from j to i, Eq. (3) expresses the rate of change in p_1; in some expressions (as in the second one below) the initial subscript for the origin state will be dropped:

$$\begin{aligned} dp_1/dt &= -q_{10}p_1 + q_{01}p_0 = -q_{10}p_1 + q_{01}(1 - p_1) \\ &= -q_0p_1 + q_1p_0 = -q_0p_1 + q_1(1 - p_1). \end{aligned} \tag{3}$$

The effects of the different categories of the explanatory attribute change the transition rates. For these data the model assumes that if an employee is working in a target office then this would increase the employee's transition rate q_{01} toward being against computerized fraud detection (in the diagram, α denotes this effect). It also assumes that if an employee would be working in a control office, then this increases the employee's transition rate q_{10} toward favoring computerized fraud detection (in the diagram, β denotes this effect). The random shocks are stochastic, representing the effects on the transition rates of variables that the model does not explicitly consider. The random shock toward being against computerized fraud detection is ε_1 and the random shock toward favoring computerized fraud detection is ε_2. For the linear decomposition, α and β add to the values of the random effects and $\alpha = -\beta$ (the effects are equal but in the opposite direction).

Coleman's model rests on these assumptions, which are similar to those made by Rubin (1974; Holland 1986, 948–949; Cox 1992, 295–296). The system is in equilibrium and the cell frequencies represent the equilibrium values (explicit for Coleman, implicit for Rubin). Both models are vulnerable to the fundamental problem of causal inference: the individual unit cannot be exposed simultaneously to the treatment and control conditions, as depicted by the vertical barrier of panel (b). The random shocks are serially uncorrelated (the error terms are independent from one observation to another). For a fixed pattern of the explanatory variables, the transition rates do not change over time (there is temporal stability).[16] Individuals characterized by the same pattern of the explanatory variables will have the same transition rates (there is unit homogeneity and constant effects, which exhibit unit-treatment additivity).[17] The transition rates are dependent on the current state of the individual and not on the individual's past history (there is causal transience).[18] The treatments affect the rates uniformly and independently of what happens to other individuals (there is no interference, i.e., the SUTVA or stable-unit-treatment-value assumption holds).[19]

In Eq. (3), when dp_1/dt is set to 0 (to portray equilibrium) and the resulting expression is solved for p_1, Eq. (4) results:

$$p_1 = q_{01}/(q_{01} + q_{10}). \tag{4}$$

More generally, taking into account the $k = 1$ or 0 categories, the equations that partition the q_{ij}s are (Coleman 1964, 120–123):

$$q_{01k} = \alpha_x x_k + \varepsilon_1 \tag{5}$$

$$q_{10k} = \beta_u u_k + \varepsilon_2 \tag{6}$$

where x_k and u_k are the same attribute, namely, claims office; but oriented in opposite directions for the two conditions of work: $x_1 = 1$ and $u_1 = 0$ for working in target offices and $x_0 = 0$ and $u_0 = 1$ for working in comparison offices.

For employees who would work in the target offices ($k = 1$), the proportion against computerized fraud detection at equilibrium (e) is estimated to be

$$p_{1e} = q_{011}/(q_{011} + q_{101}) = (\alpha + \varepsilon_1)/(\alpha + \varepsilon_1 + \varepsilon_2) = \mathrm{n}_{11}/(\mathrm{n}_{11}+\mathrm{n}_{01}). \tag{7}$$

For employees who would work in the control offices ($k = 0$), the proportion against computerized fraud detection at equilibrium is estimated to be

$$p_{0e} = q_{010}/(q_{010} + q_{100}) = \varepsilon_1/(\alpha + \varepsilon_1 + \varepsilon_2) = \mathrm{n}_{10}/(\mathrm{n}_{10}+\mathrm{n}_{00}). \tag{8}$$

The parameter α ($= -\beta$) equals the difference between the proportions:

$$p_{1e} - p_{0e} = \alpha/(\alpha + \varepsilon_1 + \varepsilon_2) = 45/77 - 19/108 = 0.584 - 0.176 = 0.408. \tag{9}$$

The random shock ε_1 equals $p_{0e} = 0.176$ and the random shock $\varepsilon_2 = (1-p_{1e}) = 0.416$. As shown next, under the assumptions made earlier, Rubin's (1974) causal model applied to these data produces an identical estimate of the effect size.

Because the proportions with score 1 on the anti-index are means, the estimate of the Rubin (and Holland) average treatment effect δ for these data is identical to the estimate of Coleman's effect parameter: Eq. (7) for $p_{1e} = q_{011}/(q_{011} + q_{101}) = E\,(Y_t)$. Equation (8) for $p_{0e} = q_{010}/(q_{010} + q_{100}) = E(Y_c)$. Equation (9), $p_{1e} - p_{0e} = \alpha/(\alpha + \varepsilon_1 + \varepsilon_2) =$ the average treatment effect δ. The mean proportion, who are anticomputerized fraud detection among those working in the target offices, is (45 × 1 + 32 × 0) / 77 = 45 / 77 = 0.584. Among those working in the control offices, the mean proportion is (19 × 1 + 0 × 89) / 108 = 0.176. The difference = 0.584 – 0.176 = 0.408 is the average treatment effect δ, whose value equals that of the Coleman effect parameter for these data. Under the assumptions made earlier, this effect can be estimated by d_{yx}, f_{ij}, and D (Smith 1974, 212–221; Marsden 1985). However, the two types of claims offices and their employees may not be perfectly matched, thus requiring controls that would strengthen the precision of the estimated causal effect (Gelman and Hill 2007, 169–170).

Linear decomposition of effects: several explanatory attributes

Gelman and Hill (2007, 167) distinguish between the predictive inference of the usual regression analysis, which relates to comparisons between individual units, and causal inference, which focuses on comparisons of different treatments when applied to the same units. Offering causal interpretations of the regression coefficients, Coleman's model is more in the spirit of regression as predictive inference. Offering insights about what would have happened if the different treatments are applied to the same units, Rubin's model is more in the spirit of causal inference. Given the earlier assumptions about the data, these approaches produce the same estimates of the effects.

Derivation of the model

When Coleman develops his model of the effects of several explanatory variables on the transition rates between the two states of a dichotomous attribute (1981, 19–29), he changes the notation from that of his earlier model. The transition rate q_{ji} from state h (home) to state j for person i is now:

$$q_{ji} = \sum_{k=0}^{n} b_{kj} x_{ki} \tag{10}$$

where the x_{ki} ($k = 0, 1, \ldots, n$) may be dichotomous attributes (0, 1) or continuous variables. Since Coleman assumes that $x_{0i} = 1$ for all persons i, b_{0j} is the constant term in Eq. (10) and the parameters b_{kj} are measures of the effect of x_{ki} on q_{ji}. For the transition rate from $h = 0$ to destination state $j = 1$ for person i, Eq. (10) becomes

$$q_{1i} = \sum_{k=0}^{n} b_k x_{ki}. \tag{11}$$

For the transition rate from $h = 1$ to destination state $j = 0$ for person i, Eq. (10) becomes

$$q_{0i} = \sum_{k=0}^{n} c_k x_{ki}. \tag{12}$$

Coleman (1981, 22) assumes equal and opposite effects (i.e., $c_k = -b_k$) for all parameters except the constant terms c_o and $b_{\mathrm{o},}$ which may differ.

Using Eq. (3) and the expressions above for q_{1i} and q_{0i}, the following equation for the effects of n explanatory variables on the rate of change in the probability of being in state 1 can be derived (Coleman 1981, 22):

$$dp_1/dt = -p_i(c_o + b_o) + b_o + \sum_{k=1}^{n} b_k x_{ki}. \tag{13}$$

At equilibrium, $dp_1/dt = 0$ and Eq. (13) becomes

$$p_i = b_0/(c_0 + b_0) + \sum_{k=1}^{n} b_k x_{ki}/(c_0 + b_0). \tag{14}$$

In Eq. (14), when β_0 is substituted for $b_0\,/(c_0 + b_0)$ and β_k for $b_k\,/(c_0 + b_0)$, and the limits on the summation are changed to reflect these changes, Eq. (15) is derived; it is the linear probability model for a dichotomous response variable:

$$p_i = \sum_{k=0}^{n} \beta_k x_{ki}. \tag{15}$$

The parameters of effect in this linear probability model can be estimated using ordinary least squares (Coleman 1970, 231–232) or explicit maximum likelihood methods (Coleman 1981, 57–62). SAS's Proc Genmod (1997a, chapter 10) is a convenient computer program for obtaining maximum likelihood estimates for generalized linear models.[20] To calculate the unstandardized effects in a linear probability model, simply assume for the random component either a binomial or a normal probability distribution and an identity link function (Agresti 1996, 74–76); the systematic component specifies the set of explanatory variables. Path coefficients can be obtained by simply multiplying each effect by the ratio of the standard deviation of that explanatory variable to the standard deviation of the response variable, or fully standardize the variables prior to estimation.

A convergence

Coleman's model of the effects of several causes can produce an estimate of average treatment effect that is mathematically equivalent to that of Rubin's widely accepted causal model. When several control variables are operating on the transition rates, an estimate of the Rubin and Holland average treatment effect δ that focuses on *the effect of a cause* can be obtained by first expanding Eq. (15) to get:

$$p_i = \beta_{0a} + \beta_1 X_{1i} + \sum_{k=2}^{n} \beta_k x_{ki} + \varepsilon_i. \tag{16}$$

In this equation, consider the effects on the proportion p_1 in state 1 of the anti-index of the unstandardized dichotomous treatment variable X_1 ($X_1 = 1$ for the target offices; $X_1 = 0$ for the comparison offices) and the standardized-by-their-mean control variables x_2 through x_n. That is, X_1 is not centered by its mean, whereas the other explanatory variables are centered by their respective means. By standardizing each of the covariates by their mean, their effects appear in the intercept term $\beta_{0\mathrm{a}}$.[21] Since this parameter appears in both the treatment and control cells of the design, and is based on the full number of units, it provides a basis for the causal inference: The average treatment effect δ, controlling for the other variables, can be obtained from these expressions:

$$\text{Mean } \hat{p} \text{ given} X_1 \ \mathit{is}\ 1 = \beta_{0a} + \beta_1 \tag{17}$$

$$\text{Mean } \hat{p} \text{ given } X_1 \text{ is } 0 = \beta_{0\mathrm{a}}$$

$$\text{Average treatment effect } \delta = (\beta_{0\mathrm{a}}{+}\beta_1) - \beta_{0\mathrm{a}} = \beta_1.$$

When the controls are not standardized by their means, β_0 is the intercept and β_1 is the Coleman parameter for the effect of the treatment variable; its value will equal the average treatment effect δ, controlling for the other variables. For example, the regression of the anti-index on X_1 = working in the target office, X_2 = workers concerned with bodily injury claims, and X_3 = working in offices with a new administrative computer system (all Xs assumed to be on equal footing) is:

$$\hat{p} = .03 + .373X_1 + .140X_2 + .117X_3. \tag{18}$$

Apparently, working in a target office induces more discontent than working on bodily injury claims or being exposed to the new administrative computer system; when the analysis focuses only on the effect of the pivotal causal variable (i.e., understanding the effect of a cause) such comparisons may be lost.

When the two control variables are centered by their means, the analogous equation is:

$$\hat{p} = .186 + .373X_1 + .140x_2 + .117x_3. \tag{19}$$

Using Eq. (17) and the values from Eq. (19), the average treatment effect δ, controlling for x_2 and x_3, = (0.186 + 0.373) − (0.186) = 0.373. This value exactly equals the effect parameter for the treatment in (18).

The estimate of $\beta_{0\mathrm{a}}$ results when the control variables are evaluated at their means; their effects appear in the intercept term:

$$\beta_{0a} = .1865 = \beta_0 + \beta_2\bar{X}_2 + \beta_3\bar{X}_3 = .03 + .140 \times .5625 + .117 \times .6648 = .1865. \quad (20)$$

Conceptually, Eq. (17) suggests a hypothetical experiment in which all the cases are exposed to the target treatment ($X_1 = 1$) and then all the cases are exposed to the control treatment ($X_1 = 0$); the difference between the two proportions is the average treatment effect δ. Here, the average treatment effect δ (conditioned on controls) equals the Coleman parameter (conditioned on controls).

A linear probability model has the advantage of easy interpretation of its effect coefficients, but it also has shortcomings. The effects on a dichotomous response variable of dichotomous and ordinal attributes can be readily conceptualized as weighted averages across the component tables that compose a higher-order cross-tabulation (Boyle 1966; Somers 1974, 238–240). However, probabilities fall between 0 and 1, whereas this model may predict proportions less than 0 or greater than 1; and all transition rates must be nonnegative, while a linear sum need not obey this constraint.[22] Coleman corrected these weaknesses, and others concerning tests of significance and model fit noted by Goldberger (1964, 248–251) and Goodman (1972a, 1978, 23–24), by respecifying his model using the logit transform, ln [p/(1 − p)].

Multiplicative decomposition of effects

Coleman states ([1973] 2006, 18): "Logit analysis appears to have the greatest net advantage. It has good statistical properties, with a variance approximately constant over the range of p, and calculation is not difficult." Goodman (1972a, 1978) and Gelman and Hill (2007, 70) agree, the latter state: "logistic regression is the standard way to model binary outcomes (that is, data y_i that take on values 0 or 1)." Because of its advantages over the linear probability model, this subsection sketches Coleman's derivation of his causal interpretation of the coefficients of the logit model and relates his notion to Rubin's model.

Derivation of model

Coleman develops his explanatory model on the basis of an exponential decomposition of q_j's by first solving Eq. (3) for the ratio of the two transition rates:

$$p/(1-p) = q_1/q_0. \quad (21)$$

Equation (22) expresses the ratio of the transition rates as an exponential decomposition of the effect parameters for individual i (Coleman 1981, 22–23):

$$q_{1i}/q_{0i} = \exp\{b_o - c_o + 2\sum_{k=1}^{n} b_k x_{ki}\}. \tag{22}$$

After substituting this expression into Eq. (21) and taking the natural logarithm of both sides, Eq. (23) gives:

$$\ln[p/(1-p)] = b_{\mathrm{o}} - c_{\mathrm{o}} + 2\sum_{k=1}^{n} b_k x_{ki}. \tag{23}$$

Now substitute into (23) $\beta_0 = (\mathrm{c}_0 - \mathrm{b}_0)$ and $\beta_k = 2b_k$ for $k = 1, \ldots, n$, and change the limits in the summation, to obtain the log of the odds, the logit

$$\ln[p/(1-p)] = \mathrm{logit}(p) = \sum_{k=0}^{n} \beta_k x_{ki}, \tag{24}$$

this is the exponential linear decomposition. Let $u = \sum_{k=0}^{n} \beta_k x_{ki}$. Then, this expression for the probability of occupancy of state 1 can simplify the interpretation of the effects:

$$p_i = e^u/(1+e^u) = 1/\{1+\exp(-u)\} = 1/[1+\exp(-\sum_{k=0}^{n} \beta_k x_{ki})], \tag{25}$$

this is the multivariate logistic equation.

Maximum likelihood estimates of the β coefficients in (24) readily can be obtained using SAS's Proc Genmod or Proc Glimmix; simply assume a binomial random component and a logit link transformation. Or, use standard logistic regression programs such as those included in SAS, Stata, or SPSS.

One explanatory attribute

Under the assumptions made earlier, the equivalence between the estimates of an effect parameter and the Rubin (and Holland) average treatment effect, δ also holds for the multiplicative decomposition of effects. For example, using the terms of Eq. (25), Table 3.2 defines the proportions in each cell of the fourfold table that relates working in the target office or working in a control office to the employees' attitudes on the anticomputerization index. Applying (24) to the data of panel (a) of Table 3.1, the estimate of the intercept $\beta_0 = -1.544$ and the estimate of the effect of the intervention $\beta_1 = 1.885$. The Rubin-esque average treatment effect δ is

Table 3.2 Multiplicative decomposition of effects

Computerized Fraud Detection	Employees in Target Offices	Employees in Comparison Offices
State 1 Against	$p(1) = e^{(\beta0 + \beta1)}/(1 + e^{(\beta0 + \beta1)})$	$p(0) = e^{(\beta0\ )}/(1 + e^{(\beta0)})$
State 0 For	$1 - p(1) = 1/(1 + e^{(\beta0 + \beta1)})$	$1 - p(0) = 1/(1 + e^{(\beta0\ )})$

The maximum likelihood estimates are $\beta_0 = -1.544$ and $\beta_1 = +1.885$.

estimated by the difference between these two proportions (i.e., means): the proportion in state 1 of the anti-index for those who would be working in the target offices minus the proportion in state 1 of the anti-index for those who would be working in the control offices. Substituting the estimates of β_0 and β_1 into the expressions for the proportions in state 1 of the anti-index in Table 3.2, the proportion for the target group = odds / (1 + odds) = $e^{.341} / (1 + e^{.341}) =$ 1.4064 / 2.4064 = 0.584, and the proportion for the control group = odds / (1 + odds) = $(e^{-1.544}) / (1 + e^{-1.544}) = 0.2135 / 1.2135 = 0.176$. Consequently, $\delta = 0.584 - 0.176 = 0.408$, which is the same value as found earlier for the linear probability model.

The effect parameters for the multiplicative decomposition of effects are defined as the coefficients on the logit scale of (24). Using the data of panel (b) of Table 3.2, for the employees who would be working in comparison offices, the estimated odds of being against computerized fraud detection are 19/89 = 0.2135; β_0 is the natural log of these odds; its value is ln(0.2135) = −1.544, as reported above. For the employees who would be working in the target offices, the estimated odds of being against computerized fraud detection are 45/32 = 1.406; the natural log of these odds is ln(1.406) = 0.3409. The logit difference (or log of the odds ratio) is 0.3409 − (−1.544) = 1.885, which equals the value of β_1, as reported above.

The proportions in each cell can be obtained from the odds by again using this expression: p = odds / (1 + odds). Thus, the Coleman effect parameters are: $p_{0\mathrm{e}} = (\varepsilon_1) / (\alpha + \varepsilon_1 + \varepsilon_2) = 0.2135 / (1 + 0.2135) = 0.176$; $\mathrm{p}_{1\mathrm{e}} = (\alpha + \varepsilon_1) / (\alpha + \varepsilon_1 + \varepsilon_2) = 1.4064 / (1 + 1.4064) = 0.584$; and $\alpha = p_{1e} - p_{0e} = 0.408$; which equals the Rubin-esque treatment effect above.

Several explanatory variables

When there are covariates, an estimate of the Rubin and Holland average treatment effect δ for the effect of the cause of primary interest can be obtained from a logistic regression by first expanding (25):

$$p_i = \exp(u)/(1 + \exp(u)) \text{ where } u = \beta_{\mathrm{oa}} + \beta_1 X_{1i} + \sum_{k=2}^{n} \beta_k x_{ki} + \varepsilon_i \qquad (26)$$

In this equation, consider the effect on the proportion p in state 1 of the anti-index of an unstandardized dichotomous treatment variable X_1: $X_1 = 1$ for employees working in the target offices; $X_1 = 0$ for employees working in the comparison offices.

If variables x_2 (= bodily injury workers) through $x_{n=3}$ (offices with the new administrative computer system) are evaluated at their mean values of zero and variable X_1 is not, then an estimate of the average treatment effect δ can be obtained from Eq. (27) as follows:

$$p_{\mathrm{i}} = \mathrm{exp(u)} / (1 + \mathrm{exp(u)})\,\mathrm{where}\ \mathrm{u} = -1.5812 + 1.7771X_1 + .8011x_2 + .7290x_3 \tag{27}$$

$$\begin{aligned}\bar{p}\ \text{given}\ X_1\ \text{is 1} &= e^{(\beta 0a\,+\,\beta 1)} / (1 + e^{(\beta 0a\,+\,\beta 1)}) = \mathrm{e}^{.1959} / (1 + \mathrm{e}^{.1959})\\ &= 1.2164 / 2.2164 = 0.549.\end{aligned}$$

$$\begin{aligned}\bar{p}\ \text{given}\ X_1\ \text{is 0} &= e^{(\beta 0a)} / (1 + e^{(\beta 0a)}) = \mathrm{e}^{\cdot -1.5812} / (1 + \mathrm{e}^{-1.5812})\\ &= 0.206 / 1.206 = 0.171.\end{aligned}$$

The average treatment effect $\delta = 0.549 - 0.171 = 0.378$, which is about the same as the $\delta = 0.373$ from the linear probability model.

When all the variables in Eq. (25) are not standardized by their means, then the expression for u is similar to (27), but the intercept is different:

$$\begin{aligned}p_{\mathrm{i}} &= \mathrm{exp(u)} / (1 + \mathrm{exp(u)})\ \mathrm{where}\ \mathrm{u}\\ &= -2.5101 + 1.7771X_1 + .8011X_2 + .7290X_3\end{aligned} \tag{28}$$

If u is evaluated at the means of $X_2 = 0.5625$ and $X_3 = 0.6648$, and X_1 can be either 1 for employees who would be working in the target offices or 0 for employees who would be working in the control offices, then $u = -2.5101 + 0.8011\ (0.5625) + 0.7290\ (0.6648) + 1.7771X_1 = -1.5748 + 1.7771X_1$. When $X_1 = 0$ for workers in the comparison offices, then $e^u = \mathrm{e}^{-1.5748} = 0.207$, and the proportion $p_0 = \mathrm{odds}/(1 + \mathrm{odds}) = 0.207/1.207 = 0.1715$. When $X_1 = 1$ for workers in the target offices, then $e^u = \mathrm{e}^{-1.5748 + 1.7771} = \mathrm{e}^{.2029} = 1.225$, and the proportion $p_1 = \mathrm{odds}/(1 + \mathrm{odds}) = 1.225/2.225 = 0.551$. The Coleman effect parameter is the difference between these proportions: $0.551 - 0.1715 = 0.38$, which is the same value as the treatment effect δ.[23]

Causal inference and Coleman's models

These examples applied Coleman's stochastic process model to survey data in which the variables are conceptualized as attributes. His model can quantify the direct effects of several explanatory attributes on a response variable (i.e., *the causes of an effect*), thereby facilitating the comparison of the sizes of the implied causal forces. By standardizing the covariates by their means, Coleman's model can be transformed to focus on the *effect of the pivotal cause*. This transformation suggests a formal equivalence, given the same data and the same assumptions, between the estimated Coleman effect parameter and the Rubin-esque average treatment effect.

This convergence suggests that Cox and Wermuth's notion of level-one causality (2004, 289–290) includes the examples developed here. But, for these examples the statistical models do not represent the underlying substantive reality; not one of the various estimates of the causal force of the attribute “working in a target office” compared with “working in a control office” is a valid estimate of the causal effect. In actual fact, the study design was broken because workers in two offices were exposed to two treatments simultaneously—the fraud detector and a new administrative computer system, referred to as M2K. This violates the stable-unit-treatment-value assumption (SUTVA) because the design called for exposure of the workers to only one treatment per office; the second computer system interfered with the effects of the first. Moreover, in one office workers were exposed only to the fraud detector, in another office workers were exposed only to M2K, and in two other offices the workers used only the established computer system. The large effect of working in an office in which the fraud detector was installed on the workers’ negative attitudes is false. It disappears when the effect of the simultaneous introduction of two computer systems is controlled using either an instrumental variable correction for selection bias ($\delta = -0.02$, $p = 0.98$), see Smith (1999a), or hierarchical modeling ($\delta = 0.06$, $p = 0.89$), see Chapter 9.

Applications of statistics derived from Coleman’s process model do not necessarily produce estimates of causal effects; the underlying data to which his models are applied must fulfill criteria for causal inferences. This finding applies as well to Coleman’s (1981) innovative development of his basic model to cover longitudinal data analysis. But, careful applications of Coleman’s model may produce appropriate estimates of causal effects of manipulated and nonmanipulated variables for cross-sectional and longitudinal data.[24] If confounding variables are taken into account, and the model correctly formalizes the substantive reality, then the estimates from applications of Coleman’s models will be notional causal effects. Thus, there can be multiple causal forces on a response without manipulation of these causes (for further analysis and examples, see Goodman ([1972b] 1978, 57–109). As Cox and Wermuth (2004, 287) imply, endogenous attitudes and prior exogenous variables on equal footing like age, level of education, gender, family size, and occupation should not appear in the same single-equation model if the goal is to estimate the effects of these exogenous variables. However, if the hypothesized causal variable is an endogenous variable, then its effect on the response should be conditioned on the exogenous variables, thus controlling for potentially spurious effects.[25]

If the effects of the manipulated treatment and the fixed covariates are estimated with *allowable* controls for background variables and for selection bias, and measurement error is minimal, then causal rather than associational inferences can be made from applications of Coleman’s model, especially since its estimates of the average treatment effect are mathematically equivalent to those from Rubin’s widely accepted causal model. Both models can be conceptualized as $yx.T^p$ and can produce implied causal inferences. The general form of the explanation is similar to that of single-equation regression model; the explanatory variable influences the response controlling for a wide range of factors that could produce spurious or independent effects.[26]

When the explanatory variables in such models are at different levels of priority, models of dependency networks that are rooted in path analysis and structural equation models (SEMs) are more appropriate. These models require detailed knowledge of the subject matter, a closed system of variables in which the effects of extraneous variables appear in the stochastic terms, strong evidence for the directions of the assumed causal effects, and replications of the effects in other studies. The next chapter focuses on dependency networks.

Acknowledgements The author derived *Causal Models for Attributes*, a section of this chapter, from portions of his earlier writings: Introduction to James S. Coleman's *Mathematics of collective action* (*vii–li*, © Aldine Transaction 2006) and Inferential causal models (*Quality* & *Quantity*, vol. 37, issue 4, 337–361; © Kluwer2003). The author thanks these publishers for their permission to reuse this material here.

Endnotes

[1] Morgan and Winship (2007) and Greenland (2004) provide overviews; Pearl (2000, 2009, second edition) and Heckman (2005a) are more difficult; and Sobel (2005) and Heckman (2005b) are contentious.

[2] Cox and Wermuth have referred to the categories of their three-category classification using different terms. Their 2004 article defines zero-, first-, and second-level causality. Their 2001 article defines causality as stable association, causality as the effect of an intervention, and causality as explanation of a process. Their 1996 book distinguishes these categories as follows (page 220):

> There are essentially three interrelated senses in which causality arises in the contexts considered in this book:
>
> 1. As a statistical dependence, which cannot be removed by alternative acceptable explanatory variables;
> 2. As the inferred consequences of some intervention in the system;
> 3. As the above, augmented by some understanding of a process or mechanism accounting for what is observed.
>
> All these notions are valuable; we favor restricting the word to the final and most stringent form.

They emphasize (1996, 219) that causality rarely can be inferred from the results of one study, especially observational studies.

[3] Morgan and Winship (2007, 286–287) refer to the first aspect of assessment of stable association as Mode 1: Associational Analysis. It is a prerequisite to causal analysis in that there is no correlation without association. They distinguish this mode from their Mode 2: Conditional Association Analysis, which researchers apply after establishing that an association exists in order to eliminate obvious sources of spuriousness. They note that conditioning variables may be determinants of the response variable y and not causes of the key explanatory variable x (such variables are on equal footing with x) or they may be causes of both x and y (such variables are prior to x and to y). The controls for variables on equal footing with x purify the nonspurious effect of x on y; the controls for variables prior to both x and y test directly for spurious association.

[4] Cox (1992, 292–293) refers to this operation as causality via association (i.e., zero-level causality) and formalizes testing for spuriousness as follows (1992, 292–293); this logic applies to the procedures for assessing causality of Suppes and Granger discussed later in this chapter:

> Let C and E be binary events and B be a third variable or collection of variables. We may say that C is a candidate cause of E if C and E are positively associated, i.e. if

$$P(E|C) > P(E|not\, C).$$

We can regard the cause as spurious if B explains the association in that

$$P(E|C\ and\, B) = P(E|not\ C,\ and\ B),$$

i.e. if E and C are conditionally independent given B. The cause is confirmed if C is a candidate cause that is not spurious, i.e. which cannot be so explained via any B.

In contrast, Pearl (personal communication, 5/2/2010) defines spurious correlation in terms of his $do(x)$ operator as follows:

C is spuriously related to E if you find correlation without causation:

1. C and E are dependent, and
2. $P(e|do(c)) = P(e)$ for all c
3. $P(c|do(e)) = P(c)$ for all e

For dichotomous C and E: $P(e|do(C = 1) = P(\mathrm{e}|do(C = 0) = P(e)$ and $P(c|do(E = 1) = P(c|do(E = 0) = P(c)$. The probabilities in each set of tables are equal and equal the marginal probability.

Then, by implication C causes E if you find causation (Pearl 2002, 208):

1. $\mathrm{P}(e|do(c)) \neq \mathrm{P}(e)$ for some values c and e

For dichotomous C and E: $P(e|do(C = 1) \neq P(e)$ or $P(e|do(C = 0) \neq P(e)$. A probability in each set of tables does not equal the marginal probability.

[5] Lazarsfeld ([1946] 1972) used panel data to form 16-fold tables that enabled him to ascertain the mutual effects of x and y: whether the effect of x on y is larger than the effect of y on x. Coleman (1964, 162–188) developed causal models for such two-attribute systems for over time data; his computer program provided parameter estimates. Goodman ([1973] 1978, 173–229) developed causal models based on loglinear models for such data.

[6] Susser (1973, 111–135) clarifies classic epidemiological approaches for assessing spuriousness and elaborating analyses. Greenland, Pearl, and Robins (1999a) show how investigators can apply graphical methods to determine which variables in a system are sufficient when controlled so as to remove bias from the estimate of the causal effect of exposure on a disease. Greenland, Robins, and Pearl (1999b) provide an overview of confounding based on a counterfactual model of causation. Advancing these introductory expositions, Robins and his colleagues apply highly abstract, state-of-the-art Bayesian networks and graphical methods to contemporary epidemiological problems. Greenland (2004, 6–8) discusses Robins's g-estimation approach; for a list of Robins's publications on this topic, see his website: http://www.biostat.harvard.edu/~robins/research.html

[7] Let Y_i be the realized empirical outcome for individual i, $Y_i(1)$ be the potential outcome if individual i would receive the experimental treatment, $Y_i(0)$ be the potential outcome if individual i would receive the control treatment, and W_i be an indicator such that $W_i = 1$ implies that the individual would receive the experimental treatment and $W_i = 0$ implies that the individual would receive the control treatment. Then (Imbens and Wooldridge 2009, 9): $Y_i = Y_i(W_i) = Y_i(0)(1 - W_i) + Y_i(1)(W_i)$. When $W_i = 1$, then $Y_i = Y_i(1)$. When $W_i = 0$, then $Y_i = Y_i(0)$. A specific manipulation W_i would make one of the potential outcomes $(Y_i(0), Y_i(1))$ the realized outcome Y_i.

[8] Cox highlights limitations to the applications of Rubin's causal model, distinguishing treatments from intrinsic properties of individuals under study, and from nonspecific factors such as countries and organizations (1992, 296):

> In this discussion, only those variables which in the context in question can conceptually be manipulated are eligible to represent causes, i.e. it must make sense, always in the context in question, that for any individual the causal variable might have been different from the value actually taken. Thus in most situations gender is not a causal variable but rather an intrinsic property of the individual. The study of sex-change operations and of possible discriminatory employment practices would be exceptions. Again, the passage of time as such is not a causal variable.

[9] Note that $R_1 - R_0 = \Delta$ implies that $R_1 = R_0 + \Delta$; the model assumes unit-treatment additivity, as does Coleman's model for attributes. That is, "the response that would be observed under $C = 1$ differs by a constant Δ from the response that would be observed on that same unit were it to receive $C = 0. \ldots \Delta$ [is] the causal effect of changing C from 0 to 1" (Cox 1992, 295).

[10] Imbens and Wooldridge (2009, 10–11) discuss five advantages that the potential outcomes approach (POA) has over the analysis of realized outcomes. They essentially state that the POA: (1) allows the definition of causal effects before specifying the assignment mechanism, and without making functional form or distributional assumptions; (2) links the analysis of causal effects to explicit manipulations; (3) separates the modeling of the potential outcomes from that of the assignment mechanism; (4) allows formulation of probabilistic assumptions in terms of potentially observable variables, rather than in terms of unobserved components; and (5) clarifies where the uncertainty in the estimators comes from.

[11] Harding (2003) established that people who are exposed to different levels of neighborhood poverty between the ages of 11 and 20 have different rates of dropping out of school and, for females, teenage pregnancies. Being exposed to poverty as a teenager can be conceptualized as a contextual attribute of the person. With this conceptualization, Harding's study exemplifies a causal analysis of the effects of an attribute.

[12] Explication of overlaps between Coleman's and Rubin's models of causality appeared in the author's unpublished paper (Smith 1999a). Portions of this present explication appear in Smith (2003a, 346–353; 2006, xi-xxviii) and are used here, respectively, with the permission of Kluwer Academic Publishers and AldineTransaction, who graciously allow their authors to reuse their work in their own publications. These earlier expositions and Chapter 9 conceptualize the introduction of a new computer system as a treatment.

[13] Greenland (2004, 4) cites this passage from David Hume that is perhaps the original statement of the potential outcomes notion of causality; it bears on Rubin's and Coleman's models:

> "We may define a cause to be an object, followed by another,…if the first object had not been, the second had never existed."

Here, if the person has the attribute "working in a target office" then that person has a high probability of being against computerized fraud detection. However, if that person had the attribute "working in a control office," then that person would have a lower probability of being against computerized fraud detection. All things being equal, the causal effect of the attribute "working in a target office" is the difference between these probabilities.

[14] Coleman's model is thus clearly in the tradition of Kurt Lewin; the movement depends on relative sizes of these forces (Lewin [1951] 1997, 136). Coleman was very knowledgeable about Lewin's notions of psychological forces: He had written an unpublished paper with Allen Barton that explicated the concept of cohesiveness as developed by Festinger et al. (1950).

[15] Heckman's (2005, 12) simplest statement of the potential outcomes model is that: "the *individual-level treatment effect* for person w comparing outcomes from treatment s with outcomes from treatment s' is $Y(s, w) - Y(s', w)$, $s \neq s'$. Initially, Rubin restricted the treatments to be manipulated variables, whereas Coleman allows the treatments to be different categories of an attribute.

[16] Temporal stability asserts the constancy of the response over time (Holland 1986, 948).

[17] The assumption of unit homogeneity implies that $Y_t(u_1) = Y_t(u_2) = Y_t(u_n)$ and $Y_c(u_1) = Y_c(u_2) = Y_c(u_3)$ (Holland 1986, 948). In Coleman's model, this means that each person in one treatment cell is characterized by the same causal force. The assumption of constant effect implies that the effect of t on every unit is the same: $T = Y_t(u) - Y_c(u)$ for all u in U (Holland 1986, 949). For the treatment cell, unit-treatment additivity implies that the treatment t adds a constant amount T to the level of the control response for each unit, as in Coleman's model.

[18] Causal transience asserts that there is a "washout effect," the person's response to the treatment c at an earlier time does not affect the person's response to the treatment t at a later time.

[19] SUTVA implies: "that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatment the other units receive." (Rubin 1986, 961). If in fact there are two treatments being applied to some units in the treatment group, or if there is leakage of the treatment t to the units in the control group, then SUTVA is violated.

[20] Hellevik (2009) offers a spirited defense of the use of the OLS linear regression model rather than a logistic regression model when the response is a dichotomy. However, he does not consider the transformation of odds ratios into probabilities p using the inverse link function; briefly $p = \text{odds}/(1 + \text{odds})$.

[21] The propensity scores standardized by their overall mean can be included as a covariate that would appear in the intercept term along with the other covariates, see Chapter 12 for an example.

[22] To avoid these problems, prior to obtaining the final predicted proportions the data can be made linear by applying Goldberger's method; see Achen (1986, 41) for instructions. Chapter 12 also provides an example; it calculates propensity scores using ordinary least squares after using Goldberger's method and then compares these scores to those obtained from logistic regression. After conducting an OLS regression, a first step is to change a unit's predicted value p greater than 1 to .99 and p less than 0 to .01. For each unit set, $q = 1\text{–}p$ and $s = (pq)$½. Then, for each variable on a unit divide each variable's score by that unit's s. In the subsequent "no intercept" regression, the intercept is set to $1/s$. Then, apply OLS to the transformed data to get the "Goldbergerized" regression coefficients and their standard errors. Logistic models that do not converge may do so when applied to linearized data if Firth's (1993, 1995) penalized likelihood method is applied; also see Chapter 12 for an example. The OLS and logistic estimates correlated .999.

[23] Given an equation similar to (289), Coleman developed this strategy for obtaining the effect parameter from a logistic regression. He asks (1981, p. 31): "What amount is added to p when $x_1 = 1$ if, when $x_1 = 0$, p is 0.5?" He defines this transformation α_k of β_k such that $0.5 + \alpha_\mathrm{k} = 1/(1 + \mathrm{e}^{-(0 + \beta k)})$. By solving for α_k, Coleman's transformation is obtained:

$$\alpha_k = \left(\mathrm{e}^{\beta k} - 1\right)/2\left(\mathrm{e}^{\beta \mathrm{k}} + 1\right)$$

When $\beta_k = 1.7771$ is substituted in the expression above, $\alpha_\mathrm{k} = 4.9127/13.8254 = 0.36$, which is about equal to the average treatment effect $\Delta = 0.38$.

[24] Coleman (1964, 145–149) applies a variant of his causal model to experimental data in which one group receives the treatment and the other group a null treatment. The measure of the causal effect is similar to those in this chapter.

[25] Gelman and Hill (2007, 186–194) strongly advise against controlling for post-experimental-treatment variables that intervene between the randomized treatment and the response because this weakens the causal (i.e., potential outcomes) interpretation of the effect of the treatment. Regarding observational studies, they state:

> Researchers often already know to control for many predictors. So it is possible that these predictors will mitigate some of the problems we have discussed. On the other hand, studying intermediate outcomes involves two ignorability problems to deal with rather than just one, making it all the more challenging to obtain trustworthy results.

Ignorability implies that conditional on the confounding covariates used in a regression, the distribution of units across treatment conditions is, in essence, random. That is, the distribution of the potential outcomes is the same across levels of the treatment variable, when the confounding covariates are controlled (Gelman and Hill 2007, 182–184).

[26] For a recent discussion of the general form of additive linear models, see Moreno and Martinez (2008, 597–604).

Chapter 4
Dependency Networks

> *Scientific research has an aim of understanding the effect of one variable on another, and some notion of causality is involved in this. Our reason for caution is that it is rare that firm conclusions about causality can be drawn from one study, however carefully designed and executed, especially when that study is observational. The thrust of our discussion, however, especially the use of univariate recursive regression graphs, is to provide representations of data that are potentially causal, i.e., which are consistent with or suggestive of causal interpretation, but this is well short of actually establishing causality in a single study.*
>
> —Cox and Wermuth (1996, 219–220)

Dependency networks—models of networks of potentially causal relationships—capture the spirit of what Cox and Wermuth refer to as level-two causality, or causality as an explanation of a process (1996, 220; 2001, 70–74; 2004, 288, 299–300). Social scientists can develop such models on the basis of their knowledge of the subject matter and by synthesizing relationships that are invariant when tested through the procedures of robust dependence (i.e., stable association) or controlled intervention (i.e., potential outcomes). This chapter discusses four types of dependency networks: graphical models, association graphs for loglinear models, generative mechanisms, and structural economic models. Graphical models emphasize linear relationships; loglinear models detect interaction effects; path-analytic generative mechanisms allow the quantification of total, direct, and indirect effects; and structural economic models underscore the importance of theory. This chapter's detailed explications supplement contemporary discussions of notions of causality that do not provide extensive empirical examples. This chapter leaves open the question "Are these multivariate dependencies truly causal?"

Graphical Models

Cox and Wermuth (1996, 25–60); Cooper (1999, 3–62); Pearl ([2000] 2009, second edition); and Lauritzen (2001, 63–107) introduce the concepts, terms, and notations of graphical models. These models depict a linkage between two variables y and x

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_4, © Springer Science+Business Media B.V. 2011

in different blocks as a directed edge (i.e., arrow) that signifies a relationship that is conditionally dependent, when prior variables in the system are controlled (e.g., $b_{yx.abcde} \neq 0$). Contrarily, the absence of an edge between two variables in different blocks indicates that the two variables are conditionally independent, when prior variables in the system are controlled (e.g., $b_{yx.abcde} = 0$). A straight line depicts an edge between two variables in the same block.

This section illustrates Cox and Wermuth's graphical modeling strategy by focusing on the factors that influence voting; the data cover the 1992 presidential election that Bill Clinton won. Ongoing research is assessing the extent to which a similar model holds for the 2008 election.

Strategy

In their various expositions, Cox and Wermuth (1996, 135–170; 2001, 70–74; 2004) develop their recursive strategy for building graphical models, which is similar to the recursive path analytic and structural equation modeling of variables at different levels of measurement. Boyle (1966), Alwin and Hauser (1975), Davis (1975, 1980, 1985), and numerous other sociologists have applied path analysis recursively to develop what the earlier literature referred to as "causal models." If investigators have thoroughly studied the substantive process, the blocks of variables have unambiguous ordering, and the blocks contain theoretically relevant variables; then the pathways form a network of dependency relationships that are potentially causal.

For a specific substantive problem, the first step organizes the measured variables into blocks, which are depicted by boxes that are ordered by precedence: the response variable is located on the left, intervening variables are in the middle, and intrinsic

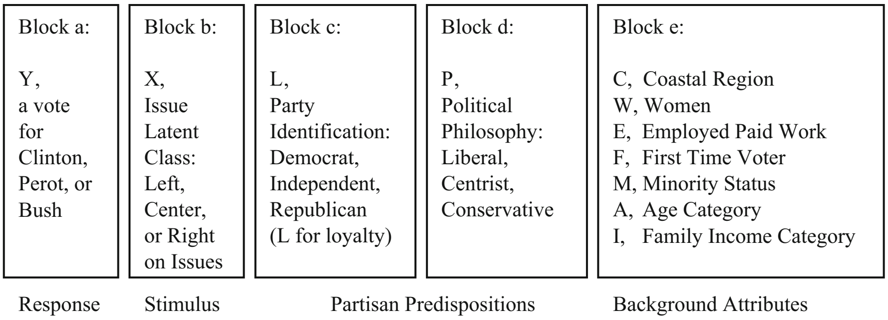


Fig. 4.1 The precedence ordering of blocks of variables bearing on electoral voting
Note: The data are from the Fredericks/Schneiders survey of the 1992 election in which the democrat candidate Bill Clinton was victorious over the republican candidate George Bush and the independent candidate Ross Perot. Political philosophy (P) and Party identification (L) are both aspects of partisan predispositions with the former influencing the latter

background variables are on the right. The voting example has five blocks of variables, a through e, see Fig. 4.1. The response variable in block a is coded Y for voting for Bill Clinton, Ross Perot, or George Bush. The explanatory variable in block b is coded X for a three-class latent structure that groups the issues of the campaign forming a left, center, and right ordinal continuum (for the details see Smith 2003b, reprinted 2004). As in the 2008 election, the economy, health care reform, the environment, and a candidate's character were salient issues. There are two partisan predispositions: Block c contains party identification—Democrat, Independent, or Republican—coded L (for loyalty to a party); and block d contains political philosophy—liberal, centrist, or conservative—coded P (for philosophy); a two-stage least-squares analysis found that party identification is a consequence of political philosophy (Smith 1999b). Block e contains five dichotomized and two trichotomized background attributes: residence in a coastal region (C), women (W), paid workers (E for employed), first-time voters (F), ethnic minorities (M), trichotomous age category (A), and trichotomous income category (I).

The second step explores the data, reporting in a table the matrix of partial and marginal correlations, the range of the variables, and their means and standard deviations. Inspection of the partial correlations provides clues about which relationships may be conditionally independent or dependent. (Because these data have been explored in other publications, here this step will be skipped.)

The third step estimates the effects recursively (Simon [1953] 1957, 10–49); the regression methods can vary depending on the levels of measurement of the response variables. Because all of the response variables in this example are ordinal trichotomies, these analyses apply logistic regression procedures; the proportional-odds model is applied first (Agresti 1996, 212–215; Stokes et al. 2000, 243–252). For a cross-tabulation of political philosophy—liberal, centrist, conservative—with Democrat or Republican party identification, this model assumes that the odds ratio in the fourfold table with political philosophy dichotomized as liberal versus [centrist + conservative] cross-tabulated with Democratic or Republican is the same as in the fourfold table formed by this alternative collapsing of political philosophy: [liberal + centrist] versus conservative cross-tabulated with Democrat or Republican. The model's goodness of fit is tested against the null hypothesis H_0: $\beta_k = \beta$ for all k, that is, H_0 states that the log odds ratios are the same in each fourfold table; probabilities less than 0.05 reject this hypothesis (Stokes et al. 2000, 249–250). If that model does not fit well, then continuation-ratio logits are used to decompose the ordinal trichotomies: the first category relative to the other two categories, and, given that the first category is now ignored, the second category relative to the third category (Agresti 1996, 218–220). For reasons of parsimony, variables exhibiting nonsignificant effects on a response variable will be deleted from that set of covariates and the resulting model reestimated.

The fourth step portrays the results in regression graphs, one for each regression analysis, in which all of the relevant prior variables appear in a doubly edged box in which continuous variables are depicted as circles and discrete variables are depicted as solid circles (dots). Arrows are drawn from the consequential variables to a response variable and the results are reported in tables; relationships between variables in the same block that are on equal footing may be depicted by an undirected edge, a straight line. Investigators synthesize their findings by depicting the salient relationships in a dependence graph and in parental and ancestor matrices, which form the basis of their interpretations.

Results

The recursive modeling procedure begins here by regressing Y, the voting choice, on all variables that are antecedent, those in blocks b to e. The first set of results in Table 4.1 indicate that the proportional-odds model does not fit the data ($p = 0.01$); so the analysis focuses on the determinants of the vote for Clinton (1) relative to that for [Bush + Perot] (0), and then vote for Perot (1) to that of Bush (0). The odds ratios clearly indicate that ethnic minorities, liberals and centrists, Democrats and Independents, and those with issue positions on the left and center were more likely to vote for Clinton. Of these determinants, Democratic party identification exhibits the largest odds ratio, 34.2. A somewhat similar pattern characterizes the vote for Perot versus Bush, with the exception of the much weaker effect of Democratic party identification and the null effect of ethnicity.[1]

The second analysis regresses the issue continuum on all of the variables in the prior blocks c to e; the proportional-odds model just fits. Liberal political philosophy and Democratic party identification determine the level of support for the more left positions on the issues.

The third analysis regresses party identification on political philosophy (block d) and on the background variables (block e). The proportional-odds model does not fit well, so the Clinton and Perot votes are analyzed separately. Liberals and centrists tended to vote for Clinton as did minorities, older or middle-aged people compared with younger people, and people with lower income compared with those with higher income. Liberals and centrists, low income people, and the middle-aged tended to vote for Perot rather than Bush.

The fourth analysis regresses political philosophy (block d) on the background variables (block e). For reasons of parsimony, the proportional-odds model is preferred to the two continuation-ratio logit models, even though the fit is questionable ($p = 0.044$). Residents of coastal regions, women, paid workers, and first-time voters tend to say they are liberals.

Interpretive Graphs

The regression graphs of Fig. 4.2 visually depict the results qualitatively. Graph (a1) portrays the regression of Clinton versus [Perot + Bush] and shows that four binary

Table 4.1 Logistic regression equations forming a recursive model of electoral voting

(1) Determinants of vote	Proportional odds (PO)		Clinton vs. All Others		Perot vs. Bush	
Clinton, Perot, or Bush	B	Exp(B)	B	Exp(B)	B	Exp(B)
Intercept1	−4.10 (0.30)		−3.80 (0.41)			
Intercept2	−2.40 (0.28)				−2.74 (0.34)	
M10, Minority Ethnicity vs. White	0.68 (0.27)	2	0.85 (0.30)	2.3	–	–
P13, Liberals vs. Conservatives	1.12 (0.20)	3.1	1.22 (0.24)	3.4	0.75 (0.29)	2.1
P23, Centrists vs. Conservatives	0.97 (0.19)	2.6	0.84 (0.23)	2.3	1.17 (0.25)	3.2
L13, Democrats vs. Republicans	3.68 (0.22)	39.8	3.53 (0.27)	34.2	1.96 (0.33)	7.1
L23, Independents vs. Republicans	1.77 (0.19)	5.9	1.46 (0.27)	4.3	1.82 (0.24)	6.2
X13, Left vs. Right Class	1.58 (0.31)	4.9	1.57 (0.41)	4.8	1.05 (0.43)	2.9
X23, Center vs. Right Class	0.88 (0.26)	2.4	0.74 (0.37)	2.1	0.90 (0.33)	2.4
Tests of fit:	PO $\chi^2 = 18.4$, df = 7, $p = 0.01$		H&L not applicable		H&L not applicable	
	Deviance/DF = 0.72, $p = 0.98$		Deviance/DF = 0.64, $p = 0.97$		Deviance/DF = 0.58, $p = 0.93$	
	BIC 1884.1 reduced to 1341.8		BIC 1220.9 reduced to 797		BIC 664.6 reduced to 567.2	
	−2*LL* 1868.3 reduced to 1270.9		−2*LL* 1213 reduced to 733.9		−2*LL* 657.3 reduced to 516	
(2) Determinants of issues latent structure	Proportional odds (PO)					
Left, Center, Right	B	Exp(B)				
Intercept 1	−2.71 (0.18)					
Intercept 2	1.05 (0.14)					
P13, Liberals vs. Conservatives	1.14 (0.19)	3.1				
P23, Centrists vs. Conservatives	0.79 (0.18)	2.2				
L13, Democrats vs. Republicans	1.11 (0.19)	3				
L23, Independents vs. Republicans	0.39 (0.19)	1.5				
Tests of fit:	PO $\chi^2 = 9.49$, df = 4, $p = 0.05$					
	Deviance/DF = 1.07, $p = 0.38$					
	BIC 1561 reduced to 1483					
	−2*LL* 1545.2 reduced to 1435.39					

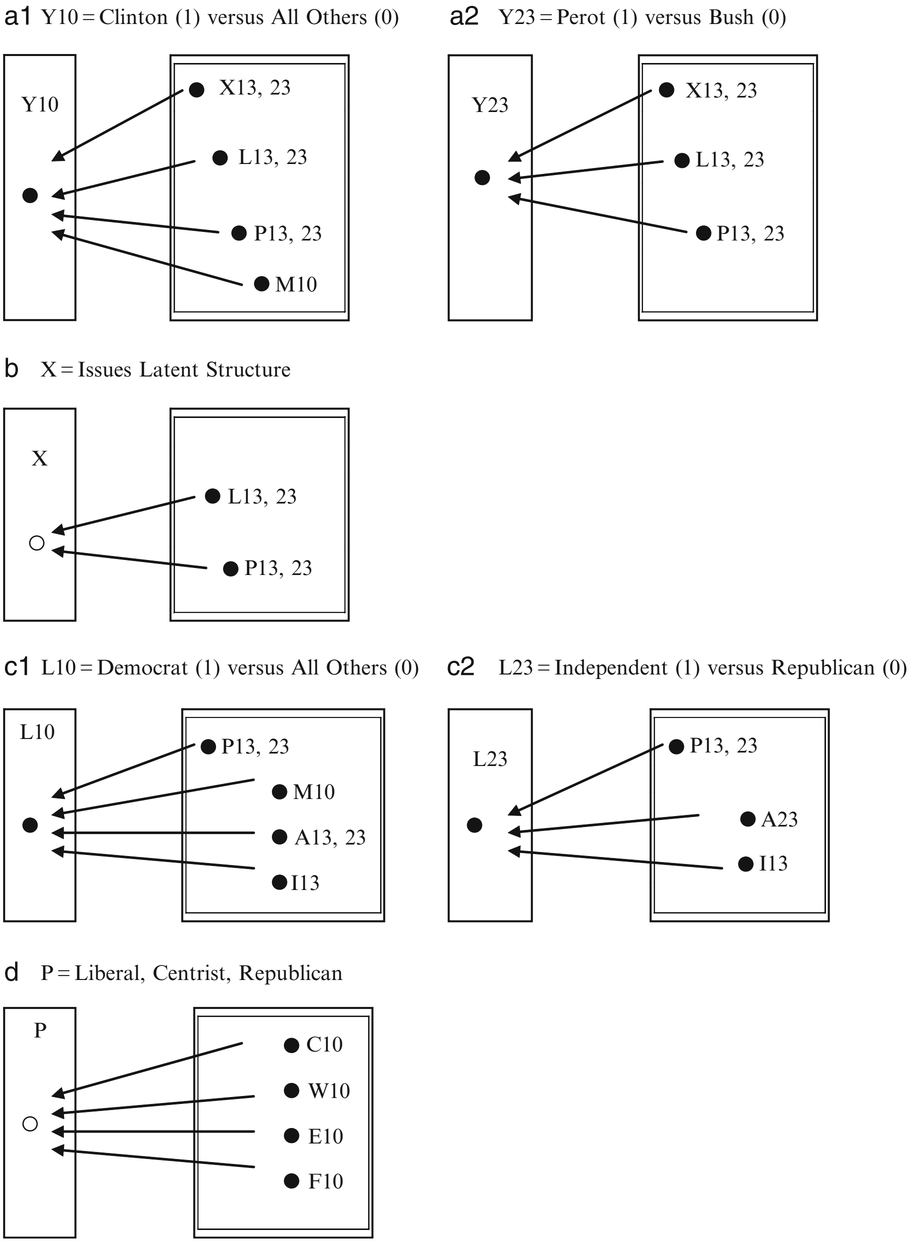
(continued)

Table 4.1 (continued)

(3) Determinants of party identification	Proportional odds (PO)		Democrats vs. All Others		Independent vs. Republican	
Democrat, Independent, Republican	B	Exp(B)	B	Exp(B)	B	Exp(B)
Intercept1	−2.42 (0.22)		−1.93 (0.22)			
Intercept2	−0.94 (0.21)				−1.13 (0.17)	
M10 Minority Ethnicity vs. White	1.92 (0.22)	6.8	1.93 (.23)	6.9	–	–
A13, Older age (50+) vs. Younger (18–29)	0.56 (0.17)	1.8	0.53 (0.20)	1.7	–	–
A23, Middle age (30–49) vs. Younger (18–29)	0.56 (0.17)	1.8	0.41 (0.20)	1.5	0.41 (0.17)[a]	1.5
I13, Low income (<$30,000) vs. High ($50,000+)	0.67 (0.16)	1.9	0.47 (0.14)	1.6	0.40 (0.18)[b]	1.5
I23, Middle income ($30,000 –$49,999) vs. High	0.29 (0.15)	1.3	–	–	–	–
P13, Liberals vs. Conservatives	1.67 (0.15)	5.3	1.37 (0.17)	3.9	1.73 (0.23)	5.6
P23, Centrists vs. Conservatives	1.12 (0.14)	3.1	0.81 (0.17)	2.2	1.23 (0.19)	3.4
Tests of fit:	PO $\chi^2 = 22.7$, df = 7, $p = 0.002$		H&L $\chi^2 = 2.8$, df = 7, $p = 0.91$		H&L $\chi^2 = 4.1$, df = 7, $p = 0.77$	
	Deviance/DF = 1.22, $p = 0.07$		Deviance/DF = 1.20, $p = 0.21$		Deviance/DF = 0.85, $p = 0.54$	
	BIC 2387.5 reduced to 2185.6		BIC 1471.7 reduced to 1334.3		BIC 919.8 reduced to 858	
	−2*LL* 2373.5 reduced to 2122.1		−2*LL* 1464.7 reduced to 1285.3		−2*LL* 913.3 reduced to 825.6	
(4) Determinants of political philosophy	Proportional odds (PO)		Liberal vs. All Others		Centrist vs. Conservative	
Liberal, Centrist, Conservative	B	Exp(B)	B	Exp(B)	B	Exp(B)
Intercept1	−1.60 (0.14)		−1.75 (0.16)			
Intercept2	−0.14 (0.13)				−0.17 (0.07)	
C10, Coastal Region vs. Midwest or South	0.28 (0.11)	1.3	0.41 (0.13)	1.5	–	–
W10, Women vs. Men	0.45 (0.11)	1.6	0.53 (0.13)	1.7	–	–
E10, Paid Work vs. Not Paid	0.43 (0.12)	1.5	0.58 (0.14)	1.8	–	–
F10, First-time voter vs. not first time	0.43 (0.17)	1.5	–	–	0.69 (0.24)	2
Tests of fit:	PO $\chi^2 = 9.8$, df = 4, $p = 0.044$		H&L $\chi^2 = 6.9$, df = 6, $p = 0.33$		H&L not applicable	
	Deviance/DF = 1.11, $p = 0.32$		Deviance/DF = 1.74, $p = 0.14$		Deviance/DF not applicable	
	BIC 2557.1 reduced to 2546.8		BIC 1405.1 reduced to 1389.8		BIC 1163.8 reduced to 1161.6	
	−2*LL* 2543 reduced to 2504.5		−2*LL* 1398.1 reduced to 1361.5		−2*LL* 1157 reduced to 1148.1	

Standard errors are enclosed in parentheses. A dash (–) indicates that the effect lacks statistical significance and was deleted from the model, which was then reestimated. Tests of fit: *PO* Proportional odds, *H&L* Hosmer & Lemeshow, *BIC* Schwarz's Bayesian Criterion, −2*LL* = 2 Log Likelihood.

[a] The base category is (Younger + Older), [b] The base category is (Middle + High Income).



a1 Y10 = Clinton (1) versus All Others (0)

a2 Y23 = Perot (1) versus Bush (0)

b X = Issues Latent Structure

c1 L10 = Democrat (1) versus All Others (0)

c2 L23 = Independent (1) versus Republican (0)

d P = Liberal, Centrist, Republican

Fig. 4.2 Results of logistic regression analyses depicted by regression graphs

Note: Solid circles depict binary (0,1) indicator variables; open circles depict ordinal variables.

indicators of substantive variables influenced the voting choice. Because the response variable is a binary indicator variable (i.e., a dummy variable) coded (0,1), it is portrayed as a solid circle. Because the four predictors are also coded here as binary (0,1) indicator variables, they are also portrayed by solid circles. The variables are arrayed so that those in the block closest to the response is on the left and those in blocks farthest from the response are on the right; the variables in intermediate blocks are located between these extremes. In graph (d), all of the background variables are on equal footing; so they are aligned vertically. Next to each circle is the descriptive label for the variable, and the coding pattern. In graph (a), X is the explanatory variable, the ordinal issue continuum. Its categories are coded 1 versus 3 and 2 versus 3, the standard coding pattern for indicator variables. Comparison of graph (a1) with graph (a2) shows that ethnic minorities tended to vote for Clinton over [Perot + Bush], but not for Perot over Bush.

Graph (b) portrays the latent structure of the issues as an ordinal variable, which it signifies by a circle. It shows that the indicator variables for party identification (loyalty = L) and for political philosophy (P) influenced the voters' positions on the issues.

In the third regression analysis, because the proportional-odds model did not hold, graph (c1) and graph (c2) depict, respectively, the vote for Clinton over the other candidates and the vote for Perot over Bush. These graphs clearly show that minorities did not support Perot, they supported Clinton. Additionally, both categories of age and income favored Clinton, but only the middle-aged and lower income people supported Perot over Bush.

In the fourth analysis, because the proportional-odds model was assumed to hold, graph (d) depicts political philosophy as an ordinal variable and relates it to four dichotomous background attributes: coastal residence, women, paid work, and first-time voter.

At this point, the big picture is obscure: how do these various relationships form a system? The dependence graph of Fig. 4.3 provides an answer; it is very similar to a summarizing path diagram. This graph simplifies the detailed findings by depicting the variables as either ordered trichotomies (circles) or as dichotomies (solid circles). It depicts the conditional dependencies by the arrows linking the blocks of variables; a missing edge between two variables in different blocks indicates that their relationship is conditionally independent. Reading the diagram from left to right, it shows that the latent classes of the issues directly determine the voting choice, as do party identification and political philosophy; additionally, minority ethnicity has a direct effect on vote. The latent classes of the issues are directly determined by the partisan predispositions of party identification and political philosophy, with the latter shaping the former, along with age and family income. Political philosophy is in turn shaped by area of residence, gender, employment, and first-time voting. Apparently, political philosophy is a pivotal variable: it directly influences party identification, positions on the issues, and the voting choice; change a person's political philosophy and this will change the person's party identification, positions on issues, and vote.

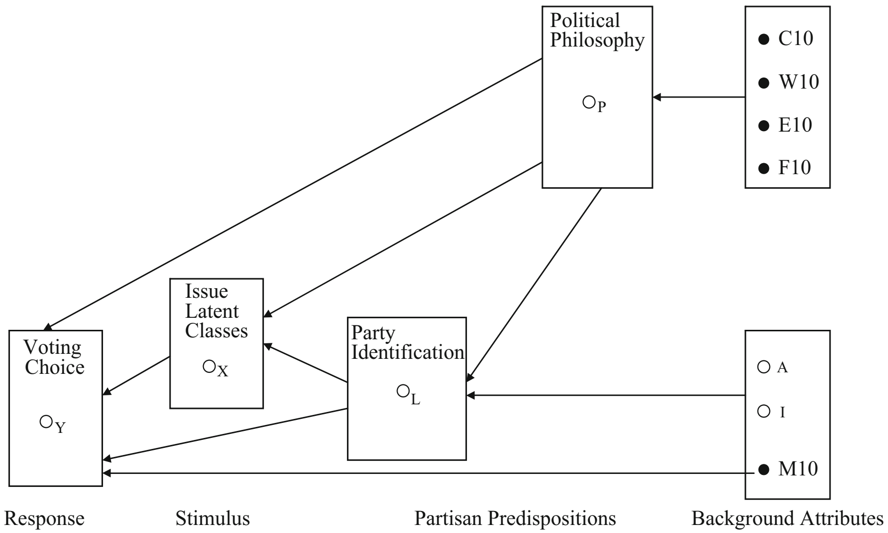


Fig. 4.3 Dependence graph for network of ordinal and dichotomous variables
Note: Ordinal trichotomies are depicted as circles ○ and dichotomies as solid circles ●. The codes for the background attributes are: C = coastal region; W = women; E = employed paid work; F = first-time voter; A = age category, I = family income category; and M = minority status.

Interpretive Edge Matrices

Facilitating interpretation, parental and ancestor edge matrices portray these dependencies (Wermuth 2003, 50–52). On the rows, the parental edge matrix of Table 4.2 lists the ordinal and dichotomous response variables (graphical modelers refer to these responses as children), and on the columns it lists the prior ordinal and dichotomous explanatory variables (graphical modelers refer to these as parents). A direct conditional dependency between a prior explanatory variable (a parent) and a response (a child) is indicated by 1; two variables that are conditionally independent of each other in the model are indicated by a 0, as are the relationships among the background variables in the same block. For vote (Y), the pattern of 1s indicates that the latent classes of the issues have direct influence, along with party identification, political philosophy, and minority ethnicity. The pattern of 0s indicates that coastal residence, women, employment, first-time voters, age, and income do not directly influence vote. Among the other interesting relationships are those between the social attributes and the variables of political partisanship. The patterns of 1s and 0s indicate that age, income, and minority ethnicity directly influence party identification but not political philosophy; however, coastal region, women, employment, and first-time voting directly influence political philosophy but not party identification. These two patterns enabled these social attributes to be used as instrumental variables in an earlier two-stage least-squares analysis that found that political philosophy is prior to party identification in these data (Smith 1999b).

Table 4.2 Parental edge matrix summarizing direct dependencies on prior variables

Priors		Y	X	L	P	C	W	E	F	A	I	M
Responses		○	○	○	○	●	●	●	●	○	○	●
Y	○	1	1	1	1	0	0	0	0	0	0	1
X	○		1	1	1	0	0	0	0	0	0	0
L	○			1	1	0	0	0	0	1	1	1
P	○				1	1	1	1	1	0	0	0
C	●					1	0	0	0	0	0	0
W	●						1	0	0	0	0	0
E	●							1	0	0	0	0
F	●								1	0	0	0
A	○									1	0	0
I	○										1	0
M	●											1

This matrix summarizes finding from Table 4.2 and Fig. 4.3 treating the variables as ordinals symbolized by circles ○ or as dichotomies symbolized by solid circles ●.
The codes are Y = vote, X = three-class latent structure, L = party identification, P = political philosophy, C = coastal region, W = women, E = paid employment, F = first-time voter, A = age, I = income, M = minority status.

Table 4.3 Ancestor edge matrix summarizing direct and indirect dependencies on prior variables

Priors		Y	X	L	P	C	W	E	F	A	I	M
Responses		○	○	○	○	●	●	●	●	○	○	●
Y	○	1	1	1	1	1	1	1	1	1	1	1
X	○		1	1	1	1	1	1	1	1	1	1
L	○			1	1	1	1	1	1	1	1	1
P	○				1	1	1	1	1	0	0	0
C	●					1	0	0	0	0	0	0
W	●						1	0	0	0	0	0
E	●							1	0	0	0	0
F	●								1	0	0	0
A	○									1	0	0
I	○										1	0
M	●											1

This matrix summarizes findings from Table 4.2 and Fig. 4.3 treating the variables as ordinals symbolized by circles ○ or as dichotomies symbolized by solid circles ●.
The codes are Y = vote, X = three-class latent structure, L = party identification, P = political philosophy, C = coastal region, W = women, E = paid employment, F = first-time voter, A = age, I = income, M = minority status.

The overall ancestor edge graph of Table 4.3 has a similar structure to that of the parental graph, but the ancestors of parents that influence the child indirectly through their influence on the parent are also indicated by 1s; no indirect or direct influence is indicated by 0. For vote (Y) all of the prior variables have either direct or indirect influence as indicated by the universal pattern of 1s. Moreover, variables

L through M directly or indirectly influence positions on the issues' latent structure, and variables P through M directly or indirectly influence party identification. The pattern of-relationships for the social attributes that influence political philosophy are the same as in the parental matrix of Table 4.2.

How best to quantify such direct and indirect dependencies in matrices of ordinal and dichotomous variables is a problem for further research. However, Smith (1972a) provides an example of a path analysis of ordinal variables based on Kendall's tau-b, and a number of alternative statistical methods can be applied. These include the assignment of equal-interval scales to the ordinals (e.g., Smith 1999b); the use of rank correlation methods in the Kendall's tau and Spearman-rho families of statistics (e.g., Smith 1978, 1985b, 1986); weighted least-squares methods based on Goodman and Kruskal's gamma or Somers's d_{yx} both using Mann-Whitney statistics (Carr et al. 1989 ; Stokes et al. 2000); structural equations and path analysis for discrete data (Winship and Mare 1983; Pearl [2000] 2009 second edition, 133–172); LISREL modeling based on polychoric and tetrachoric correlation coefficients (Jöreskog and Sörbom 1993, 44–50); and loglinear models (Goodman 1978). The latter are especially appropriate for detecting interaction effects among the variables, as the following example illustrates.

Association Graphs for Loglinear Models

Loglinear models can uncover relationships in multivariable contingency tables that are composed of dichotomous attributes and ordinal variables; association graphs can clarify the relationships that loglinear models detect. These graphs represent the factors (i.e., variables) of a loglinear model by vertices; the edges that connect two factors are the two-factor effects. Explicating the usefulness of such graphs, this section: (1) applies the backward selection method to identify the best-fitting loglinear models for a cross-tabulation, (2) uses association graphs to depict the key interactions among the variables, (3) summarizes the findings in a path-analytic diagram, and (4) quantifies the effects using logistic regression. Illustrating these procedures, it uses the same data set about the 1992 presidential election that the previous section modeled; ongoing research is testing these interactions in the 2008 election.

The Cross-Tabulated Variables

For the Clinton-Perot-Bush election, a cross-tabulation relates four measures of the issues with the final vote. Single questionnaire items gauge the trichotomous voting choice (Y) and the dichotomous environmental (N) and character (C) issues; valid two-item ordinal indices gauge the economic (E) and health care issues (H). The latent class model of the previous section is based on these measures of the issues; their separate effects are analyzed here.

The environmental issue (N)

This item gauges the environmental issue: the environment was a very important factor in determining which candidate to vote for (47%) versus it was not very important (53%). Responses indicating environmental concern have positive associations with agreeing that the president should protect the environment even if there is loss of some jobs (Kendall's tau-b $= \tau = 0.20$); with agreeing that a company's environmental record is important in forming an opinion about it ($\tau = 0.21$); and with the index composed of these two items ($\tau = 0.24$). Environmental concern is associated with left positions on the political continuum, namely, liberalism, Democratic identification, and vote, but not strongly with the character issue.

Character of the candidates (C)

About 52% said the character of the candidates was a very important factor in determining their choice of candidate; 48% said the opposite. Most likely, this question assessed the voters' perceptions of differences between Bush and Clinton. Because of the negative campaign waged by some Republicans against Clinton (they accused him of "slickness" and "waffling" on issues, adultery, draft evasion, and marijuana use) and the pro-family and pro-life campaign of the Republicans, those most concerned about the character of the candidates voted for Bush and those less concerned voted for Clinton; the Perot voters were in the middle. Those less concerned about character leaned toward the left: they were liberals and Democrats, and supported governmental interventions for economic equity, the environment, and health-care reform. Voters who favored universal access to health care (which included women's health services) were more likely to say that character was not an important determinant of their vote ($\tau = 0.11, p < .0001$). Apparently, the character issue in part reflects a candidate's position on women's choice. Since a vote for Clinton was given a plus (+) sign, the positive category (+) of this item about presidential character will be "not very concerned" and the negative category (−) will be "very concerned."

The economy (E)

Two items about presidential economic interventions measure the economic issue: "On the economy, should he concentrate on economic expansion and jobs even if that means a higher deficit (41%) or should he concentrate on first getting the deficit under control (59%)?" The other asked: "On regulation, should he concentrate on regulating industry to protect consumers (41%) or reducing regulation to make American businesses more competitive (59%)?" The first alternative answer to each question was the more liberal response; the second, the more conservative response. Their additive index assesses support for economic expansion and regulations; it classifies about 19% of the respondents as wanting both presidential economic interventions, 44% as wanting one, and 37% as opposing both interventions; the latter favored bringing the deficit under control and reducing regulation of businesses. This index has positive associations with support for governmental interventions about economic equity, social equality, and public health; liberalism; Democratic identification; and a vote for Clinton. Those on the right and those in the center tended to prefer a reduced deficit and less regulation of industry.

Health care reform (H)

In 1992, as in 2008, competitive health care plans were salient: Then, the political left proposed a national health care system similar to Canada's; in the center, Clinton endorsed "managed competition," but stipulated that the plan must provide universal access and limits to spending; on the right, Bush offered his voucher-based, private-sector plan. A three-category typology taps this reform continuum: those who favored comprehensive reform trusted federal participation and desired radical change (+ + = 29%); those who supported some reform wanted one aspect of reform (+ − or − + = 39%); and those who opposed comprehensive reform rejected both aspects (− − = 32%). The left supported comprehensive reform ($\tau = 0.28$): pro-reformers were more likely than anti-reformers to support governmental interventions aimed toward economic equity, social equality, and the public's health. Compared with the political right, liberals, Democrats and voters for Clinton were more likely to support reform ($p < .001$); Independents and Perot voters held intermediate positions. Compared with the anti-reformers, the pro-reformers were more likely to say that a candidate's character was *not* very important in determining their vote.

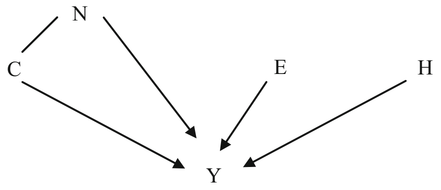
A Loglinear Model

The backward selection procedure starts with the highest level interaction term, here it is EHNCY, and successively eliminates from the model insignificant interactions. After ten steps, the hierarchical loglinear module of SPSS found this best-fitting model ($\chi^2 = 45.3$, $df = 56$, $p = 0.85$): (EHNC), (NCY), (EY), (HY). The four-way interaction (EHNC) interrelates the four explanatory variables, the three-way interaction (NCY) suggests that N and C directly influence Y and that there is an association between N and C. The two-way terms suggest that E and H have independent effects on vote.[2]

Association Graphs

Putting aside the interrelations among the four explanatory variables and focusing first on their effects on vote, this association graph depicts the above relationships

(Goodman [1972b] 1978, 57–109; Agresti 1996, 174–180; Christensen 1997, 178–210):



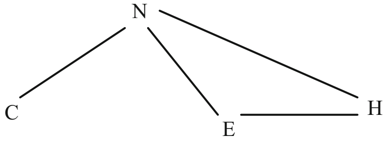
On the basis of the meaning and time ordering of the variables, a person's vote (Y) is the response variable. It is directly influenced by her positions on the issues of the campaign, namely, a candidate's character (C), the environment (N), the economy (E), and health care reform (H).[3] The arrows express the direct effects of the issues on the voting choice. The solid line between the character (C) and the environmental (N) issues along with their effects on vote indicate that these issues interact; the (CNY) interaction is statistically significant. If it is deleted from the model (i.e., the null hypothesis is H_0: no significant effect), then the gain in degrees of freedom (*df*) would be 2, the change in χ^2 would be 8.54, and the probability that an effect this large could happen by chance would be $p = 0.014$ (H_0 is rejected decisively). In comparison, the (EHY) interaction is not statistically significant. When it is deleted from the model, the gain in *df* would be 8, the χ^2 change would be 10.53 and $p = 0.230$ (H_0 is not rejected); thus, the fit of the model is not significantly improved by the inclusion of this three-way interaction.

The interaction effects of the character and the environmental issues on vote appear when these variables are cross-tabulated and measures of association are calculated for the conditional tables. When the environment is a very important determinant of one's vote (+), the association between saying that character is very important (−) and Republican vote (−) was much lower than that association when the environment was not very important (−). Using Somers's d_{yx} and Kendall's τ to measure the asymmetric and symmetric associations between these ordinal variables, the coefficients are as follows: When the environment is very important, $d_{yx} = 0.19$ and $\tau = 0.18$; when the environment is not very important $d_{yx} = 0.42$ and $\tau = 0.37$. Thus, being pro-environment weakened the effect of concern about Clinton's character on Republican vote.

These same cross-tabulations show that the effect of the unimportance of environmental issue (−) on Republican vote (−) was stronger when Clinton's character was considered important (−), compared with that association when Clinton's character was considered to be unimportant (+). When character was a very important consideration, $d_{yx} = 0.34$ and $\tau = 0.30$; when character was not very important $d_{yx} = 0.13$ and $\tau = 0.12$. Thus, when the Democrats successfully countered attacks on Clinton's character, the effect of lack of environmental concern on Republican vote weakened.

The four-variable interaction (EHNC) contains only a few statistically significant component effects, namely, (NEH) ($df = 4$, $\chi^2 = 10.73$, $p = 0.03$), (NH) ($df = 2$, $\chi^2 = 10.06$, $p = 0.07$), and (NC) (*df* 1, $\chi^2 = 6.6$, $p = 0.01$). Of these two-way interactions, (NH) is a component of (NEH) and (NC) is a component of

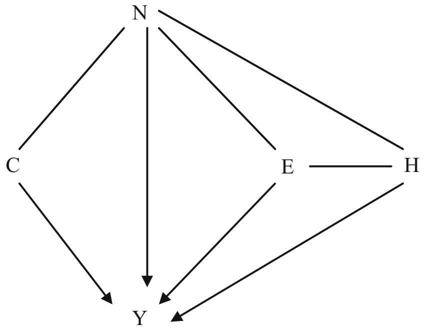
(CNV), which was studied earlier. Because parsimonious models are preferred to the more complex, it is appropriate to re-estimate the earlier best-fitting model, deleting from it (EHNC) and instead including the simpler three-way interaction (NEH). The resulting model has this structure (NEH), (NCY), (EY), (HY); this parsimonious model indeed fits rather well: $df = 72$, $\chi^2 = 74.14$, $p = 0.41$. The following association graph focuses on the interrelationships among the explanatory variables in this simpler model:



The straight lines indicate relationships that are associational. C is linked to N via the earlier three-way interaction (NCY), and N, E, and H are linked via (NEH). The latter suggests that the relationship between supporting interventions in the economy (E) and supporting health care reform (H) varies with the pivotal variable, the environmental issue (N). The cross-tabulation of E and H, controlling for N, indicates the following: When the environment is very important (+), then the correlation between not favoring governmental interventions in the economy (−) and not favoring health care reform (−) is lower (Kendall's $\tau = 0.14$) than when the environment is not very important ($\tau = 0.25$). Apparently, environmental concern (+) reduces the consistency of opposition to governmental interventions in the economy (−) and opposition to health care (−). But the absence of environmental concern (−) enhances the consistency of opposition to governmental interventions in the economy (−) and opposition to health care (−).

The synthesized model

This association graph synthesizes the two graphs depicted earlier:



It suggests these relationships: (NCY), (NEH), (EY), (HY) and the three-way interaction (EHY) that is not statistically significant. The relevant two-way associations are (CY), (NY), (EY), (HY), (NC), (NH), (EH), and (NE); the latter is not statistically significant ($df = 2$; $\chi^2 = 3.39$, $p = 0.18$), but it cannot readily be deleted from the model because it is included in (NEH).

Quantification of effects

Because the four issues directly influence the voting choice, a logistic regression model that assumes asymmetric effects can provide appropriate estimates of effects (Goodman [1972a] 1978, 7–25). In these models, the character variable is coded so that concern about character (−) reduces the tendency to vote for Clinton (+) and the environmental, health care reform, and economic intervention issues are coded (+) so that favorable index scores enhance the tendency to vote for Clinton. On the basis of the loglinear analysis and the association graphs, the character × environment interaction is expected to have a positive effect on voting for Clinton, any other interactions among the issues are expected to be negligible, and they are.

Table 4.4 presents the results of logistic regressions for the dichotomous choice, Clinton or Bush (Perot voters are deleted), and the ordinal choice, Clinton, Perot, or Bush. In both sets of data, the direct effects on vote are consistent with expectations derived from the association graphs; the character issue has a negative effect on voting for Clinton. However, when the character × environment interaction is introduced into the model, it has a strong positive effect. Thus, the environmental issue weakened the effect of negative aspersions about Clinton's character. Separate analyses indicate that the other two-way interactions among the issues are not statistically significant.[4]

The reasons for this positive interaction effect that weakens the negative effect of a candidate's character on Democratic vote are speculative, but may be interpreted as follows: Some evangelical Christians are both anti-women's choice concerning abortions and also believe that human beings are the custodians of God's earth, and that we humans should take good care of it; they are pro-environment, at least implicitly. Clinton favored women's choice; evangelical Christians deemed this stance among others of his to be a character flaw. However, Gore and Clinton favored protection of the environment. This created a cross-pressure that may have moderated the anti-Clinton fervor of evangelical Christians.[5] Ongoing research on the 2008 election is testing this interpretive mechanism.

Generative Mechanisms

Generative mechanisms typically link a variable of social structure x to a response y via a series of theoretically relevant intervening variables that form a chain of relationships: $x \rightarrow t_1 \rightarrow t_2 \rightarrow y$; for example, neighborhood poverty (x) influences family social class (t_1), which influences their children's intellectual achievement (t_2), which influences the children's college plans (y). The following

Table 4.4 The perception of a candidate's character interacted with a person's environmental concern, 1992 election data

	Clinton vs. Bush		Clinton–Perot–Bush	
	Direct effects only	Interaction effect added in	Direct effects only	Interaction effect added in
Variables				
Intercept 1	−0.873	−0.635	−1.282	−1.106
	$p < .0001$	$p = .0079$	$p < .0001$	$p < .0001$
Intercept 2	–	–	−0.164	0.021
			$p = 0.325$	$p = 0.906$
Character	−1.325	−1.786	−1.032	−1.364
	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$
Environment	1.026	0.437	0.874	0.480
	$p < .0001$	$p = 0.106$	$p < .0001$	$p = 0.015$
Character × Environment	–	1.045	–	0.747
		$p = .0042$		$p = .006$
Economy 1	1.411	1.421	1.276	1.275
	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$
Economy 2	0.751	0.75	0.567	0.553
	$p < .0002$	$p = .0002$	$p < .0001$	$p = .0002$
Health care 1	1.403	1.381	1.047	1.015
	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$
Health care 2	1.016	1.009	0.833	0.817
	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$
Measures of fit				
Deviance	24.9	16.8	67.81	60.2
	$p = 0.681$	$p = 0.953$	$p = 0.349$	$p = 0.576$
AIC	757.2	763.3	1700.2	1694.6
BIC	793.8	795.4	1738.7	1737.9
$-2LL$	741.2	749.3	1684.2	1676.6
Proportional odds Assumption	–	–	$p = 0.018$	$p = 0.034$

The proportional odds assumption is rejected, but the data for the dichotomized response support the notion of a statistically significant interaction effect.

reasons justify the ordering of these variables: A child's intellectual achievement (t_2) is a more general construct than her specific college plans (y), so the former is thought to be prior to the latter.[6] Parental social class (t_1) is a variable of social structure that is prior in time to the child's intellectual achievement (t_2), so family social class is thought to be prior to the child's intellectual achievement. Neighborhood poverty (x) is a more macrovariable than the social class of the families who live in the neighborhood (t_1); the latter is thought to provide a mechanism through which neighborhood poverty (x) affects the child's intellectual achievement (t_2).

Hypothesized generative mechanisms are flawed if they include variables that are not truly independent constructs or if they exclude theoretically important intervening variables. For example, what variables intervene between a student's intellectual achievement (t_2) and her college plans (y)? Cicourel and Kitsuse (1963) hypothesize that school administrators differentially label students on the basis of their social background and then assign them to different tracks, college going or not; it is the tracking of the students by guidance counselors and teachers that is a proximate influence on her college plans. Assuming that this hypothesis holds, one could introduce into this model a variable $t_{3,}$ "a student's low expectations about success in college due to the low expectations of the school staff." This variable is thought to moderate the relationship between her intellectual achievement (t_2) and college plans (y); thus, $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow y$.[7]

An investigator can circumvent criticism of such models by grounding them in first-hand qualitative observations that trace how the social process being studied unfolds. On the basis of these data, the investigator can develop a theoretically and experientially informed questionnaire that includes indicators of the relevant constructs. When the survey data are gathered and the measures constructed, statistical methods can be applied to quantify and test the hypothesized direct and indirect relationships. There are few social scientific examples of generative processes that are based on cumulative research studies.[8] Addressing this gap, the following section provides a detailed example. Using survey data on undergraduate exchange students, whose home universities are the Massachusetts Institute of Technology (MIT) or the University of Cambridge in the UK, it probes the sources of their consideration of novel uses for technology. This exposition is based on collaborative research the author conducted with William A. Lucas at the Cambridge-MIT Institute; the empirical measures were jointly developed.

Consideration of Use

Consideration of use is an important component of what Donald Stokes (1997, 73) refers to as use-inspired basic research, which he distinguishes from basic and applied research. Basic research is primarily motivated by the investigator's curiosity and quest for fundamental understanding, whereas applied research is primarily motivated by utilitarian goals, with little or no emphasis on developing a fundamental understanding or theory of the phenomena. In use-inspired basic research, the investigator tackles a pressing practical problem, but also seeks a fundamental understanding that can be expressed in a theory that explains the results. Here are some examples (Stokes 1997, 79–80): "Pasteur wanted to understand *and to control* the microbiological processes he discovered. Keynes wanted to understand *and to improve* the workings of modern economies. The physicists of the Manhattan Project wanted to understand *and to harness* nuclear fission. Molecular biologists have wanted to understand *and to alter* the genetic codes in DNA material."

Table 4.5 Five indicators of consideration of use

How often does it happen that you…	Almost never occurs	Once a month or less	Several times a month	Once or twice a week	Almost daily, or more often
G_1 Realize while thinking about a problem that there is a technology that could be used in a new way to provide a solution	1	2	3	4	5
G_3 Learn a new applied science concept and get excited about an application idea (whether or not the idea was right)	1	2	3	4	5
G_4 Wonder while you are in class or a lab whether something you just learned could be used to improve a product or process	1	2	3	4	5
G_6 As you learn about a principle, you realize on your own that there are special cases when the principle does not hold up	1	2	3	4	5
G_7 Think about some social problem or need that could be addressed by something you are studying	1	2	3	4	5

These questionnaire items were created by William A. Lucas at the Cambridge-MIT Institute.

The students' consideration of use is indicated by their responses to the five items of Table 4.5, which Lucas created.[9] These questions ask how often the respondent experiences sudden insights about applying one's knowledge to solve a problem. The manifestations of such insights are akin to exclaiming "eureka!" (e.g., the first and second items) or to saying "a-ha" (e.g., the third to fifth items). The five response categories range from "almost never occurs" to "almost daily, or more often." The index formed from these questions has a reliability of $\alpha = 0.83$ and has validating correlations with measures of a deeper intellectual understanding ($r = 0.48$, $p < .0001$), communication skills ($r = 0.45$, $p < .0001$), leadership skills ($r = 0.25$, $p = 0.007$), pre-entrepreneurial behavior ($r = 0.59$, $p < .0001$), technical self-efficacy ($r = 0.40$, $p < .0001$), and venturing self-efficacy ($r = 0.32$, $p = 0.0003$).

The consideration of use of these exchange students is rooted in differences between the undergraduate pedagogy of MIT and Cambridge: At Cambridge, a supervisor (i.e., a tutor) meets every few weeks with two or three students answering their questions about the lectures and homework problems, which are not graded explicitly. At the end of the academic year, the students take a demanding comprehensive examination that determines their grade. At MIT, students turn in graded homework assignments and examinations throughout their courses, which are demanding. Moreover, MIT encourages its undergraduates to participate in professors' research projects as early as their first year, whereas, at the time of these surveys (fall of 2004 and fall of 2005), Cambridge students did not have such research opportunities until after their third year. At baseline, about 70% of the MIT exchange students reported undergraduate research experience, compared with only 10% of the roughly matched Cambridge students. Participation in research projects engenders insights about applying scientific knowledge to solve practical problems; the MIT students exhibit somewhat higher levels of consideration of use than the Cambridge students, $r = 0.14$.

A Generative Process Model

The path diagram of Fig. 4.4 links a student's consideration of use to the background variables gender and home university (the latter variable is at level-2, all other variables are at level-1) via a generative mechanism composed of three elements: self-reported participation in research → self-rated confidence in one's basic research skills → self-rated innovative ability. Thus: social background → generative mechanism → student outcome. Cumulative research studies have examined in depth each linkage in this chain of relationships; these studies are described later. The path coefficients are fully standardized so that a change of one standard deviation above the mean of zero of an explanatory variable brings about $\beta \times$ standard deviation units above the mean of zero on the response variable, controlling for the other covariates in the equation.

The model-fitting program Analysis of Moment Structures (AMOS), which implements Jöreskog and Sörbom's (1979) statistical theory via an easy-to-use graphical interface, provides the estimates of the effects, goodness-of-fit measures,

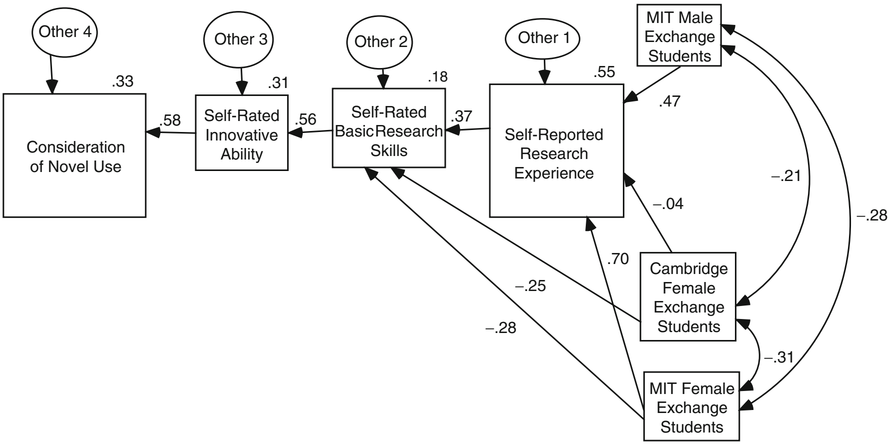

Other 4
Other 3
Other 2
Other 1
MIT Male Exchange Students
.33
.31
.18
.55
.47
Consideration of Novel Use
.58
Self-Rated Innovative Ability
.56
Self-Rated Basic Research Skills
.37
Self-Reported Research Experience
−.21
−.28
−.04
.70
−.25
−.28
Cambridge Female Exchange Students
−.31
MIT Female Exchange Students

Fig. 4.4 A generative process model producing consideration of use, standardized coefficients

and levels of statistical significance (Arbuckle and Wothke 1999; Arbuckle 2005). Guided by substantive considerations and various goodness-of-fit measures (Kline 2005), a series of recursive and nonrecursive models were estimated and the better-fitting models selected. Three such measures provide an overview of the fit: the Akaike Information Criterion (AIC); the Bayesian Information Criterion (BIC), which selects parsimonious models; and the model χ^2 test.[10] By examining the χ^2 due to the discrepancies between the predicted and actual covariances, it tests the null hypothesis that the model fits the actual covariance structure; probabilities less than or equal to 0.05 reject this null hypothesis (Kline 2005, 135–137). Competitive models that exhibit smaller AIC and BIC values, and model χ^2 probabilities of fit (p) considerably greater than 0.05, are preferred. Table 4.6 presents the correlations among the variables, their standard deviations, and their means—all variables are coded so that their measures range from 0 to 1. Consistent with the substantive ordering of the variables, Table 4.6 shows that a direct determinant of a variable (i.e., a parent) has a higher correlation with that variable than variables that are not direct determinants. The process begins with the background variables.

Background variables

Reading from right to left in Fig. 4.4, the path diagram of the basic model portrays three curved two-headed arrows that express the unanalyzed correlations among the three binary-coded background variables; Cambridge males are the base category for these indicator variables. A typical table cross-tabulates MIT males (1 = 20 students) or (0 = 68 students) with MIT females (1 = 36 students) or (0 = 68 students). Although no students are in the cell for MIT males = 1 and MIT females = 1, both binary variables have in common 68 Cambridge students who are not MIT males (0) or not MIT females (0). These cases are responsible for the nonzero correlation between these dichotomous variables. A similar logic explains the other correlations among these types of students.

Research experience

The home university and gender shape a student's score on the index of self-reported research experience, which is reliable (Cronbach's alpha = α = 0.81). The average of the responses to the following three questions forms this index. "At the university you currently attend, have you:"

> Worked as part of a research team (2 through 6 = 1, almost never = 0).
> Worked on a research project with professors or academic staff (yes = 1, no = 0).
> Talked to any professors or faculty members about working on a research project (yes = 1, no = 0).

Table 4.6 Correlations, standard deviations, and means for analysis of consideration of use

	Rowtype_	Varname_	Conuse5	Srinnoab	Srbasicr	Resexp3	Females	Mitone	Mitmales	Mitfems	Camufems	Camumale
1			124.00	124.00	124.00	124.00	124.00	124.00	124.00	124.00	124.00	124.00
2	Corr	Conuse5	1.00									
3	Corr	Srinnoab	0.58	1.00								
4	Corr	Srbasicr	0.40	0.56	1.00							
5	Corr	Resexp3	0.25	0.13	0.30	1.00						
6	Corr	Females	−0.13	−0.26	−0.22	0.24	1.00					
7	Corr	Mitone	0.14	−0.02	0.15	0.74	0.29	1.00				
8	Corr	Mitmales	0.14	0.12	0.19	0.28	−0.42	0.48	1.00			
9	Corr	Mitfems	0.04	−0.12	0.01	0.58	0.66	0.70	−0.28	1.00		
10	Corr	Camufems	−0.21	−0.19	−0.29	−0.36	0.51	0.44	−0.21	−0.31	1.00	
11	Corr	Camumale	0.03	0.18	0.08	−0.47	−0.72	−0.67	−0.33	−0.47	−0.36	1.00
12	Stddev		0.23	0.21	0.20	0.41	0.50	0.50	0.37	0.46	0.40	0.48
13	Mean		0.39	0.49	0.54	0.37	0.48	0.45	0.16	0.29	0.19	0.35

conuse5 = five-item measure of consideration of use, srinnoab = self-rated innovative ability, srbasicr = self-rated basic research skills, resexp3 = self-reported research experience, females, mitone = binary variable for MIT vs. Cambridge home university, mitmales = MIT male students, mitfems = MIT female students, camufems = Cambridge University female students, camumale = Cambridge University male students.

Compared with Cambridge males, the base category for the comparisons, MIT females ($\beta = 0.70$) and MIT males ($\beta = 0.47$) report more research experience; Cambridge females report slightly less ($\beta = -0.04$). This latter difference is not statistically significant and the arrow could be eliminated from the model, but this would change the reference category to all Cambridge students, which would obscure some findings. The variance explained by the three types of students is quantified by the $R^2 = 0.55$, which appears above the box for self-reported research experience. The residual path coefficient is $(1 - 0.55)^{½} = 0.45^{½} = 0.67$.

Basic research skills

The average of the following two indicators forms a normally distributed, reliable ($\alpha = 0.68$), and valid index of self-rated confidence in one's basic research skills; how able are you to:

> "Develop your own original hypothesis and research plan to test it" (0 to 10);
> "Understand exactly what is new and important in a ground-breaking theoretical article" (0 to 10).

The validity of this index (which was rescaled from 0 to 1) is suggested by its significant correlations with other aspects of confidence in one's research skills, being able to: "set up a demonstration of a basic scientific principle" ($r = 0.60$, $p < .0001$); "repeat a procedure described in a detailed research report to replicate its findings" ($r = 0.60$, $p < .0001$); "grasp the concept and limits of a technology well enough to see the best way to use it" ($r = 0.57$, $p < .0001$); "learn a new method or technique that you need to conduct an unfamiliar line of research" ($r = 0.50$, $p < .0001$); and "when reviewing the literature recognizing an important implication not mentioned in the readings" ($r = 0.38$, $p = 0.003$).[11]

Self-reported research experience has a strong positive effect on self-rated basic research skills ($\beta = 0.37$) but female students, either from MIT ($\beta = -0.28$) or from Cambridge ($\beta = -0.25$) exhibit strong negative effects (these effects are relative to male students). The explained variance $R^2 = 0.18$ and the residual path coefficient is $0.82^{½} = 0.905$. Why young women, when differences in reported research experience are controlled, still report less confidence in their basic research skills than their male colleagues is an unsolved problem for future research.[12]

Innovative Ability

Self-rated innovative ability has two aspects: creating novel designs and solving unstructured problems. The responses to these two items, which may range from 1 (poor) to 4 (excellent), indicate a student's self-rated ability to create a novel design:

> Design something novel and innovative.
> Create novel solutions to problems.

The responses to the following two items, which also may range from 1 (poor) to 4 (excellent), indicate a student's self-rated ability to solve unstructured problems:

> Solve an unstructured problem (that is, one for which no single right answer exists).
> Develop several methods that might be used to solve an unstructured problem.

The average of the sum of the responses to the four items, rescaled to range from 0 to 1, forms the index of self-rated innovative ability ($\alpha = 0.79$).

The path diagram for the basic model shows that self-rated confidence in one's basic research skills directly influences self-rated innovative ability ($\beta = 0.56$); the effects of the prior variables operate through confidence in basic research skills. The $R^2 = 0.31$ and the residual path coefficient is $0.69^{½} = 0.83$.

Self-rated innovative ability—being able to solve unstructured problems and create innovative designs—directly determines consideration of use ($\beta = 0.58$). The $R^2 = 0.33$ and the residual path coefficient is $0.67^{½} = 0.82$.

Figure 4.5 presents the structure of the basic model depicting its parameters as unstandardized *b* path-regression coefficients instead of fully standardized β path coefficients. These *b* coefficients suggest that a unit change (from 0 to 1) in a predictor brings about a $b \times 1$ unit change in the response, controlling for prior variables. Following the convention for reporting unstandardized coefficients, it reports covariances among the exogenous variables instead of correlations, and the variances of the extraneous and background variables instead of R^2s. As expected, the relationships are very similar to those reported earlier in Fig. 4.4. Except for the insignificant effect of Cambridge females compared with Cambridge males on research experience, all of the other effects in this model are substantial and statistically significant; it fits the data rather well: the model χ^2 $p = 0.23$, the AIC = 48.85 and the BIC = 99.6. If the insignificant path is removed and the model re-estimated, then the fit of this alternative model is improved only very

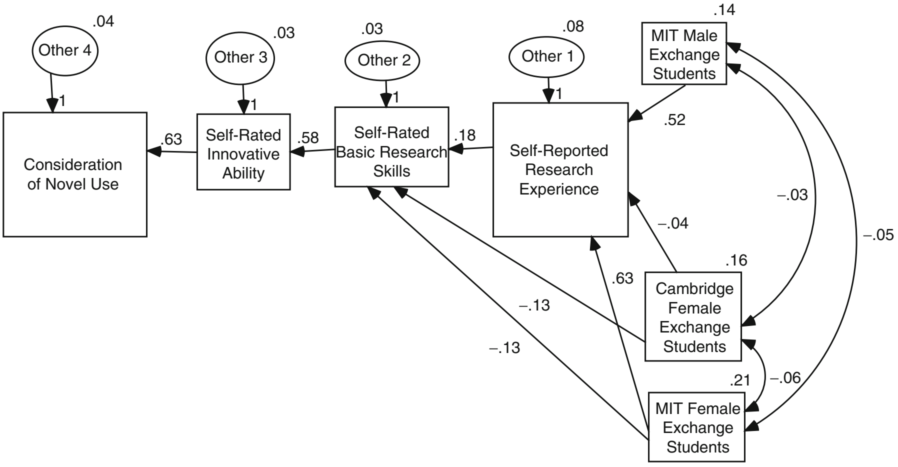

.04
Other 4
1
Consideration of Novel Use
.63
.03
Other 3
1
Self-Rated Innovative Ability
.58
.03
Other 2
1
Self-Rated Basic Research Skills
.18
.08
Other 1
1
Self-Reported Research Experience
.14
MIT Male Exchange Students
.52
−.03
−.05
−.04
.16
Cambridge Female Exchange Students
.63
−.13
−.13
.21
−.06
MIT Female Exchange Students

Fig. 4.5 A generative process model producing consideration of use, unstandardized coefficients

slightly: the model χ^2 $p = 0.28$, the AIC = 47.25 (this difference is less than the benchmark value of 2 for a significant difference), and the BIC = 95.2[13].

Correlations or Multivariate Dependencies?

The models described here are based on correlations among the variables that were measured on the students at one point in time; the models merely relate one associational property of a student to other associational properties, controlling for the effects of prior variables. Causal interpretations of these effects are risky because the evidence for causality lacks results from studies that explicitly assigned the students at random to treatment and control groups, and no experimental manipulations created the changes in the response variables. However, the models do trace a developmental sequence and quasi-experimental and other evidence supports the ordering of these variables, as explicated next, one linkage at a time.

Institutional practices influence an undergraduate's research experience

At the time of these surveys, Cambridge and MIT had different rules, programs, and customs regarding undergraduate research experience. MIT encourages its undergraduates to participate in UROPs (undergraduate research opportunities) whereas Cambridge did not. This difference in pedagogy at the two universities determines the level of research experience that the students reported. Moreover, after the MIT exchange students spend a year at Cambridge, their reported research experience for that year declines ($p = 0.0008$), whereas after the Cambridge exchange students spend a year at MIT, their reported research experience for that year increases ($p = 0.0004$). So, the first link in the chain is substantiated: the institutional differences in pedagogy between the two universities influence the difference in research experience the students report.

Research experience influences confidence in basic research skills

Because the home university influences the students' level of reported research experience, it becomes more plausible to assume that research experience directly influences confidence in research skills than it is to assume that confidence in research skills influences the level of research experience. The latter relationship assumes that students with more confidence in their basic research skills choose to acquire more research experience, but these data exhibit effects that are rooted in the universities' different pedagogies. Moreover, women report more research experience than men but exhibit less confidence in their research skills. If confidence in one's basic research skills determines the level of research experience, then men would report more research experience than women, which they do not. The university does not directly influence confidence in basic research skills when research experience is held constant. This implies that confidence in one's basic research skills is produced in part through this path: university → research experience → confidence in research skills; the latter is a specific manifestation of research experience, which is the more general of these two constructs.

Several alternative models test these conjectures. The first assumes that confidence in one's basic research skills reciprocally interacts with research experience. For this alternative model, AMOS produces stable estimates of the reciprocal effects but these effects are not statistically significant, and the overall model does not fit the data as well as the basic model reported in Figs. 4.4 and 4.5. For example, the BIC $= 103.9$ for this alternative model compared with the smaller BIC $= 99.6$ for the basic model. The second alternative model simply reverses the direction of the arrow between basic research skills and research experience. AMOS produced a range of fit measures that favor the basic model over this alternative model, respectively: model χ^2 probability, $0.23 > 0.18$; noncentrality parameter,[14] $2.85 < 3.88$; RMSEA,[15] $0.048 < 0.056$; AIC, $48.85 < 49.88$; and BIC, $99.6 < 100.64$. Thus, in these data, it is better to assume that research experience directly influences confidence in basic research skills, rather than the opposite. However, gender differences are significant. Women exhibit less confidence in their basic research skills than men do, even when the level of research experience is held constant. So, gender and research experience both influence the level of confidence in basic research skills, as depicted in Figs. 4.4 and 4.5.[16]

Confidence in basic research skills influences self-rated innovative ability

The basic models reported above assume that the student's self-rated innovative ability to create novel designs and to solve unstructured problems is driven by their confidence in their basic research skills, which is driven by their research experience, which varies according to their university's pedagogical practices. The simplest test of this generative sequence is to reverse the direction of an arrow so that self-rated innovative ability, conceptualized as a psychological disposition, is thought to shape confidence in one's basic research skills, and then to estimate the fit of the resulting structure using AMOS. This new model does not fit the data as well as the basic model: Both the AIC and BIC are now considerably higher (smaller is better) than for the reported models: AIC $58.8 > 48.85$ and BIC $109.6 > 99.6$; moreover the fit probability is near zero, $\chi^2\ p = 0.01$, compared with $\chi^2\ p = 0.23$ (larger is better).

A model that is strongly rooted in psychology would assume that consideration of use is a key psychological schema that drives self-rated innovative ability, which reciprocally interacts with confidence in basic research skills. To estimate the parameters of this model, instrumental variables are needed that determine one but not the other of the pair of reciprocally interacting variables. In this new psychological model, the instrumental variables are consideration of use, which is assumed to influence self-rated innovative ability but not confidence in basic research skills, and Cambridge and MIT women, whose experience and predispositions influence confidence in basic research skills but not self-rated innovative ability. AMOS estimated the parameters of this model: the reciprocally interacting variables are stable, but the resulting model fits the data poorly; χ^2 $p = 0.007$ compared with χ^2 $p = 0.23$ for the reported models. The AIC and BIC are much larger (smaller is better): AIC, $60.8 > 48.85$ and BIC, $114.4 > 99.6$. Consequently, it is better to assume that a student's confidence regarding basic research skills shapes self-rated innovative ability, rather than to assume the opposite. Apparently, confidence in research skills is the more general construct; self-rated innovative ability is a manifestation of confidence.

Self-rated innovative ability influences consideration of use

The thrust of this analysis assumes that consideration of use is a malleable characteristic of a student that is rooted in his or her theoretical and practical education. Change the student's education and research experience and this will change confidence in basic research skills, which will change the student's self-rated innovative ability, which will change consideration of use.[17] The alternative explanations de-contextualize these students from their university background by assuming that psychological dispositions independent of educational background are the key drivers. Thus far, this view produces models that do not fit these data on exchange students—that similar models may fit other data is a question for further empirical research. Now, a crucial test of these two perspectives involves the relationship between consideration of use and self-rated innovative ability. Does the alternative model, which assumes that consideration of use shapes innovative ability, fit better or worse than the basic model, which assumes that innovative ability shapes consideration of use? The basic model fits the data much better than this alternative model: The χ^2 probability level for this alternative model is zero, compared with χ^2 $p = 0.23$ for the basic model. The AIC and BIC for the alternative model are much larger than those for the basic model: AIC, $63.1 > 48.85$ and the BIC, $119.1 > 99.6$. Thus, in these data, self-rated innovative ability shapes consideration of use, thereby supporting the view that institutional differences in pedagogy are a key determinant of research capabilities and considerations of use of these roughly matched exchange students.[18]

A Test of the Basic Model

The computer output from AMOS indicates that a direct path toward consideration of use was left out of the reported models. On theoretical grounds, a direct path from self-reported research experience to consideration of use would buttress the institutional perspective embedded in the process model, which holds that research skills and experience are prior to consideration of use.[19] Figs. 4.6 and 4.7, respectively, present the standardized and unstandardized coefficients for the elaborated

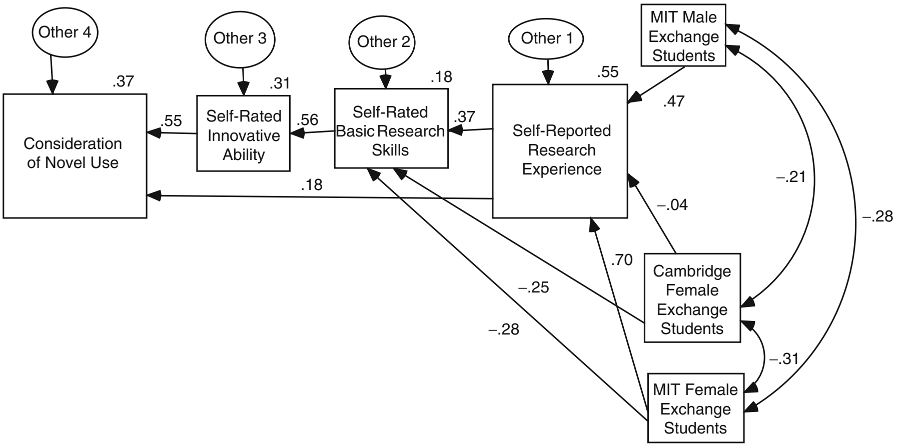

Other 4
Other 3
Other 2
Other 1
MIT Male Exchange Students
.37
.31
.18
.55
.47
Consideration of Novel Use
.55
Self-Rated Innovative Ability
.56
Self-Rated Basic Research Skills
.37
Self-Reported Research Experience
.18
−.21
−.04
−.28
.70
Cambridge Female Exchange Students
−.25
−.28
−.31
MIT Female Exchange Students

Fig. 4.6 Adding a direct effect of self-reported basic research experience on consideration of use, standardized coefficients

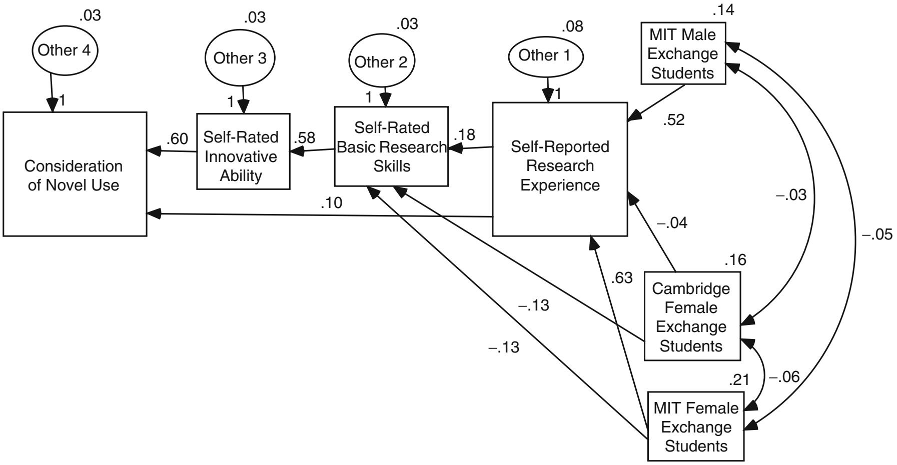

.03
.03
.03
.08
.14
Other 4
Other 3
Other 2
Other 1
MIT Male Exchange Students
1
1
1
1
.52
Consideration of Novel Use
.60
Self-Rated Innovative Ability
.58
Self-Rated Basic Research Skills
.18
Self-Reported Research Experience
.10
−.03
−.04
−.05
.16
.63
Cambridge Female Exchange Students
−.13
−.13
.21
−.06
MIT Female Exchange Students

Fig. 4.7 Adding a direct effect of self-reported basic research experience on consideration of use, unstandardized coefficients

model that has an additional direct path from research experience to consideration of use. This model fits the data very closely. The standardized path coefficient of 0.18 is in the middle range but it improves the R^2 from 0.33 to 0.37, a noticeable improvement over the earlier models. The model χ^2 $p = 0.68$, which is considerably better than the $p = 0.23$ of the basic model. Moreover, its AIC is significantly lower, $44.6 < 48.8$, and the BIC is also lower, $98.1 < 99.6$. Since the BIC selects parsimonious models, the change in BIC is less than the change in AIC, and suggests that the basic model is still acceptable. If the insignificant difference between Cambridge females and Cambridge males regarding their research experience is removed from the model, then the fit of the elaborated model is even better: the model χ^2 $p = 0.73$, AIC = 43, and BIC = 93.7, thereby further substantiating the institutional perspective that has informed this empirically specified theory of consideration of use.[20]

Decomposition of Effects

Given the correlations of Table 4.6 and the recursive path-analytic models whose coefficients are based on these correlations—the parsimonious model of Fig. 4.4 or the elaborated model of Fig. 4.6—the total effects of the variables can be calculated following the decomposition rules of path analysis (Alwin and Hauser 1975; Kline 2005, 129–131). Here, the total effect of a variable on a response variable is the sum of its direct effect and any indirect effects through other variables; the sum of this total effect and any spurious or shared effects should equal the total (i.e., zero-order) correlation between the response and the focal variable.[21] The reader is invited to apply graph-theoretical decomposition rules to Fig. 4.7 and to compare the results to the decompositions in Table 4.8. These path-regression coefficients can be interpreted as probabilities because their range is from 0 to 1.[22]

Effect of innovative ability on consideration of use

In Fig. 4.4, the students' self-rated innovative ability to create novel designs and solve unstructured problems is the only direct effect on consideration of use; the effects of all the other variables operate through innovative ability. In this diagram, the effect rendered by the path coefficient (0.58) equals the total correlation between these two variables (0.58). The path diagram of Fig. 4.6 suggests that the additional direct effect of self-reported research experience on consideration of use creates some spuriousness, which reduces this direct effect of innovative ability on consideration of use to 0.55; the spurious component (0.03) of the original correlation can be estimated by the difference between the two path coefficients $(0.58 - 0.55)$. Alternatively, this direct effect of research experience multiplied by its indirect effect on innovative ability via self-rated basic research skills creates a spurious component $0.56 \times 0.37 \times 0.18 = 0.037$ that reduces the direct effect of innovative ability.[23]

Effect of research experience on consideration of use

The diagram of Fig. 4.4 shows that self-reported research experience only influences consideration of use indirectly through the generative mechanism; this indirect effect of $0.12 = 0.37 \times 0.56 \times 0.58$ underestimates its impact. Its simple correlation with consideration of use is 0.25; the difference of 0.13 suggests that some other path has been left out, namely the direct effect of research experience on consideration of use.

In Fig. 4.6, this direct path is 0.18 and the reduced indirect effect is $0.114 = 0.37 \times 0.56 \times 0.55$; thus the predicted total effect of 0.294 overestimates the total correlation by $0.294 - 0.25 = 0.044$. This overestimation does not take into account the spuriousness due to the effects of the background variables directly on research experience and indirectly on consideration of use. For example, consider how MIT females spuriously influence the correlation between research experience and consideration of use: the attribute MIT females has a direct effect of 0.70 on research experience and an indirect effect on consideration of use of −0.086, via the mechanism MIT females, basic research skills, and innovative ability $(-0.28 \times 0.56 \times 0.55)$, the spurious component is $0.70 \times (-0.086) = -0.060$. Similar spurious components are due to Cambridge females (+0.003) and perhaps due to MIT males. Thus, given Fig. 4.6, a reasonable estimate of the total correlation of self-reported research experience is the sum of direct effect = 0.18 + indirect effect = 0.11 + spurious effect = −0.06 = 0.23, which is slightly less than the zero-order correlation of 0.25.

Effect of basic research skills on consideration of use

Figure 4.4 suggests that self-rated confidence about one's basic research skills only has an indirect effect via innovative ability on consideration of use of $0.58 \times 0.56 = 0.325$, which is considerably less than its correlation of 0.40 with consideration of use. This difference suggests that the model is not specified correctly, a spurious component is left out. In Fig. 4.6, the direct path from research experience to consideration of use creates the major missing spurious component; it equals $0.18 \times 0.37 = 0.067$, which when added to the indirect effect of $0.55 \times 0.56 = 0.308$ produces a predicted correlation of 0.38, which is slightly less than 0.40; this difference is due to spurious effects of the background variables. Given this model, the total nonspurious effect of self-rated basic research skills on consideration of use is about 0.31. The size of this total effect on consideration of use is less than the total effect of innovative ability (0.55) and slightly larger than the total effect of research experience (0.29); in these models, variables closer to the response variable have stronger effects, and this holds true for the background variables as well.

Effects of background variables on consideration of use

The indicator-variable coding of the background variables uses Cambridge males as the reference category; the reported effects are thus relative to the Cambridge males: MIT males and females have higher consideration of use and Cambridge females lower consideration of use, as follows. In Fig. 4.6, the relevant paths between MIT males and consideration of use are through research experience:

$0.47 \times (0.37 \times 0.56 \times 0.55 + 0.18) = 0.054 + 0.085 = 0.14$. Although the MIT females have more research experience, their lower self-rated confidence in their basic research skills weakens their reported consideration of use, as follows: $0.70 \times (0.37 \times 0.56 \times 0.55 + 0.18) + (-0.28 \times 0.56 \times 0.55) = 0.08 + 0.126 - 0.086 = 0.12$. On the basis of their research experience, they are ahead of the MIT males by $0.206 - 0.14 = 0.066$. But their lower confidence in basic research creates a slight disadvantage, $0.066 - 0.086 = -0.02$. However, MIT males and females both report more consideration of use than Cambridge males and females (respectively, 0.14 and 0.12). But, primarily because of lower reported confidence in basic research skills, Cambridge females report less consideration of use than Cambridge males: $-0.04(0.37 \times 0.56 \times 0.55 + 0.18) + (-0.25 \times 0.56 \times 0.55) = -0.0046 + -0.007 + -0.077 = -0.089$. As noted earlier, why these female exchange students report less confidence in basic research skills than their male colleagues is not yet known.

The calculation of total nonspurious total effects in complex path diagrams can be difficult; facilitating this endeavor AMOS can estimate the direct, indirect, and total nonspurious effects. Using AMOS's calculations for the standardized path coefficients, Table 4.7 presents quantified edge matrices similar to those for graphical models: the matrix of direct effects (Panel A) is similar to a parental graph, the matrix of indirect effects (Panel B) is similar to an ancestor graph, and the element-by-element sum of these two matrices produces the matrix of total nonspurious effects (Panel C). The spurious components can be calculated by subtracting these total effects from the total zero-order correlations. Using AMOS's calculations for the unstandardized path-regression coefficients, Table 4.8 presents the edge matrices for the direct, indirect, and total effects; apparently, these decompositions are rather similar to those for graphical models.

This example of a generative process has delineated path-analytic multivariate dependencies.[24] The estimates of the total effects of the variables are reasonable because there is cumulative subject-matter evidence that supports the models from which these effects were derived.[25] These models articulate variables of social context, a generative mechanism, and a response, thereby providing a heuristic paradigm for the study of other generative processes in basic, applied, and policy research.[26]

Causality in Policy Research

Nobel Prize winning economist James J. Heckman (2005a, 2005b) has explicated for sociologists the model of causality that informs his policy research studies; this section briefly reviews aspects of those articles. The reader can consult his writings for the nuanced details and explication of how he addresses problems of identification and estimation; see his web site http://jenni.uchicago.edu/discussion/discussion.html (available circa May 2010).[27] His econometric approach builds on the different notions of causality sketched earlier. He refers to his model as scientific because it focuses on modeling the various *causes of an effect,* as do Coleman's and

Table 4.7 Decomposition of standardized effects for path diagrams of Fig. 4.6

	Priors	Y	X	BS	RX	TF	CF	TM
Panel A: Quantified parental edge matrix: direct effects of prior variables on responses, path coefficients								
Responses		○	○	○	○	●	●	●
Y	○	1	0.551	0	0.183	0	0	0
X	○		1	0.556	0	0	0	0
BS	○			1	0.372	−0.280	−0.246	0
RX	○				1	0.696	−0.043	0.468
TF	●					1	–	–
CF	●						1	–
TM	●							1
Panel B: Quantified ancestor edge matrix: indirect effects of prior variables on responses, path coefficients								
Responses		○	○	○	○	●	●	●
Y	○	0	0	0.307	0.114	0.121	−0.088	0.139
X	○		0	0	0.207	−0.012	−0.146	0.097
BS	○			0	0	0.259	−0.016	0.174
RX	○				0	0	0	0
TF	●					0	–	–
CF	●						0	–
TM	●							0
Panel C: Quantified total effects edge matrix: direct plus indirect effects of prior variables on responses, path coefficients								
Responses		○	○	○	○	●	●	●
Y	○	1	0.551	0.307	0.297	0.121	−0.088	0.139
X	○		1	0.556	0.207	−0.012	−0.146	0.097
BS	○			1	0.372	−0.021	−0.263	0.174
RX	○				1	0.696	−0.043	0.468
TF	●					1	–	–
CF	●						1	–
TM	●							1

Cambridge University male students are the base for the indicator variable codes. A circle ○ indicates a continuous ordinal variable; a solid circle ● indicates a dichotomous variable.
The variables' names are as follows: Y = consideration of use, X = self-rated innovative ability, BS = self-rated basic research skills, RX = self-reported research experience, TF = MIT female students, CF = Cambridge University female students, TM = MIT male students.

Goodman's statistical methods, and Duncan's path-analytic models. Statistical theorists of causality focus on the various *effects of a cause* (e.g., Holland 1986, 945), but policy and social problems researchers most often want to understand the fundamental *causes of effects*. This requires a clear definition of the policy problem being addressed, a theoretical model bearing on the policy problem, determination of what parameters are required to answer the problem; the uncovering of minimal identification conditions; and the explication of the properties of various estimators (Heckman 2005b, 139).

Table 4.8 Decomposition of unstandardized effects for path diagram of Fig. 4.7

	Priors	Y	X	BS	RX	TF	CF	TM
Panel A: Quantified parental edge matrix: direct effects of prior variables on responses, path regression coefficients								
Responses		○	○	○	○	●	●	●
Y	○	1	0.605	0	0.103	0	0	0
X	○		1	0.576	0	0	0	0
BS	○			1	0.185	−0.125	−0.126	0
RX	○				1	0.627	−0.044	0.520
TF	●					1	–	–
CF	●						1	–
TM	●							1
Panel B: Quantified ancestor edge matrix: indirect effects of prior variables on responses, path regression coefficients								
Responses		○	○	○	○	●	●	●
Y	○	0	0	0.348	0.064	0.061	−0.051	0.087
X	○		0	0	0.106	−0.005	−0.077	0.055
BS	○			0	0	0.116	−0.008	0.096
RX	○				0	0	0	0
TF	●					0	–	–
CF	●						0	–
TM	●							0
Panel C: Quantified total effects edge matrix: direct plus indirect effects of prior variables on responses, path regression coefficients								
Responses		○	○	○	○	●	●	●
Y	○	1	0.605	0.348	0.167	0.061	−0.051	0.087
X	○		1	0.576	0.106	−0.005	−0.077	0.055
BS	○			1	0.185	−0.009	−0.134	0.096
RX	○				1	0.627	−0.044	0.520
TF	●					1	–	–
CF	●						1	–
TM	●							1

Cambridge University male students are the base for the indicator variable codes. A circle ○ indicates a continuous ordinal variable; a solid circle ● indicates a dichotomous variable.
The variables' names are as follows: Y = consideration of use, X = self-rated innovative ability, BS = self-rated basic research skills, RX = self-reported research experience, TF = MIT female students, CF = Cambridge University female students, TM = MIT male students.

Theoretical Models Are Necessary

According to Heckman, a model of causality for policy research should address the following policy evaluation problems (2005a, 8), each of which requires a theoretical model for its solution. The first problem (**P1)** focuses on the *internal validity* of a specific randomized trial or quasi-experiment: What were the impacts (i.e., individual-level or population–level implications and their valuations) of the treatment in the given environment? The second problem (**P2)** focuses on *external validity*: Can the model of the intervention and the findings be applied in different settings? The third problem (**P3**) involves the *forecasting* of the effects of a new policy not yet experienced: Can past experience with one policy be used to forecast the impacts of a new policy?

Heckman views causality as a property of a model of hypotheticals (2005a, 2):

> A fully articulated model of the phenomena being studied precisely defines hypothetical or counterfactual states. A definition of causality drops out of a fully articulated model as an automatic by-product. A model is a set of possible counterfactual worlds constructed under some rules. The rules may be the laws of physics, the consequences of utility maximization, or the rules governing social interactions. … Counterfactual statements must be made within a precisely stated model. Ambiguity in model specification implies ambiguity in the definition of counterfactuals and hence of the notion of causality.

Scientists create such models based on their knowledge of the subject matter, logic, and their imagination (2005b, 3). The scientific community is more likely to accept a hypothetical model (i.e., a model of counterfactuals) if it is empirically based, explicit, consistent with the implications of other theories, and agrees with established interpretations of facts. Models based on cumulative research studies are more likely to meet these criteria than models created on an *ad hoc* basis.

Given that an investigator has constructed a theoretical model, the fundamental problem of causal inference must be confronted: at any point in time, person ω (i.e. omega) can be observed in one state s but not in another state s'; these states are mutually exclusive. A solution to this problem is provided by the experimental paradigm of potential outcomes causality and Rubin's causal model: $\delta = Y(s, \omega) - \mathrm{Y}(s', \omega)$ is estimated in a population applying the stable-unit-treatment-value assumption (SUTVA) and other assumptions using means or some other measure. A second solution, which Heckman prefers, is in the tradition of structural econometric analysis: $Y(s, \omega)$ is modeled explicitly in terms of its theoretically specified determinants with the goal of understanding the factors underlying the outcomes, choice of outcome equations, and their interrelationships (2005a, 17).

The Experimental Paradigm

Operating purely in the domain of theory, Heckman develops a notation for causality in policy research that contributes to the potential outcomes perspective. He provisionally accepts and redefines in his own notation (2005a, 11) the SUTVA and the ignorability assumption that the outcome is not affected by the treatment assignment (Rubin 1986, 961). He elaborates the potential outcomes perspective by defining the individual-level treatment effect for person ω ($\omega \in \Omega$) comparing outcomes from treatment s to those from treatment s'; s and s' are elements of the universe of treatments S (i.e., $s, s' \in S$). His equation number (1) defines the basic individual-level causal effect, which can be a random variable or a constant (2005a, 12):

$$Y(s,\ \omega) - \mathrm{Y}(s', \omega), s \neq s', \tag{1}$$

In footnote 7, he presents this more complete model:

$$Y(s, \omega, \rho, \tau) - Y(s', \omega, \rho, \tau).$$

This model specifies that the policy treatment effect is defined under a specific policy regime $\rho \in P$ and for a specific mechanism of selection within the policy regime $\tau \in T_\rho$. He then defines (2005a, 12–14) other comparisons that could be made, depending upon the policy research question. These include personal or planner utilities, social welfare, cost benefit between two policies, and comparisons for different social characteristics.

Heckman also discusses population treatment effects based on means (2005a, 14–21): the average treatment effect (ATE), the average treatment effect on the treated (TT), the average treatment effect on the untreated (TUT), the effect of treatment for people at the margin of indifference (EOTM). Then, because of the limitations of means as the outcome measure, he considers criteria for scalar outcomes (2005a, 21–22): the proportion of people taking the program who benefit from it relative to an alternative; the proportion of the total population that benefits from one program compared with another program; various quantiles of the impact distribution; and distributions of the outcomes for the disadvantaged. He also discusses *ex ante* and *ex post* outcomes and the role of uncertainty.

Limitations of the Experimental Paradigm

Heckman's stipulation that policy researchers create a scientific model of the phenomena they are studying differs from the logic of some experimental studies, especially clinical trials. This example may clarify the difference. In a randomized clinical trial (RCT), the causal force resides in the chemical composition of the experimental drug, whose effect is being tested. The composition of the drug (i.e., its model) is the concern of the pharmaceutical scientists whose research has led them to formulate the drug. The experimental drug, having been tested for safety (phase 1) and dosage (phase 2), is then tested for efficacy, noninferiority, or equivalence, by comparing its outcomes to that of the control treatment (phase 3). Biostatisticians apply their extensive knowledge of statistical science to design the trial, monitor its results, and analyze its data. They focus on the outcomes of the treatments asking: Was the outcome in the group that received the experimental treatment, compared with the outcome of the group that received the control treatment, consistent with the hypothesis being tested?

Innovative contributions to statistical science bearing on the design and analysis of clinical trials are highly specialized, profound, and often brilliant. Even so, when the model of causality based on randomized clinical trials is applied more broadly in the social sciences, especially to answer research questions in policy research that involve economic science, there are numerous limitations that Heckman discusses (2005a, 5–6, 35–38); these are summarized next: (1) incomplete specification of the formal model from which counterfactual implications are derived; (2)

intuition and not explicit formal models of the substantive problem guide the constructions of counterfactual outcomes; (3) the narrow focus on the outcomes of treatment only implicitly specifies the model-selecting outcomes; (4) randomization and matching rule out alternative channels of identification of counterfactuals; (5) why observationally equivalent people make different choices and have different outcomes for the same choice is not explained; (6) the model rules out simultaneous choices of outcomes of treatment and one outcome causing another; (7) treatments are opaque boxes whose components are neither analyzed separately nor linked to an underlying theory; and (8) out-of-sample forecasts of treatment effects to new samples cannot be readily made.

Heckman's Paradigm

Heckman (2005a, 3–4) says his scientific model overcomes these shortcoming if these three tasks are fulfilled—definition, identification, and estimation: (1) *define* a model of the phenomena based on a scientific theory, the model is a set of hypotheticals or counterfactuals; (2*) identify* relevant parameters of the model using hypothetical population distributions and mathematical analysis (i.e., explore the model's implications using hypothetical data); and (3) *estimate* the parameters empirically using samples of real data to test the model and the underlying theory (i.e., ground the model empirically).[28] He presents his model of causality by covering in depth the following topics that are coordinated to the above tasks (2005b, 139):

> A clearly formulated causal model should (a) define the rules or theories that generate the counterfactuals being studied, including specification of the variables known to the agents being studied as well as the properties of the unobservables of the model where the unobservables are not known to the analyst but may be partly known by the agent; (b) define how a particular counterfactual (or potential outcome) is chosen; (c) make clear the assumptions used to identify the model (or to address the policy questions being considered); and (d) justify the properties of estimators under the maintained assumptions and under alternative assumptions. These are tasks 1 (corresponding to a and b), 2 (corresponding to c), and 3 (corresponding to d).

The next section briefly highlights conceptual aspects of Heckman's all-causes scientific model (his task 1).

Structural Economic Models

Because structural models focus on the various causes of effects, Heckman declares that such models solve all three policy evaluation problems better than models that focus narrowly on the various effects of a cause. He substantiates this assertion by showing how the schooling outcomes of students vary with patterns of inputs. Let s index the observed type of school (e.g., regular public, charter public, private secular, and private parochial). Let x index the observed characteristics of the school children (e.g., parental social class, ethnicity, and attentiveness in class). Then, the "production function" relating inputs to outputs is the "all causes" model of traditional economics, as specified by Heckman's (2005a, 27) equation (9):

$$y(s) = g_s(x,\ u_{\mathrm{s}}). \tag{9}$$

Here x and u_{s} are fixed variables for student ω; a summary of Heckman's (2005a, 27) elucidation of this equation follows:

1. The lower case notation denotes fixed variables.
2. The arguments in g_s explain in a functional sense all of the outcomes $y(s)$ in (9).
3. The "all causes" functional relationship (9) appropriately models the *ex post* realizations of outcomes because after the fact all uncertainty has been resolved.
4. The notation x and u_{s} implies that some arguments of (9) are observed by the analyst while others may be unobserved.
5. The notation allows different unobservables from a common list u to appear in different outcomes because g_s maps $(x,\ u_s)$ into y.
6. Because the domain of definition D of g_s can differ from the empirical support, (9) can map logically possible inputs into logically possible *ex post* outcomes; in a real sample, only a subset of D may be observed.

Heckman adds some complexity to this model by letting c_s index the observed characteristics of schools of type s (e.g., the ratio of students to teachers, the number of computers per students, and the effectiveness of the principal); then the model becomes Heckman's "deep structural" equation (10):

$$y(\mathrm{s}) = g(c_{\mathrm{s}}, x, u_{\mathrm{s}}) \tag{10}$$

The inclusion of c_s allows the components of the type of school s to be characterized as different bundles of the same characteristics that generate all outcomes. In one school of type s, the school reform lowers the ratio of students to teachers; in another school the ratio of computers to students is increased and the principal receives professional development that promises to improve effectiveness; and in another school all three reforms are implemented. Hypothetical causal effects of a pivotal variable can be studied by conditioning on the other variables and systematically setting the values of the pivotal variable in parameter studies and thought experiments (2005a, 32–34).

Heckman believes that research guided by this framework can solve all three types of policy evaluation problems. When (10) guides an evaluation of the effects of existing types of schools, programs, and students, it can solve **P1**, that is, the assessment of the impacts of an actual intervention. When known schooling inputs—the type of school s and mix of program characteristics c_s—are applied to different students, and either (9) or (10) is identified over the new domain of definition, then **P2** can be solved; that is, the model developed in one environment can be applied to forecast outputs in another environment. If a proposed new type of school can be defined as a new combination of different levels of x, c_s, and u and this new package identified over the domain of (10), then **P3** can be solved; that is, the effect of a new policy never before experienced can be forecast. After presenting this example, Heckman considers in depth problems of identification and estimation.

Conclusions

Chapter 3 and this chapter have sketched three general notions of causality in social inquiry that closely parallel Cox and Wermuth's distinctions (2001, 65–70): *stable association* (i.e., level-zero causality) includes classic causality and robust dependence and *potential outcomes* (i.e., level-one causality) here includes causality as an effect of an intervention and causal models for attributes. *Dependency networks* (i.e., level-two causality) includes graphical models, association graphs for log-linear models, generative processes, and causality in policy research (structural economic models). All of these notions are valuable because they enable the reader to locate the causal zones of the results of empirical inquiries, improve causal inferences in their own research, and critique the causal claims of other research studies, especially those of the subsequent chapters.

Acknowledgements The author derived the exposition of the generative mechanism, which links the students' home university to their consideration of use, from unpublished research reports by Smith and Lucas (July 15, 2005; August 12, 2005; November 7, 2005; June 19, 2006) and Lucas et al. (March 2006). This research was conducted under the auspices of the Cambridge-MIT Institute (CMI) and is reused here with the consent of former CMI Directors. The author based the exposition of graphical modeling of voting on portions of his article, Issues matter: A case study of electoral voting, in *Case Studies in Business, Industry, and Government Statistics: An Internet Journal* (vol. 2, no. 2, 127–146; © 2009 Robert B. Smith).

Endnotes

[1]

```
*This SAS code will replicate the proportional odds model for the three
voting choices, given the relevant data set ThreeClusters123
   It applies proc logistic illustrating the use of three weighted files
   to produce true estimates of the three class latent structure;
   *In addition to Springer's website, these data are also available at the
internet journal Case Studies in Business, Industry, and Government
Statistics (CSBIGS) at Bentley University. See Volume 2, number 2, the
article by Smith 2009;
*Table 4.1. Panel 1;
proc logistic;
class trueclas partyid libcon minority
   /param = reference;
model choice3 = trueclas libcon partyid minority/
scale = none aggregate lackfit ctable rsquare;
weight newvar;
run;
```

```
*This code replicates the continuation logits model for Clinton vs. All
others;
*Table 4.1. Panel 1.1;
proc logistic descending;
class trueclas partyid libcon minority
   /param = reference;
model clin1vot = trueclas libcon partyid minority/
scale = none aggregate lackfit ctable rsquare;
weight newvar;
run;
*This code replicates the continuation logits model for Perot vs. Bush;
*Table 4.1. Panel 1.2;
proc logistic descending;
class trueclas partyid libcon
   /param = reference;
model pero1vb0 = trueclas libcon partyid/
scale = none aggregate lackfit ctable rsquare;
weight newvar;
run;
```

[2] This SPSS syntax will find this best-fitting loglinear model given the relevant data set threeclusters123.sav. Just open the data file, paste the syntax into the syntax window, and run it. In addition to Springer's website, these data are available at the internet journal *Case Studies in Business, Industry, and Government Statistics* (CSBIGS) at Bentley University. See vol. 2, no. 2, the article by Smith 2009:

```
USE ALL.
COMPUTE filter_$ = (fileid = 1).
VARIABLE LABEL filter_$ ‘fileid = 1 (FILTER)’.
VALUE LABELS filter_$ 0 ‘Not Selected’ 1 ‘Selected’.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
HILOGLINEAR
econreg1(2 4) hreformX(2 4) envirvip(1 2) charnvip(1 2) choice3(1 3)/METHOD = BACKWARD
/CRITERIA MAXSTEPS(10) P(.05) ITERATION(20) DELTA(.5)
/PRINT = FREQ RESID
/DESIGN.
```

[3] Goodman ([1972] 1978, 91) notes that loglinear models do not establish causation; to give a causal interpretation to the system of variables, certain assumptions must be made: "For example, we must assume that the system under study is closed in the sense that, if variables outside the system. . .have any effect on the variables in the system, it is to produce stochastic disturbances corresponding to the usual random variation of the observed frequency f_{ijkl} from its expected value F_{ijkl} (when sampling is random). In addition, we must make other assumptions (e.g., assumptions as to which variables are prior to which other variables) in order to select one causal system from among the various causal systems that are consistent with a given model."

[4] *The following SAS syntax creates the dichotomous Clinton versus Bush response variable, calculates the main effects, and the predicted values. Then it calculates the interaction effect and the predicted values. It uses the first file of threeclusters123;

```
data trythree; set trytwo;
```

```
*this creates a dichotomous response variable Clinton vs. Bush;
if choice3 =2 then delete;
if choice3 =3 then choice3 =2;
*this calculates the main effects dichotomous response;
proc logistic;
class character envirvip econreg1 hreformx/param = ref;
model choice3 = character envirvip econreg1 hreformx
/scale = none aggregate lackfit ctable rsquare;
output out = predict pred = prob;
run;
*this calculates the predicted values for the main effects model;
data bob; set predict;
if character =. then delete;
if envirvip =. then delete;
proc sort; by character envirvip;
proc means; var prob;
by character envirvip;
run;
*this calculates the interaction effect and the main effects;
proc logistic;
class character envirvip econreg1 hreformx/param = ref;
model choice3 = character| envirvip econreg1 hreformx
/scale = none aggregate lackfit ctable rsquare;
output out = predict1 pred = prob1;
run;
*this calculates the predicted values for the interaction effects model;
data bobs; set predict1;
if character =. then delete;
if envirvip =. then delete;
proc sort; by character envirvip;
proc means; var prob1;
by character envirvip;
run;
```

[5] The predicted values provide the following interpretations for this interaction effect. For the main effects model when both the character issue and the environmental issue favor the Republican positions (C- N-) then the predicted proportion choosing Clinton (+) is 0.27. When the environmental issue changes to N+, then the proportion for Clinton is increased to 0.58, an increase of 0.31. For the model with an interaction effect the analogous proportions are C- N- = 0.23 and C- N+ = 0.63, for a larger increase of 0.40. However, when the character issue initially favors the Democratic candidate (+) and the environmental issue does not (−) the main effects model exhibits the largest gain when there is consonance: For the main effects model C + N- = 0.64, C + N+ = 0.85 for a gain of 0.21 for Clinton. For the interaction effects model C + N- = 0.68, C + N+ = 0.80 for a gain of 0.12 for Clinton. The difference between the differences = |0.09 |. Such interaction effects should be interpreted with extreme caution (Littell et al. 2006, 551).

[6] Lazarsfeld (1955b, *xi*) suggests that in addition to the time ordering of variables, the general to the specific ordering principle can be applied, e.g., patriotism is more general than saluting the flag, so the latter is a manifestation of the former. He also suggests that structural levels can be distinguished: e.g.,

membership in an occupational group, employment in a specific factory, and membership in a special clique within the shop.

[7] Hyman (1955, 326) suggests that the investigator’s subject matter knowledge governs when to stop inserting intervening variables into a chain of relationships, he states: “We stop inserting new links into this chain only when we feel psychologically satisfied that the underlying process has been clarified.”

[8] Classic examples of generative mechanisms are Hyman’s (1955, 324–326) analysis of the variables that intervene between prejudice (x) and misunderstanding of a cartoon theme (y); Glock and Stark’s (1966, 134–135) explication of how hostility (t_1) interprets the relationship between religious dogmatism (x) and anti-Semitism (y); and (Glock’s 1967) codification of this method. Currently, Morgan and Winship (2007, 219–242) review the roots of mechanistic explanations, analyze hypothetical mechanisms using graphical methods, and discuss the movement toward such explanations in sociology.

[9] In early 2004, Lucas formulated these and other survey questions and included these items in several surveys of engineering and science students at MIT and in the United Kingdom.

[10] The AIC and the BIC are essentially values of the log likelihood that have been penalized for the number of parameters estimated; the BIC imposes a heavier penalty than the AIC.

[11] An internally valid index is composed of items that are more strongly correlated with each other than with other items not included in the index. Each item should exhibit strong, expected correlations with other survey questions that logically follow from the meaning of the index. Cronbach’s alpha (α) provides an indication of the index’s internal validity. An externally valid index has strong correlations with actual behaviors that logically follow from the meaning of the index. For example, students who self-report high levels of basic research skills should actually perform better in research than those who self-report low levels of basic research skills. Students who participate in entrepreneurial competitions should exhibit higher scores on venturing self-efficacy than those who do not participate.

[12] A simple answer may be that male scientists treat their younger female colleagues less well than they treat their younger male colleagues. If this were true, then women who are treated nicely should blossom, which seems to be the case. Zeldin and Pajares (2000, 237–243) find that the self-efficacy beliefs of successful women scientists concerning their mathematics ability were nurtured by favorable familial, academic, peer, and work-related influences. The encouragement they received and the vicarious experience of watching and learning from others, which their participation in research provided, enhanced their self-efficacy beliefs.

[13] The more detailed fit statistics confirm that the alternative model fits very slightly better than the reported model. Note that the BIC statistic picks the more parsimonious of these two models whereas the AIC does not (the AIC difference of 1.6 is < 2):

	Box 4.1 Measures of goodness of fit	
	Reported	Alternative
	Model	Model
Model chi square	12.85	13.25
Degrees of freedom	10	11
Probability	0.23	0.28
Noncentrality parameter	2.85	2.25
Comparative fit index	0.988 > 0.90	0.991 > 0.90
Root mean square residual	0.005	0.005
RMSEA	0.048	0.041
LO 90	0	0
HI 90	0.115	0.108
PClose	0.46	0.52
AIC	48.85	47.25
BIC	99.61	95.19

[14] The noncentrality parameter, designated as δ, expresses the degree of misspecification of the model. It has this form (the M refers to the model being tested):

$$\hat{\delta}_M = \max\left(\chi^2_M - df_M, 0\right)$$

The estimate of $\hat{\delta}_M$ is which ever is larger: zero or the difference between the chi square for the model minus the model's degrees of freedom (Kline 2005, 137–138).

[15] The RMSEA is the root mean square error of approximation. It estimates the amount of error of approximation per model degree of freedom and takes sample size into account. The formula takes the square root of the noncentrality parameter for the model and divides it by the product of the model's degree of freedom times the sample size minus 1: RMSEA $= (\delta_{\mathrm{M}}/df_{\mathrm{M}}\ (N - 1))^{1/2}$. Values close to zero (RMSEA < or = 0.05) and a confidence interval between 0 and 0.10 indicate a close fit of the model to the data (Kline 2005, 137–140).

[16] The students' home university and gender exhibit different means on confidence in basic research skills: on a 0 to 1 scale MIT students have a mean of 0.58 compared with a mean of 0.51 for Cambridge students; but this difference is not statistically significant ($b = 0.06$, $t = 1.7$, $p = 0.09$). Young men report more confidence in their basic research skills than young women; 0.59 − 0.50, for a significant difference ($b = 0.09$, $t = 2.5$, $p = 0.01$). When these regressors are used to simultaneously predict basic research skills, both become statistically significant: MIT students have more confidence in their basic research skills ($b = 0.095$, $t = 2.61$, $p = 0.01$) and young men have more confidence than young women ($b = 0.12$, $t = 3.2$, $p = 0.002$). However, the variable "self-reported research experience" interprets the effect of the university: with this control the difference between MIT and Cambridge disappears ($b = -0.03$, $t = -0.60$, $p = 0.54$), but the gender difference in favor of men remains ($b = 0.12$, $t = 3.5$, $p = 0.0007$). However, the direct effect of prior research experience on basic research skills is considerably larger than these effects and very significant ($b = 0.21$, $t = 3.5$, $p = 0.0007$).

[17] Research on change over time for the 2004–2005 cohort of exchange students supports the view that consideration of use is malleable. MIT students after an academic year at Cambridge significantly reduced their level of reported research experience ($p = 0.0008$) and, concomitantly, their consideration of use ($p = 0.0006$). After an academic year at MIT, the Cambridge students reported a significant change in their research experience ($p = 0.0004$) but no significant drop in their level of consideration of use.

[18] Smith and Lucas's memo of November 7, 2005 tested this hypothesis. Using instrumental variables, a model with reciprocal causation between consideration of use and confidence in basic research skills was estimated. The effect of basic research skills on consideration of uses was large ($b = 0.48$) and statistically significant ($p = 0.04$) whereas the effect of consideration of use on basic research skills was negative (b = −0.11) and not significant ($p = 0.85$). Additionally, the recursive model that assumes that basic research skills drive consideration of use produced a range of fit statistics that were superior to the alternative recursive model that assumes that consideration of use drives basic research skills. Apparently, basic research skills are prior to consideration of use, as Figures 4.4 through 4.7 depict.

[19] Anecdotal evidence supports this view (Thomson 2005). When commenting on his selection for a Nobel Prize in chemistry, Richard Schrock of MIT credited basic research as catalyst to his success. He defines basic research as the exploration of an interesting new area that may have potential. As Schrock pointed out, his work had applications that he did not anticipate that time.

[20] Robins and Wasserman (1999, 306–307) stress that in an observational study, given a large enough sample, almost any $x \rightarrow y$ relationship is vulnerable to the effects of some unmeasured confounder U_{k}. Some rather obvious potential confounding background variables that may be correlated with consideration of use are the students' departmental major (no effect), age ($r = 0.19$, $p = 0.13$), female gender ($r = -0.07$, $p = 0.55$), and MIT home university ($r = 0.13$, $p = 0.15$). In contrast to these small, insignificant correlations, note that confidence in one's basic research skills is strongly and significantly correlated with consideration of use ($r = 0.39$, $p = 0.001$). These data are from the 2004–2005 cohort of exchange students at baseline as reported by Smith and Lucas (July 15, 2005, 3).

[21] Pearl ([2000] 2009, second edition, 164) defines direct and indirect as follows:

> Definition 5.4.2. (Total Effect)
>
> *The* total effect *of X on Y is given by P(y | do(x)), namely, the distribution of Y while X is held constant at x and all other variables are permitted to run their natural course.*
>
> Definition 5.4.3. (Direct Effect)
>
> *The* direct effect *of X on Y is given by P(Y | do(x), do(s_{XY})), where S_{XY} is the set of all observed variables in the system except X and Y.*
>
> He then states:
>
> In linear analysis, Definitions 5.4.2 and 5.4.3 yield, after differentiation with respect to *x*, the familiar path coefficients in terms of which direct and indirect effects are usually defined.

[22] For explication of such rules see Cox and Wermuth (1996), Cooper (1999), Pearl ([2000] 2009, second edition, chapter 5), or Morgan and Winship (2007).

[23] Figure 4.6 indicates slight spurious components due to MIT females and Cambridge females. The compound path from MIT females to basic research skills (−0.28) and then to self-rated innovative ability (0.56) and from MIT females to research experience (0.70) and then directly to consideration of use (0.18) changes the spuriousness by $-0.02 = -0.28 \times 0.56 \times 0.70 \times 0.18$. The analogous effect for Cambridge females has even less effect:$.001 = -.25 \times .56 \times -.04 \times .18$. Because of the binary coding, these indirect effects and those involving MIT males may be uncertain.

[24] This exposition of path-analytic models estimated by AMOS of determinants of consideration of use built on earlier analyzes followed a recursive modeling strategy. Recent work by Pearl ([2000] 2009, second edition, 133–171), Morgan and Winship (61–74, 182–183), and others have developed “back door” and “front door” strategies for determining what variables in a system should be conditioned on for identifying causal effects of variables in directed acyclic graphs (DAGs) such as those in this chapter. The reader is invited to apply these approaches to the data and models of this chapter and to compare the differences.

[25] Robins and Wasserman (1999, 318–319) emphasize the necessity of prior subject matter knowledge as a guide to the inference of causal effects.

[26] However, generative mechanisms must be carefully specified. If no variables were hypothesized to intervene between research experience and consideration of use, then the direct effect of research experience on consideration of use would be larger. Intervening variables weaken the direct effects of antecedent variables. If a student’s research experience and innovative ability were dropped from the model, then there would be a direct effect of confidence in one’s basic research skills on consideration of use. Also, prior variables may create spurious components of effects of intervening variables.

[27] Heckman conducts use-inspired basic research as defined by Stokes (1997, 73). See his Nobel Prize acceptance letter present on his web site ca. May 12, 2009.

[28] Heckman’s three tasks of definition, identification, and estimation also apply to the development of computer simulation models. For example, the author defined his model of attitude change in Nazi Germany on the basis of a formalization of Goldhagen’s detailed study (1996) and data about the time period being simulated (Smith 1998). Parameter studies of the model identified key parameters and hypothetical relationships (Smith 2010b). Ongoing research will attempt to estimate the parameters of the model empirically and to generate and test its new implications.

Chapter 5
Uses for Multilevel Models

> *A fitted model has the form* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *where* $\mathbf{Y}$ *is a vector of responses,* $\mathbf{X}$ *is the fixed-effects design matrix and* $\boldsymbol{\varepsilon}$ *is a vector of residual errors. In this model we assume that* $\boldsymbol{\varepsilon}$ *is distributed as* N[0,**R**], *where* **R** *is an unknown covariance matrix. A common belief is that* $\mathbf{R} = \sigma^2\mathbf{I}$. … *A mixed-effects model, in general, has the form* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{\gamma} + \boldsymbol{\varepsilon}$ *where the extra term* $\mathbf{Z}_{\gamma}$ *models the random effects.* **Z** *is the design matrix of random effects and* $\boldsymbol{\gamma}$ *is a vector of random effects parameters.*
>
> —SPSS Mixed Models White Paper (2002, 6, 10)

Drawing on the contextual, evaluative, and summarizing studies of this book, this chapter explicates 11 uses for multilevel models and defines relevant vocabulary, concepts, and notational conventions. Because multilevel models are composed of both fixed and random components, statisticians refer to them as *mixed models* (Littell et al. 2006). Because multilevel models focus on data at different hierarchical levels, educational researchers refer to them as *hierarchical models* (Raudenbush and Bryk 2002). In the contextual analyses of data at one point in time, the level-1 response variable and its covariates are conceptualized as being grouped (or contained) within the level-2 units. In the analyses of data at several points in time, the level-1 response variable is an observation at a time point on an entity, and the repeated observations on that entity are said to be grouped or contained by that entity. The entity (e.g., a person, an organization, a country) is the level-2 unit. Ideally, multilevel models assess change on disaggregated data at several points in time (e.g., the scores on the repeated assessments of individual students who are grouped into classrooms). When the chapters of this book model aggregated data, it is because the disaggregated data are not available.[1] By applying special cases of generalized linear mixed models—the Poisson and logit—some chapters model response variables that are not normally distributed. By applying multilevel models, all of the core chapters address the clustering of level-1 units when they are contained within level-2 units.

A multilevel model includes covariance parameters, that is, the random effects, and such fixed, structural components as the design variables and their covariates. When the response variables are normally distributed, SAS can estimate the parameters of the multilevel models using expected mean squares from the analysis of

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_5, © Springer Science+Business Media B.V. 2011

variance (ANOVA), maximum likelihood (ML), or restricted maximum likelihood (REML) procedures (Littell et al. [1996] 2006, 7, 61–62).[2] The ML procedure results in biased estimates of the covariance parameter, whereas the REML procedure does not (Claeskens and Hjort 2008, 271–272).[3] REML obtains the estimates of the covariance parameters by minimizing the likelihood of residuals from fitting the fixed portion of the model, or equivalently minimizing twice the negative of the residual log likelihood. When the candidate models have the identical fixed-effects structure (i.e., the same set of fixed design variables and covariates, the same *X* design matrix), then the REML estimates are preferred for determining which of the several alternative covariance structures provide the best estimates of the random effects. But REML estimates are not always superior to ML estimates: If the candidate models have different sets of fixed components, and goodness-of-fit statistics are used to select the set that produces the model that fits the data best, then the maximum likelihood estimates are superior and should be used. After the ML estimation procedure has pointed out the best set of covariates, then REML can find the best estimates of the variance components, the effects of the covariates, and the standard errors of the estimated parameters.

Contextual Studies

These three chapters probe the multiple determinants of a response variable, that is, the various *causes of an effect*.[4] Substantively, they focus on uncovering some determinants of human development, contemporary violence against European Jewish people, and worker discontent stemming from the computerization of their workplace. Moving from the macro to the micro, the units of analysis are, respectively, regions of the world grouping countries; countries grouping measures at different points in time; and organizational units grouping their employees. The following uses, which are drawn from these chapters, illustrate how multilevel modeling can: (1) address clustered macrolevel units, (2) determine the statistical significance of a level-2 variance component, (3) account for a level-2 variance by classifying (i.e., grouping or nesting) the levels of the random variable, (4) analyze counts of events, (5) address clustered microlevel units, and (6) apply macro and micro explanatory variables. The models of categorical response variables and counts of events require estimation by Proc Glimmix (a generalized linear mixed models procedure); the models of normally distributed response variables are estimated by Proc Mixed (a mixed models procedure).

Clustered Macrolevel Units

Social researchers sometimes analyze macrovariables such as countries and their properties by applying the same methods they might use to analyze variables on people; they relate a property of a country to other properties. Examples of such studies are Lipset and Lenz's (2000) regression analysis that relates a country's level of corruption to its level of familial clientalism and to its citizens' use of illegitimate means to attain culturally proscribed goals, and Lipset's ([1959] 1981) earlier study that relates a country's level of democracy to its level of economic development. By applying econometric techniques to test the latter hypothesis, Barro (1999) has advanced the quantitative paradigm for the study of such macrosocial relationships.

Multilevel models can quantify relationships among macro- and microunits while adjusting for the clustering of data. In Chapter 7, for example, the response variables on 138 countries are their rank and scores on a human development index, which combines measures of economic development, literacy, and life expectancy. The covariates on the countries are their cultural zone and indicators of the following instrumental factors: democracy, slavery, debt, internal chaos, and corruption; different social and economic policies could change the values of these factors. Using the single-equation specification of the multilevel model and qualitative typologies as potential explanatory factors, the analyses focus on quantifying the effects of the covariates on human development, and uncovering the reasons why some geographic regions exhibit disparities in human development compared with the other regions. Because nearby countries contained within the same region usually are more similar to each other than to countries contained within distant regions, the observations on countries exhibit clustering within the 18 regions. This clustering violates a stipulation of regression analysis, namely that observations should be independent; multilevel modeling provides a solution.

In this example, the estimate of the unexplained variance between the r regions is symbolized by $\hat{\sigma}_r^2$, and the estimate of the unexplained variance among the c countries within a region is symbolized by $\hat{\sigma}_c^2$. These quantities exemplify three notational conventions: the hat signs (^) above these variances indicate that these quantities are estimated values; lower case subscripted letters denote level-2 and level-1 variance components; but, reflecting the different substantive variables of the chapters, the subscripted letters most often will differ from those of this example.

Based on estimates of these variances, the intraclass correlation coefficient ρ (rho) quantifies the amount of clustering. It is the quotient of the unexplained level-2 variance of the response variable ($\hat{\sigma}_r^2$) when divided by the sum of the level-1 ($\hat{\sigma}_c^2$) and level-2 unexplained variances ($\hat{\sigma}_r^2 + \hat{\sigma}_c^2$); that is, the estimate of $\hat{\rho} = \hat{\sigma}_r^2/(\hat{\sigma}_r^2 + \hat{\sigma}_c^2)$. When SAS Proc Mixed estimated these variance components using its default method of restricted maximum likelihood (REML), the initial value of $\hat{\rho}_{\text{REML}} = 0.46$ (i.e., 1,878.35/(1,878.35 + 2,214.68)) and it is statistically significant ($p = 0.003$); that is, the null hypothesis (H_0) of no clustering is rejected decisively. The full maximum likelihood (ML) estimates are very similar: $\hat{\rho}_{ML} = 0.44$ (i.e., 1,769.37/(1,769.37 + 2,214.54)).[5]

If clustering is ignored and the fitted regression model does not partition the variance of the response variable into its components (e.g., between the regions and within the regions), then the standard errors will be incorrect and the model will fit the data less well than a model that takes into account the clustering that the intraclass correlation quantifies. However, if $\hat{\rho}$ is not statistically significant and near zero, then the data are not seriously clustered; regular regression procedures may be appropriate.

Significance of a Level-2 Variance

In addition to testing for clustering using the intraclass correlation, some researchers want to account for the level-2 unexplained variance. To do so, they may classify the level-2 random variable by an explanatory typology (see use 3), or introduce other level-2 variables into the multilevel equation (see use 6). In either case, the change in the size and significance of the level-2 variance component is of interest. Holding constant the fixed structure of the model, the deviance statistic, or a variant of the deviance adjusted for the number of parameters in the model, can be used to compare the significance of the level-2 covariance parameters in models that have all or some of these parameters. For this application that applies REML estimation, the deviance is defined as the quantitative difference between $-2 \times$ [residual log likelihood of a current model minus the residual log likelihood of a saturated model that fits the data perfectly] (Singer and Willett 2003, 117).[6] Since the value of the latter is zero, the deviance is simply the $-2 \times$ residual log likelihood of the current model, which the SAS output reports as "−2 Res Log Likelihood" (abbreviated here as $-2LL_R$). For either REML or ML estimates of the parameters, likelihood-ratio tests and such goodness-of-fit statistics as the AIC (Akaike information criterion) and Schwarz's BIC (Bayesian information criterion) can help the modeler determine the importance of a variance component at level-2.

Again using data from Chapter 7 to illustrate these tests, the response variable is a country's rank on the human development index, the countries are grouped into 18 regions of the world, region is the random macrovariable, the fixed covariates are indicators of a country's civilization zone and this set of dichotomized (0,1) instrumental factors: emancipative employment (i.e., "emancip," low or no slavery = 1), a fully democratic political system (i.e., "dichofre," fully free = 1), low national debt (i.e., "lodebt," not a heavily indebted poor country = 1), no internal chaos (i.e., "nochaos," the absence of extreme disorder =1), and elite integrity (i.e., "integrit" = low corruption =1). The estimation method is REML.

For this model that does not nest the levels of the random regions by a typology, let y_{ij} denote the level on the HDI of the j^{th} country of the i^{th} region; μ denote the intercept level of the HDI; $X_{\mathrm{k}ij}$ denote a covariate k; and β_{ij} denotes its regression coefficient. Then,

$$y_{ij} = \mu + \Sigma\beta_{\mathrm{k}} X_{\mathrm{k}ij} + a_i + e_{ij},\ i = 1, 2, \ldots,\ \mathrm{r},\ j = 1, 2, \ldots, c_i,\ k = 1, 2, \ldots,\ K \quad (1)$$

where $a_i \sim iid\ N(0,\ \sigma_r^2)$ and $e_{ij} \sim iid\ N(0,\ \sigma_c^2)$. (*iid* means independent and identically distributed.) This syntax tells SAS Proc Mixed to estimate the parameters of this model:

```
Title 'Proc Mixed, Weights, No Nesting, HDI Rank, Dichotomized Factors';
Proc Mixed Data = UNdata covtest ratio cl = Wald;
Class region culture;
Model hdirank = culture emancip dichofre lodebt nochaos integrit/solution;
Random region/solution; Weight sqrtpop; Parms/nobound;
Run;
```

The meanings of these statements follow: The first line prints out a title for the SAS run. The second line calls Proc Mixed; specifies the analytic data set; and requests tests of significance for the variance parameters (covtest), their ratio (ratio), and their Wald confidence limits (cl = Wald). The Class statement defines region and culture as attributes for which SAS will create (0,1) indicator variables; for each set of indicator variables, the highest alphanumeric code of the untransformed variable will be used as the base. The model statement defines the fixed-effects structure of the model and requests a solution for the effects of the variables and their statistical significance. The random statement defines regions as a random variable and requests a solution for the regional random effects. The weight statement requests that the contribution of each country to the residual sum of squares is to be weighted by the square root of the country's population; that is, the weighted residual sum of squares is $\Sigma_\mathrm{i}\, w_\mathrm{i}\, (y_i - \hat{y}_i)^2$, where w_i is the square root of the population of country i, y_i is the observed value of the response variable, and $\hat{y}_i$ is the predicted value. The Parms/nobound statement removes the requirement of positive bounds on the confidence intervals for the variance components and requests that the level-2 variance component be tested for significance using the likelihood-ratio test; current versions of Proc Mixed implement this test automatically. Because no other covariance structure is specified, SAS uses its default variance components (VC) structure.

Likelihood-ratio test

Using the above syntax, the significance of a level-2 variance component can be tested following this strategy: Holding constant the fixed covariates, the specified model that includes a level-2 variance parameter is compared with a similar model in which the level-2 variance parameter is made absent by deleting the random statement. If the difference in $-2LL_R$ between the two models is statistically significant using a likelihood-ratio χ^2 (chi square) test with one degree of freedom (for the missing parameter), then the null hypothesis H_0 of no difference can be rejected. Rejection of H_0 implies that the alternative hypothesis H_1 is preferred; that is, the better model includes a level-2 variance component. In this test, the two candidate models have the same fixed covariates and are hierarchically nested because the more complex model has two covariance parameters and the simpler model has only one.

Applying the likelihood-ratio test, the $-2LL_R$ for the more complex model is 1130.8 compared with a $-2LL_R$ of 1136.6 for the simpler model. The difference in χ^2 of 5.8 is statistically significant (degrees of freedom (df) = 1, $p = 0.017$); the null hypothesis H_0 of no difference between the two models is thus rejected. The significance of the level-2 variance component indicates that the model's fixed variables do not account for the disparities between the regions. Because the fixed covariates of the two models are identical, the REML estimates are appropriate; ML estimates are not required.

Information criteria

The AIC and BIC also can be used to test the relative goodness of fit of two normal models that have the same covariates but different random effects. Although these statistics are derived from rather complex statistical theories, their computational formulas are rather straightforward. Both statistics adjust the maximized log likelihoods by a penalty factor that favors more parsimonious models; the BIC's penalty factor usually is more severe than the AIC's. Claeskens and Hjort (2008, 271) define these information criteria for multilevel models on the basis of the maximized residual log likelihood LL_R so that higher values indicate the better fit; here are their formulas:

$$\mathrm{AIC_{REML}} = 2LL_R(\hat{\theta}) - 2\mathrm{length}(\theta),$$

$$\mathrm{BIC_{REML}} = 2LL_R(\hat{\theta}) - \log(N - r)\ \mathrm{length}(\theta).$$

In these expressions, $LL_R(\hat{\theta})$ is the maximized REML log-likelihood with $\hat{\theta}$ being the maximizer, length(θ) denotes the number of estimated parameters, and $(N - r)$ is the rank of the transformation matrix for the REML estimates (N is the total number of observations and r is the length of the β vector).

SAS, however, now defines these statistics on the basis of the deviance so that the candidate model that exhibits the smaller value of the adjusted deviance fits the data better than the other models:

$$\mathrm{AIC_{REML}} = -2LL_R(\hat{\theta}) + 2\mathrm{length}(\theta),$$

$$\mathrm{BIC_{REML}} = -2LL_R(\hat{\theta}) + \log(N - r)\ \mathrm{length}(\theta).$$

In these expressions, $-2LL_R(\hat{\theta})$ is the deviance, which is then corrected by 2 times the number of model parameters (for the AIC) or by the rank of the transformation matrix times the number of model parameters (for the BIC).

For the illustrative data, the model with the two variance components has a deviance-based AIC = 1134.8, whereas the model that includes only the residual variance has a deviance-based AIC = 1138.6; the difference of −3.8 favors the more complex model because its AIC is smaller than that for the simpler model.

Using the deviance-based BIC, the more complex model has a BIC = 1136.6, compared with a BIC = 1141.4 for the simpler model, the difference of –4.8 also favors the more complex model. For these data, both the likelihood-ratio test and the fit statistics indicate that the between-region variance component is statistically and substantively significant.

Accounting for a Level-2 Variance by Classifying the Levels of the Random Variable

A statistically significant level-2 variance component can be tested for robustness by classifying (i.e., nesting) the level-2 random variable by theoretically relevant typological test factors that appear in the Class statement of SAS. A nesting variable is usually a fixed covariate; the variable that is nested never is a covariate. Continuing the above example, both the likelihood-ratio test and the model-fit criteria have established that the variance component that summarizes the regional random effects is statistically significant. This variance component is an expected mean squares that SAS estimates in statistically arcane ways (Littell et al. [1996] 2006, 60–61). Here, it is based on and summarizes the random-effect parameters for each of the 18 regions. When all of these regional random-effect estimates lack statistical significance, then the variance component between the regions will approach zero and will not be statistically significant. Inspection of the 18 regional random-effects estimates indicates that four regions have statistically significant parameters, these are Eastern Africa (14.5, $t = 2.08$, $p = 0.04$), Western Africa (13.4, $t = 2.1$, $p = 0.038$), Caribbean (16.9, $t = 1.95$, $p = 0.054$), and Western Asia (−16.84, t = −2.75, $p = 0.007$). These random effects contribute to the level-2 variance of 149.3 and to its statistical significance, as gauged earlier by the likelihood-ratio test and the measures of goodness of fit.

However, when the levels of the regional random variable are classified by this typology—the political system of the regions' constituent countries are fully democratic (1) versus not fully democratic (0)—then the regional means and their random effects exhibit less variability than the un-nested model described earlier, and the nested variance between the regions becomes not statistically significant using the likelihood-ratio test. In the statistical model that nests the random regions by the democracy typology, let y_{ijk} denote the value of the HDI of the k^{th} country of the j^{th} region of the i^{th} category of TYPE (here type is full democracy or not); $X_{\mathrm{m}ijk}$ denotes a covariate m (m not equal to TYPE), and β_{ijk} denotes its regression coefficient (these m covariates are culture, emancip, lodebt nochaos, and integrit). Then,

$$y_{ijk} = \mu + \beta_{m+1}\text{TYPE}_i + \Sigma\beta_m X_{\mathrm{m}ijk} + a_{j(i)} + e_{ijk}, \quad i = 1, 2, \ldots, \text{ t}, \quad j = 1, 2, \ldots, r, \quad k = 1, 2, \ldots, n_{ij} \tag{2}$$

where $a_{j(i)} \sim iid\ N(0,\ \sigma^2_{r(TYPE)})$ and $e_{ijk} \sim iid\ N(0,\ \sigma^2_c)$. The random effect $a_{j(i)}$ indicates that region j is nested in typology category i, it is assumed to be *iid* normally distributed with a mean of zero and a level-2 variance component $\sigma^2_{r(TYPE)}$. Except for the random statement, which now specifies that the type of democracy (dichofre) of the countries nests the region random variable, the syntax is identical to that specified earlier; the random statement now is:

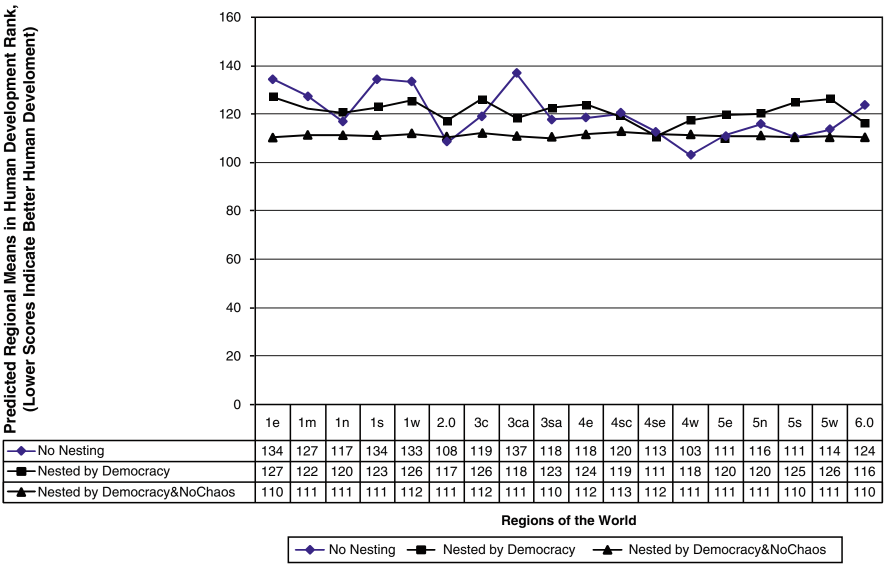
Random region(dichofre)/solution;

Nesting the regional random effect by an appropriate test factor here creates homogeneous region × test factor units, each one of which has a solution that includes the estimate, standard error, *df*, *t*-value, and significance probability. When the regional random effect is classified using the typology full democracy (1) versus not full democracy (0), SAS forms 30 region × democracy subgroups. Of these, only one, Western Asia × Not Full Democracy, exhibits a statistically significant random effect (-10.22, $t = 2.09$, $p = 0.04$). The between-region variance component nested by democracy, which is based on the effects in the 30 subgroups, is now only 46.8, whereas in the model with the un-nested random effect, the variance between the 18 regions was 149.3. The likelihood ratio test, which compares this nested model that includes two variance components with a near identical model that lacks the level-2 variance component, indicates that there is no difference between the models; the null hypothesis H_0 of no difference is not rejected ($\chi^2 = 1.89$, $p = 0.17$). The classification of the regional random variable by a country's level of full democracy reduces the between-region variance so that it becomes not statistically significant. This implies that differences in fullness of democracy among the countries grouped by the regions partly account for the regional differences in human development. Moreover, when the random regions are nested jointly by both full democracy and no chaos, the nested between-region variance disappears altogether.

Here, the un-nested model produces predicted regional means and their random effects that are more variable than the predicted means and random effects the nested models produce. The nesting of the regions by democracy flattens the means compared with the un-nested model, and the joint nesting by democracy and no chaos produces predicted means that are almost perfectly flat; see Fig. 5.1.[7] Furthermore, the estimates of the regional random effects depicted in Fig. 5.2 corroborate this pattern of increasing flatness when the random effects are nested. The flattening of these regional means and their random effects are reflected in the estimates of the level-2 variance component, respectively: un-nested = 149.3 ($p = 0.017$); nested by full democracy = 46.8 ($p = 0.17$); and nested jointly by democracy and no chaos = zero ($p = 0.85$), the latter results from the use of this random statement:

Random region(dichofre nochaos)/solution;

In sum, when the regional means and random effects reflect the empirical world, there are regional disparities in human development. When through statistical manipulations the countries' various amounts of democracy are used to form counterfactual homogeneous regional subgroups of countries, either singularly or jointly with no chaos, then the regional disparities in human development evaporate; apparently, these factors thus account for the regional differences.



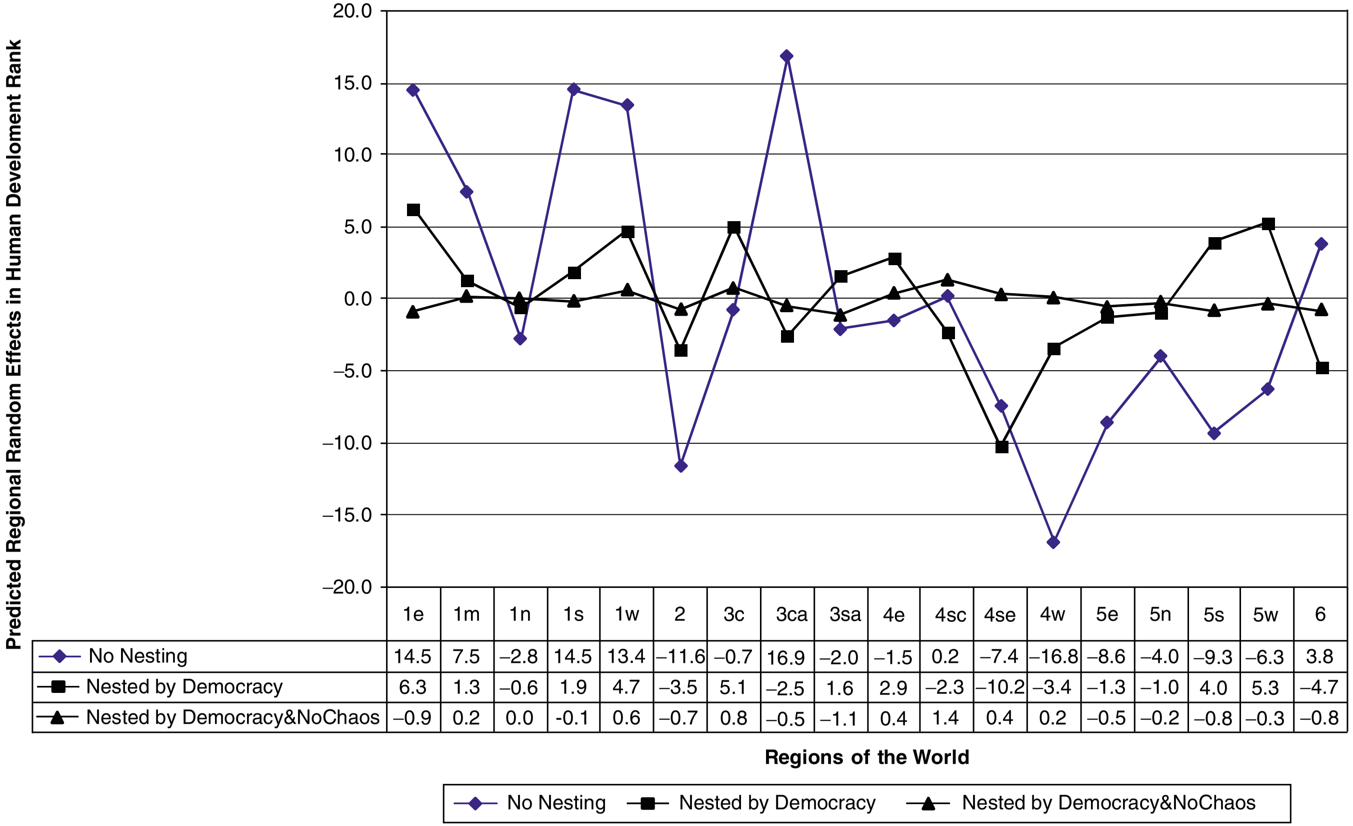
	1e	1m	1n	1s	1w	2.0	3c	3ca	3sa	4e	4sc	4se	4w	5e	5n	5s	5w	6.0
No Nesting	134	127	117	134	133	108	119	137	118	118	120	113	103	111	116	111	114	124
Nested by Democracy	127	122	120	123	126	117	126	118	123	124	119	111	118	120	120	125	126	116
Nested by Democracy&NoChaos	110	111	111	111	112	111	112	111	110	112	113	112	111	111	111	110	111	110

Fig. 5.1 The predicted regional means of human development are flattened by nesting (i.e. classifying) the regions by their countries' full democracy or by full democracy and no chaos

Counts of Events

The Poisson distribution can model such counts and rates of events as births, citations of articles, sales of technical books, accidents, murders, riots, bombs hitting targets, and hate crimes. Chapter 8 applies longitudinal multilevel Poisson models that test theoretical interpretations of the reasons why anti-Jewish violent events were taking place in contemporary European countries during the period of the second Palestinian uprising (i.e., intifada). These violent events are clustered within a country and are not normally distributed—the multilevel algorithm addresses the clustering and the Poisson model the lack of normality. In addition to the structural factor of the country's Muslin and Jewish population sizes, the chapter hypotheses three attitudinal factors that could engender these violent events: mobilization of the perpetrators by the events in the Middle East, cognitive ambivalence, and bystander indifference. In these analyses, the country is the level-2 unit and the repeated measures on the country are at level-1. A country's Jewish and Muslim population category—many Jews (100,000 or more) and many Muslims (500,000 or more), versus not many Jews and many Muslims—is a country-level attribute that has important effects on the level of violence.

Poisson regression models often are fit to data that are overdispersed; such models have sample variances that are larger than the sample means. An extra dispersion factor corrects the parameters of the tests of significance (e.g., the standard errors and confidence intervals) for over- or underdispersion, but even so a model with a dispersion factor very close to 1 is usually preferred to models with



	1e	1m	1n	1s	1w	2	3c	3ca	3sa	4e	4sc	4se	4w	5e	5n	5s	5w	6
◆ No Nesting	14.5	7.5	–2.8	14.5	13.4	–11.6	–0.7	16.9	–2.0	–1.5	0.2	–7.4	–16.8	–8.6	–4.0	–9.3	–6.3	3.8
■ Nested by Democracy	6.3	1.3	–0.6	1.9	4.7	–3.5	5.1	–2.5	1.6	2.9	–2.3	–10.2	–3.4	–1.3	–1.0	4.0	5.3	–4.7
▲ Nested by Democracy&NoChaos	–0.9	0.2	0.0	-0.1	0.6	–0.7	0.8	–0.5	–1.1	0.4	1.4	0.4	0.2	–0.5	–0.2	–0.8	–0.3	–0.8

Fig. 5.2 The predicted regional random effects are flattened by nesting (i.e. classifying) the regions by their countries’ full democracy or by their full democracy and no chaos

dispersion factors distant from 1. In Chapter 8, as additional explanatory variables are added to the model, the dispersion factor is reduced to an acceptable level for the small number of cases, 20 (10 countries at two points in time) and, in the replication, 50 (10 countries at five points in time). The initial intercepts-only model has an extradispersion factor of 3.81. When social structural and attitudinal variables are added to the model, the dispersion scale factor is reduced considerably so that it ranges from 1.16 to 1.32 for the three candidate final models. In the assessment of which of these three models fits the data the best, there is no clear-cut winner. But, considering a number of factors, this chapter concludes that of the three competitive explanatory factors, the mobilization of the perpetrators by the conflict in the Middle East and the cognitive ambivalence of ordinary Europeans are more pivotal than bystander indifference, although the latter is also relevant.

Clustered Microlevel Units

Sometimes, for reasons of expediency and tight budgets, researchers randomly select subunits of a large organization (level-2), and then administer a survey to all the members of those subunits (level-1). Most likely, the level-1 data are clustered and the standard errors will not take into account this clustering. The survey reported in Chapter 9 was administered to employees in six insurance claims offices. Because the employees in one office are more similar to each other than to employees in the other offices, these data are clustered. Model 1 does not take into account this clustering; it pools all of the data and merely regresses the employees' negative attitudes about computerization of fraud detection (the anti-index) on two fully standardized and orthogonal (i.e., uncorrelated) explanatory variables, the workers' receptivity to organizational innovations and their formal status in the organization. Symbolically,

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + e_i, \tag{3}$$

In this equation, Y_i is the score on the anti-index for the i^{th} employee; Z_{qi} is a fully standardized explanatory variable observed on the i^{th} employee; β_{q} (q $=$ 0, 1,. . ., Q). The constant term is β_0; β_1 and β_2 are regression coefficients, e_i is the residual or error term; and σ_{e}^2 is the variance of e_i, symbolically: $e_i \sim iid\ N(0,\ \sigma_{\text{e}}^2)$—normally distributed with a mean of zero and a constant variance of σ_{e}^2. When this model is estimated using REML $\sigma_{\text{e}}^2 = 0.50$ ($p < .0001$) with a confidence interval of 0.41 to 0.61. The effect of receptivity to innovation on the anti-index is -0.216 ($p < .0001$) and the effect of a higher status occupation is 0.094 ($p = 0.064$), which is not statistically significant. The $-2LL_R$ of this model is 421.2 and the BIC goodness-of-fit statistic is 426.5. As mentioned earlier, this statistic adjusts the $-2LL_R$ of the test statistic so that analysts can more readily select parsimonious models; when "smaller is better," a model with a smaller value of BIC is more acceptable than a similar model with a higher value of BIC.

Model 2 takes into account the clustering by specifying that the employees are grouped into one of the six offices symbolized by the subscript j (=1, 2, 3, 4, 5, 6):

$$Y_{ij} = [\beta_{0j} + \beta_{1j}Z_{1ij} + \beta_{2j}Z_{2ij}] + [\mathrm{u}_j + e_{ij}] \tag{4}$$

where Y_{ij} is the score on the anti-index of the i^{th} employee in the j^{th} office, $\beta_{\mathrm{q}j}$ (q = 0, 1,. . ., Q) are employee-level coefficients; $Z\mathrm{q}_{ij}$ is a fully standardized employee-level predictor q for employee i grouped in office j; u_j is a residual that varies randomly between offices; and e_{ij} is a residual that varies randomly for the i^{th} employee within the j^{th} office, symbolically: $e_{ij} \sim iid\, N\, (0, \sigma^2_{\mathrm{e}})$ and $u_j \sim iid\, N\, (0, \sigma^2_{\mathrm{o}})$. Equation (4) has two parts: the fixed part that comprises the covariates and their βs, and the stochastic part that comprises the random errors, one for each of the two-levels of data. A number of authors refer to the u_j as the random effect and the e_{ij} as the residual.

When SAS estimates this model by applying REML, the between-office variance $\sigma^2_{\mathrm{o}} = 0.1026$, with a lower bound of 0.036 and an upper bound of 0.89 (i.e., 0.036, 0.89); the within-office variance $\sigma^2_{\mathrm{e}} = 0.41$ (0.34, 0.51). The fixed parameters are about the same as in the regression model: –0.20 ($p < .0001$) and 0.095 (0.003, 0.187); but now the latter effect is also statistically significant ($p = 0.043$). Equation (4) uses one more parameter than Eq. (3), but the estimates of the deviance = 396.6 and BIC = 400.2 are considerably lower than that of Eq. (3)—by 24.6 ($p < .0001$, likelihood-ratio test) and 26.3, respectively—thus confirming the better fit of this two-level model. But the intraclass correlation is still substantial, $\rho = 0.20$ (0.1026/0.5139), and explanatory macrovariables are needed to explain this variance; see use 6.

Macro and Micro Explanatory Variables

Contextual analysts assess the joint effects on a response variable of variables that characterize higher and lower level units. Ideally, they quantify the cross-level interactions, but such contextual effects most often require large data sets and a fairly large number of level-2 units. To continue the above example, the number (zero, one, or two) of new computer systems being simultaneously installed in a claims office, which specifies the Office Type, will be introduced as a macrolevel explanatory variable. Chapter 9 provides the SAS syntax and develops the relevant model by combining micro- and macrolevel equations, resulting in Eq. (5).

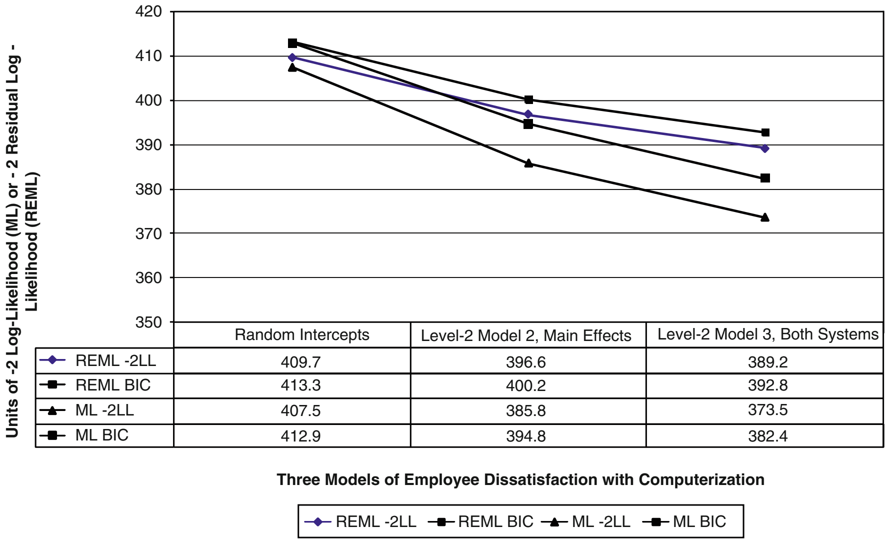
$$Y_{ij} = [\gamma_{00} + \gamma_{01}\mathit{OfficeType}_{1j} + \gamma_{10}Z_{1ij} + \gamma_{20}Z_{2ij}] + [u_{0j} + r_{ij}] \tag{5}$$

where $u_{0j} \sim iid\, N(0,\ \sigma^2_{\mathrm{o}})$ and $r_{ij} \sim iid\, N(0, \sigma^2_{\mathrm{e}})$—the latter parameters compose the stochastic part of this model and are enclosed in the second set of brackets of (5.5). This equation is analogous to Eq. (4) with only one exception: it includes the fixed effect of the macrolevel variable $\mathit{OfficeType}_{1j}$ which, depending on the office, may be an office with zero, one, or two new computer systems.

Two new computer systems were being installed simultaneously in two of the six claims offices. When the Office Type variable is set to 1 for those two offices and 0 for the other four offices, and the model estimated using REML, then the variance that is between office disappears. The $\sigma_{\rm o}^2 = 0.001$ $(p = 0.46)$ and the intraclass $\rho = 0.002$ (zero). The fixed effects on the anti-index are Office Type = Both New Systems = 0.61 $(p < .0001)$, receptivity to innovation = −0.19 $(p < .0001)$, and higher rank = 0.094 $(p = 0.04)$. This Model 3 fits the data much better than Model 2; its $-2LL_R = 389.2$ and its BIC = 392.8, compared with 396.6 and 400.2 for Model 2; smaller values are better.

When Model 3 is elaborated by including the cross-level interactions, the effects are small and not statistically significant. The Office Type × receptivity to innovation interaction = 0.02 $(p = 0.83)$ and the Office Type × rank in office interaction = 0.07 $(p = 0.49)$. These insignificant cross-level interactions imply that the negative effect on employee morale of the two new computer systems holds universally across the individual employees. Moreover, that the variable Both New Systems explains the variance between the offices suggests that the simultaneous introduction of two new computer systems may cause the employees' discontent.

The estimates of Eqs. (4) and (5) are based on REML, not ML, even though the covariates in these two models differ slightly. Model 3 includes the office typology whereas Model 2 does not; the fit statistics could be misleading. To test if the use of REML has led to an erroneous selection of Model 3 over Model 2, Fig. 5.3 plots the REML $-2LL_R$, the ML $-2LL$, and BIC statistics for three



	Random Intercepts	Level-2 Model 2, Main Effects	Level-2 Model 3, Both Systems
◆ REML -2LL	409.7	396.6	389.2
■ REML BIC	413.3	400.2	392.8
▲ ML -2LL	407.5	385.8	373.5
■ ML BIC	412.9	394.8	382.4

Fig. 5.3 Comparison of full maximum likelihood (ML) and restricted maximum likelihood estimates (REML) for three models: Random intercepts, employees in offices, and multilevel office effects with both systems installed in two offices

candidate models: the random intercepts model; Model 2, the multilevel model specified by equation (4); and Model 3, the multilevel model specified by Eq. (5). Although the REML statistics have higher values than the ML statistics, using any of these measures of fit, Model 2 is preferred to Model 1, and Model 3 is preferred to Model 2. Here, but not always, the ML and REML estimation procedures produce rather similar results. The chapters usually report the REML estimates because these are more accurate, especially for these small samples. If there is any doubt about the results, the analysis is replicated using ML and if there are changes in the covariates, then the new model is re-estimated under REML.

Evaluative Research

These three chapters on evaluations of comprehensive school reforms focus on how the test scores of underperforming elementary school students can be improved via the successful implementation of comprehensive reforms of their schools; these evaluations measure the various *effects of a cause*. Change agents from Co-nect, an educational consultancy, implemented these reforms by providing training, consultation, professional development, and new instructional strategies. However, the evaluators could not randomly select either the units that received the target program or the units that received the alternative treatments, so selection bias threatens the conclusions.

Estimating Treatment Effects in Difference-in-Differences Designs

To address this problem of selection bias, these chapters pool their data and apply difference-in-differences (DID) designs in which the various treatment groups serve as their own control. The schools in the target group receive the reform treatment, and the closely similar schools composing the comparison group receive the null treatment. The response variables are the students' achievement test scores aggregated to the level of the school at a point in time. As Chapter 10 explicates, the basic DID design includes observations on the two treatment groups at two time points, pre and post the intervention. Change is measured on the response variable in the target group by subtracting its mean-preperiod test score from its mean-postperiod test score; let this difference be symbolized as $d(1)$. Similarly, change is measured on the response variable in the comparison group by subtracting its mean-preperiod test score from its mean-postperiod test score; let this difference be symbolized as $d(0)$. Then, the DID average treatment effect δ_t is measured as the difference between these two differences: $\delta_t = d(1) - d(0)$.

Few, if any, examples in the research literature explicitly apply multilevel models to estimate the DID average treatment effect δ_t. Addressing this gap, the subsequent two chapters illustrate this procedure, which is use (7). Each chapter elaborates the basic DID design: The models in Chapter 11 have three treatment groups and illustrate borrowing strength (use 8); Proc Mixed provides the estimates for the models that have a normally distributed response variable. The models in Chapter 12 include propensity scores that aim to reduce bias (use 9); Proc Glimmix provides the estimates for the multilevel logistic regression models.

Borrowing Strength

Analyzing information on 15 of 31 elementary schools in the school system of Harford County, Maryland, Russell and Robinson (2000) compared the performance of each of the five target elementary schools that received the school reform treatment with the performance of two closely matched schools that received the null treatment. They matched the two comparison schools to their target school by taking into consideration these six characteristics: grades served by the school, total enrollment, the ethnic composition of enrolled students, percentage of students receiving free or reduced-price lunch (i.e., socioeconomic status), the percentage of students classified as Limited English Proficiency (LEP), and mobility rate.

The multilevel modeling of Chapter 11 borrows strength (use 8) by pooling the data for all 31 schools; grouping the schools on the basis of the presence or absence of the reforms as target, matched, or not matched; and estimating one equation per response variable that quantifies the effects of the covariates and the design variables.[8] The use of two comparison groups rather than one enables Proc Mixed to provide better estimates of program effects and the additional comparison group facilitates causal inferences.[9] The equation contains two program effect coefficients: one coefficient compares pre to post change in the target school with that of the matched schools; the other compares pre to post change in the not-matched schools with that of the matched schools. The significant difference in effects between the target and matched schools and the minimal difference in effects between the matched and not-matched schools strengthen the causal inferences. Compensating for the shrinkage of the effects of a school toward the overall solution for its treatment group, the data for each target school are weighted by its score for the quality of the implementation of the reforms.

Using Propensity Scores to Reduce Bias

Chapter 12 exemplifies use (9) by using propensity scores to reduce bias in the estimated effects of school reforms in elementary schools in Houston, Texas. Initially, there were a number of methodological problems: Because of the absence of random assignment of schools to the reform and comparison treatments, this study was vulnerable to uncontrolled selection bias, that is, the target schools may have been selected because they had prior characteristics that would enhance the performance of their students relative to the comparison schools. Because a school is a level-2 unit and the repeated aggregated achievement measures on the students of a school are the level-1 units, the data on each school are clustered. Because the response variables are binomial—conceptualized as the number of successes divided by the total number of trials—they are not normally distributed. Because the data are aggregated to describe the school in a unique school year, the patterns of interschool mobility and the achievements of the individual students are lost. Because the set of response variables includes the proportions passing standardized tests of reading, writing, mathematics, and the average of reading and mathematics, the response variables exhibit multiplicity that should be corrected.

This chapter ameliorates these problems by applying the following strategy. To mitigate selection bias, a propensity score (i.e., the probability that a given school received the reform treatment) was calculated for each school using data on all of the Houston elementary schools. Then, because the small number of target and comparison schools made matching on the propensity scores unfeasible, this variable was included with caution in the multilevel models as an additional mean-centered covariate (Winship and Morgan 1999, 677–678; Imbens and Wooldridge 2009, 33). Propensity scores used to match units or as a statistical control may reduce selection bias because these scores take into account the effects on the relationship between the treatment and the response variable of the covariates that produced these scores (Rosenbaum and Rubin [1983] 2006; Morgan and Winship 2007, 74–77).

Because the initial logit and probit models for deriving the propensity scores did not converge to a solution, the data were first linearized by using Goldberger's method (Achen 1986, 40–41). Then, by simply adding FIRTH as an option on the model statement of Proc Logistic, SAS applied Firth's (1993) penalized likelihood estimation procedure to the "Goldbergerized" data; the estimates now converged when the binary attribute for the reform treatment (0,1) was regressed on 16 prior school characteristics. After transformation of the logit-scale predicted values to form probabilities, the resulting propensity scores for the various units and the values of the other covariates were then centered by their overall means and used as mean-centered covariates in the multilevel models. This mean-centering modifies the intercept term so that it includes the effects of the covariates; the design variables are not centered by their means.

The design variables include a cross-sectional treatment (1) versus comparison school (0) indicator variable that distinguishes the cross-sectional difference between the two groups of schools and a typology that indicates the time period as 0, 1, or 2. This setup can define three DID treatment effects: The first gauges change from the Time 0 to Time 1; the second overall change from Time 0 to Time 2; and the third from Time 1 to Time 2.

The results (using corrections for the multiplicity of the response variables) suggest that the target schools generally improved their performance from Time 0 to Time 1, and then held steady from Time 1 to Time 2. Contrariwise, the comparison schools experienced a decline in achievement from Time 0 to Time 1. Extra teachers were assigned to low-performing schools to help the students prepare for the fifth grade tests. These extra teachers reduced the student-to-teacher ratios in the comparison schools and improved the performance of these schools from Time 1 to Time 2. By Time 2, there was little difference in test scores between the target and comparison schools except for the fourth grade writing test scores. Because the extra teachers did not influence the students' preparation for the fourth grade writing tests, the comprehensive school reforms improved these tests scores across the evaluation period. These findings suggest that either comprehensive school reforms a la Co-nect or smaller class sizes indicated by reduced student-to-teacher ratios can improve the performance of inner-city students.

Research Summaries

To guide social policies and to develop theories that describe empirical realities adequately, the scattered concepts and findings of social inquiries are often in need of consolidation and responsible critique. This book thus offers several examples of research summaries, two of which bear directly on problems of health care. By applying the fixed- and random-effects paradigm of meta analysis (use 10), Chapter 14 consolidates findings about the effectiveness of nurses who preauthorized and concurrently reviewed inpatient care. By comparing the procedures for assessing causality health scientists applied in Europe and the USA, Chapter 15 summarizes the research findings concerning the null causal effects of childhood vaccinations on autism. Building on this explication, the concluding chapter applies similar procedures to review and critique the evidence for the assumed causal relationships produced by the multilevel models of this book (use 11).[10]

Meta-analysis

To consolidate research results across studies, some meta-analysts apply regression analysis with the study (level-2) and its result (level-1) as the units of observation; the regression coefficient estimates the average effect size of the results across the studies. This approach has several weaknesses: the standard error of the summarizing regression coefficient will be incorrect and the variability of the findings from one study to another may indicate that the studies are not homogeneous. For example, a major meta-analysis tested for homogeneity of effects and found several studies whose effects were outliers. The analysts then deleted these outliers from their analysis and summarized only those studies whose effects were homogeneous. A better procedure may be to include these outlying observations in a random-effects model, as outlined next.

Rather than deleting the outlying observations, Chapter 14 applies the fixed- and random-effects meta-analytic paradigm of DerSimonian and Laird (1986) to consolidate findings bearing on health care costs. Along with the practical methods of Lipsey and Wilson (2001), this generic paradigm could be applied broadly in the social sciences.[11] The fixed-effects statistical model ignores variation that is between the effect coefficients of the studies; it assumes that there is one common program effect and that each study provides an estimate of that effect. The pooled effect is a weighted average of each study's effect; each study is weighted by the inverse of the squared standard error of its effect divided by the sum of such components across the studies.

The random-effects statistical model quantifies the interstudy variation by assuming that there is a population of studies from which those in the consolidation were sampled. It assumes that each study has its own true treatment effect that would result if that study were replicated an infinite number of times. The pooled estimate is the weighted average of the effects of each individual study, but now the weight for an individual study is the inverse of the sum of its squared between-study and within-study variance parameters, divided by the sum of such summed components across the studies.

This chapter consolidates the findings from eight studies of precertification of inpatient medical care and four studies of onsite review of inpatient medical care. These studies share a common DID study design. This design takes the difference between the pre- to post-period change in the means of the treatment group and the pre- to post-period change in the means of a similar comparison group. Many of the original reports presented their results in a four-cell table that takes into account the effects of numerous test factors and controls. The latter variables are centered by their overall sample means and this allows the intercept term of the regression equation to express their effects. The predicted means result from different combinations of four parameters: the intercept, the effect of time (postintervention versus preintervention), the cross-sectional effect (the target group versus the comparison group), and the interaction between the target group and the postintervention time period. To the extent that these studies controlled for selection bias and measurement error, they report the implied average causal effects of the programs and the confidence bounds of these effects.

Evidence for Causal Inferences

How does a conscientious person evaluate the evidence bearing on such claims of causal associations? Providing some guidelines, Chapter 15 examines the procedures for assessing the evidence for causality applied by a group of health scientists affiliated with the Cochrane Collaboration in Europe, and by a committee affiliated with the Institute of Medicine (IOM) in the USA. Using slightly different procedures, both groups reached similar conclusions about the null relationship between childhood vaccinations and such adverse effects as autism, atypical autism, and Asperger's syndrome; these maladies compose autism-spectrum disorders (ASD).[12]

The European reviewers applied meta-analytic procedures similar to those of Chapter 14. They conducted an exhaustive search resulting in 5,000 studies that loosely conformed to their selection criteria. From these, two-person teams narrowed the studies to 139 bearing on the relationship between the combined measles–mumps–rubella (MMR) vaccination and any adverse consequences. The investigators further reduced the pool of eligible studies to 31: They eliminated studies that reused the same basic data set several times resulting in redundant publications, and those that did not meet other selection criteria based on their assessment of the quality of a study. The Cochrane Collaboration prefers epidemiological evidence from published peer-reviewed randomized control trials, cohort studies, case-control studies, time series analyses, and so forth. Given these 31 studies, the reviewers scrutinized their designs, measures, and findings, identifying any biases. Across each study grouped by its methodology, the investigators reported that: "No credible evidence of an involvement of MMR with either autism or Crohn's disease was found "(Demicheli et al. 2005, 2).

The IOM committee combined a meta-analytic search for relevant studies with an informal Bayesian approach in which they initially took a neutral position with respect to a study's contribution to the causality argument. If warranted by the evidence, they would modify their initial position of neutrality one way or the other on the basis of their assessment of the validity of a study's design, measures, and findings. Prior to their review, they carefully formulated statements that they would use to summarize their conclusions about causality and the methodological requirements that the studies must fulfill in order to trigger a particular statement. For example, in order for the statement "The evidence favors acceptance of a causal relation" to be selected, then the following evidential requirements must hold: "The balance of the evidence from one or more case reports or epidemiological studies provides evidence for a causal relation that outweighs the evidence against such a relation." As the latter implies, these reviewers critiqued studies that reported no causal relation between childhood vaccinations and adverse consequences, as well as those that reported a positive causal association. Weighing the resulting balance of the evidence across the studies, they changed their initial position of neutrality. They unequivocally concluded that: The scientific evidence does not support a causal relationship between vaccinations and the MMR vaccine. Moreover, the scientific evidence does not support a causal relationship between vaccinations with thimerosal-containing vaccines and autism-spectrum disorders (IOM 2004, 7).

The concluding chapter of this book draws upon the procedures of these review groups to gauge the causal zone achieved by the evidence produced by the multilevel modeling of the chapters. It first differentiates a finding of no causality from a finding that a relationship is not informative with respect to causality: The former implies that the study is conducted with competence and is valid, but the relationship being examined does not indicate a causal relationship, most likely because the average treatment effect is not statistically significant. The latter finding implies that the study is not conducted with competence and is not valid; consequently, its results are irrelevant to the causality argument. Then, it briefly reviews the definitions of stable association, potential outcomes, and dependency networks, the three zones of notions of causality and their special cases.

To establish whether the results of a multilevel model are informative with respect to causality, the study should be valid. This concluding chapter explicates five aspects of validity: fit, construct, internal, external, and statistical conclusion. For each chapter that presents a multilevel model, the chapter initially assumes a position of neutrality with respect to causality. Then, it applies the criteria for assessing validities determining the causal zone achieved by each multilevel model. The reader is invited to apply these criteria, evaluating the evidence for causal inferences and forming his or her own conclusions.

Endnotes

[1] Cox and Wermuth (1996, 51) note that aggregating data: "over individuals at each of a number of time points, or over time for each of a number of individuals… can appreciably simplify analysis,… but care is needed in relating the properties of the aggregated data back to the behavior of individuals."

[2] Littell et al. (2006, 7) state:

> Of these [three procedures], ML is usually discouraged, because the variance component estimates are biased downward, and hence so are the standard errors that are computed from them. This results in excessively narrow confidence intervals whose coverage rates are below the nominal 1-α level, and upwardly biased test statistics whose Type I error rates tend to be well above the nominal α level. The REML procedure is the most versatile, but there are situations for which ANOVA procedures are preferable; Proc Mixed in SAS uses the REML approach by default, but provides optional use of ANOVA and other methods when needed.

[3] Claeskens and Hjort (2008, 271) state:

> The maximum likelihood method for the estimation of variance components in mixed models results in biased estimators. This is not only a problem for mixed models, even in a simple linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathrm{N}_N\,(0, \sigma_\varepsilon^2 I_N)$ the maximum likelihood estimator of σ_ε^2 is biased. Indeed $\hat{\sigma}_\varepsilon^2 = N^{-1}$ SSE$(\hat{\beta})$, while an unbiased estimator is given by $(N - r)^{-1}$SSE$(\hat{\beta})$, with r being the length(β). The method of restricted maximum likelihood produces unbiased estimators of the variance components.

[4] Holland (1986, 945) distinguishes the study of the "causes of effects" from "measuring the effects of causes." The potential outcomes perspective of the evaluative research studies of Part 3 focus on the multiple effects of a cause, whereas the more basic research studies of Part 2 focus on the multiple causes of a response variable.

[5] Philip Gibbs of the SAS Institute offers this advice (personal communication, April 12, 2007):

> You can only compare the likelihoods (and by extension the associated IC [information criteria] statistics) under REML if the models involved have the same set of fixed effects…. The usual progression in MIXED usually goes something like this. Come up with a set of fixed effects (your MODEL statement) that makes sense to you. You do not have to be statistically rigorous when you do this. Then, under REML and NOT changing the MODEL statement in this phase, make changes to the RANDOM and/or REPEATED statement(s) until you find a model that has a "best" AIC value. Now, switch over to ML estimation and change the effects on the MODEL statement until you come up with a "best" AIC value. When that is done, switch back to REML to report the final model.

[6] The term "deviance" most often is used in the context of models with a categorical response variable. This book sometimes uses this word in the context of models with normal response variables, as done here. At other times it just reports the −2 × residual log likelihood of REML estimation or the −2 × log likelihood of ML estimation.

[7] The following Estimate statements ask SAS to calculate estimates of the predicted regional grand mean, intercept 1 | region (dichofre), and the 18 predicted regional means when the levels of the random regions are classified by full democracy. There are 18 spaces on the right-most part of each estimate statement, one for each region; a 1 in a space tells SAS to estimate that region’s parameter whereas a 0 tells SAS not to estimate that region’s parameter. For the grand mean, there is a 1 for each region so SAS estimates the mean across all 18 regions.

```
estimate ‘mean′        intercept 1 | region (dichofre) 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1;
estimate ‘region 01e′  intercept 1 | region (dichofre) 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 01 m′ intercept 1 | region (dichofre) 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 01n′  intercept 1 | region (dichofre) 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 01s′  intercept 1 | region (dichofre) 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 01w′  intercept 1 | region (dichofre) 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 02′′  intercept 1 | region (dichofre) 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 03c′  intercept 1 | region (dichofre) 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 03ca′intercept 1 | region (dichofre) 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0;
estimate ‘region 03sa′intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0;
estimate ‘region 04e′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0;
estimate ‘region 04sc′intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0;
estimate ‘region 04se′intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0;
estimate ‘region 04w′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0;
estimate ‘region 05e′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0;
estimate ‘region 05n′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0;
estimate ‘region 05s′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0;
estimate ‘region 05w′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0;
estimate ‘region 06′′  intercept 1 | region (dichofre) 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1;
```

run;

[8] Singer and Willett (2003, 136) clarify the notion of borrowing strength as follows:

> Because OLS trajectories differ markedly from person to person, the model-based trajectories differ as well, but their discrepancies are smaller because the population average trajectories are more stable. Statisticians use the term “borrowing strength” to describe procedures like this in which individual estimates are enhanced by incorporating information from others with whom he or she shares attributes. In this case, the model-based trajectories are *shrunk* toward the average trajectory of that person’s peer group (those with the same predictor values). This combination yields a superior, more precise estimate.

[9] When commenting on William Cochran’s many contributions to observational studies, Donald Rubin (2006, 10) states that he and Cochran agree that several comparison groups are better than only one:

> A second theme in design is the need for a control group, perhaps several control groups (e.g., for a within-hospital treatment group, both a within-hospital control group and a general-population control group). The rationale for having several control groups is straightforward: If similar estimates of effects are found relative to all control groups, then the effect of the treatment may be thought large enough to dominate the various biases probably existing in the control groups, and thus the effect may be reasonably well estimated from the data.

[10] The ongoing controversies about the alleged causal effects of childhood vaccinations on engendering autism, and the ignoring of scientific evidence regarding global warming, suggest that ordinary citizens need practice in assessing evidence about causality.

[11] The present-day fragmentation of the social sciences leads to numerous interest groups, each one of which may have their own unique standards for judging the quality of a piece of work.

Consequently, intra-disciplinary conflicts over styles of research are common; to advance cumulative knowledge, shared standards are needed. Their absence leads to personal attacks that inhibit progress rather than collegial peer reviews that improve the work.

[12] The logical structure of this explication is similar to that of a multilevel model: The review groups in the two different geographical areas, Europe and the USA., are level-2 variables that encompass the studies that they review, which are level-1 variables, as are the conclusions reached by these two review groups. In essence, the research summary of Chapter 15 exemplifies a qualitative meta-analysis.

Part II
Contextual Studies

Chapter 6
Contexts and Covariates

Macroeconomists, like medical scientists, use case studies to teach their students about the maladies in which the system is susceptible. For supply shocks and stagflation, the example is the 1970s. The financial dislocations that occur when bubbles burst are illustrated by the Great Depression and Japan's problems in the 1990s. The importance of central bank credibility in resisting inflation emerges from discussions of the experience of the late 1960s and the 1970s.

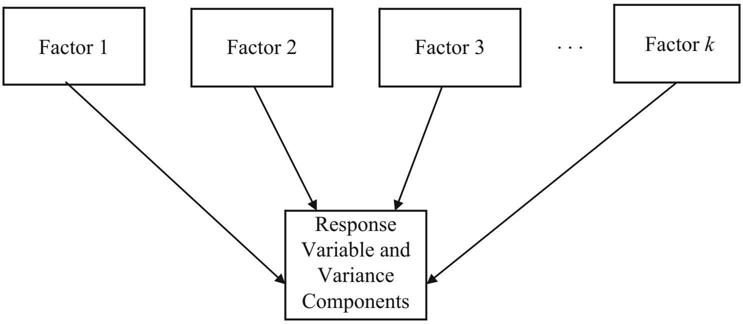
—Lawrence Summers (2008, 5)

Let us begin by describing the data that will be analyzed here for illustrative purposes.

—Leo Goodman (1972a, 28)

Exemplifying the general applicability of multilevel modeling for the analysis of contexts and their covariates, the three subsequent chapters in Part 2 examine how social contexts can influence human development, interpersonal violence, and technological change. Although the multilevel models of these chapters each have two levels, their constituent level-2 and level-1 units range from the very macro to the very micro. The first of these chapters studies individual countries that are grouped by the regions of the world; the second, repeated measures taken at different times that are grouped by their country; and the third, measures on employees that are grouped by their organizational unit. A similar multilevel study design unifies these diverse studies: the response variable and the covariance parameters are thought to depend on a number of antecedent factors (i.e., the contexts and their covariates), as illustrated by the following diagram.

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_6, © Springer Science+Business Media B.V. 2011



Each chapter applies a different approach for formulating appropriate multilevel models that can estimate the effects of the factors. Chapter 7 directly specifies multilevel models in which regions group the human development scores of their constituent countries. Chapter 8 estimates multilevel Poisson regression models in which European countries group their yearly counts of anti-Jewish violence. Chapter 9 combines an equation for an organization with an equation for the employee to derive the final equation for multilevel models in which claims offices group their employees and their employees' characteristics, especially negative attitudes toward computerization. Each chapter also illustrates a different tactic for uncovering implied causal effects: homogeneous subgroups, testing of competitive theoretical models, and causal equations.

For each topic of inquiry this present chapter discusses the motivating social problem, relevant theory, the study design, measures of the variables, results of the multilevel modeling, and implications for policy.

Global Human Development

The Social Problem

The unjust inequality of nations and regions of the world is perhaps the major social problem facing humanity at this time; some countries are very rich, others are very poor: One in every five persons—1.2 billion people—survive on less than one dollar a day. Many suffer from hunger and lack access to clean water and sanitation; child mortality rates are high and increasing; many people suffer from malaria, HIV/AIDS, and other diseases. These grim facts constrain the development of vast numbers of people (United Nations Development Program 2003, 1–13). Chapter 7 aims to uncover some pivotal determinants of a country's human development as defined by the UNDP.

Theory

The UNDP defines human development conceptually as the expansion of capabilities that widen people's choices to lead lives that they value (UNDP 2002a, 53–58). It measures this capacity by a country's position on the human development index (HDI). The HDI has three dimensions: *a long and healthy life*, indicated by life expectancy at birth; *knowledge,* indicated by a combination of adult literacy (weighted 2/3) and the combined gross primary, secondary, and tertiary enrollment ratio (weighted 1/3); and *a decent standard of living*, indicated by real Gross Domestic Product (GDP) per capita (PPP$)—purchasing power parity per capita in dollars. On the basis of the scores on this index, the UNDP ranks the countries by assigning equal-interval ordinal numbers to the composite measure based on the underlying scores.

After grouping the countries into the 18 regions of the world—these regions are the contexts of the countries—Chapter 7 uses measures of human development as the response variables in its multilevel models. Drawing upon the conceptions of Amartya Sen and Samuel Huntington, this chapter examines the effects on human development of six nominal and ordinal factors. Following Huntington, these include a country's rather stable and intrinsic dominant civilization zone defined as Western, Latin American, African, Islamic, Sinic, Hindu, Orthodox, Buddhist, or Japanese. Following Sen, these covariates also include the indicators of these five potentially changeable instrumental factors. These factors and their indicators (which are in parentheses) are: social opportunity (the absence of slavery); political freedom (full democracy); transparency guarantees (low corruption); protective security (the absence of conflict and social unrest); and economic facilities (not a heavily indebted poor country, i.e., not an HIPC).

Study Design

Because neighboring countries often have more similar HDI scores and rankings than countries that are distant, the observations are clustered and not independent, thus calling for multilevel modeling. Taking advantage of this clustering, Chapter 7 groups the 138 countries of the world for which the data are complete into one of 18 geographical regions; these regions (level-2) and the countries within a region (level-1) are the units of analysis. The chapter's descriptive and explanatory analyses then assess the estimates of the variance in HDI that is between regions ($\hat{\sigma}_r^2$), the variance between countries within a region ($\hat{\sigma}_c^2$), and the effects of the correlates and implied causes of the HDI. In the descriptive analyzes of HDI rank, the models simply link one response of the country, its HDI score, to the six covariates. These descriptive models quantify, but do not account for, the between-region variance in human development.

The explanatory models aim to account for the variance in a country's position on the HDI by using typologies to classify (i.e., nest) the random regions. The resulting region × typology subgroups most often exhibit less variability in the HDI than the unclassified regional groups. If this nesting by a typological variable reduces to statistical insignificance the random effects for these subgroups, then that typology accounts for the relationship between the different regions and the HDI. This procedure is similar to Cochrane's (1968) use of categorical test factors to form homogeneous groups that reduce bias and spuriousness, and to Lazarsfeld's (1955a) testing of an $x \to y$ relationship for stable association using contextual variables in an elaboration procedure.

Measures

The UNDP refers to the HDI and the regions of the world as objective measures; it refers to the other indicators as subjective measures. To form subjective measures, rating agencies gather qualitative and quantitative data for each country from such sources as nongovernmental organizations, the State Department's desk officers, travelers, business people, and newspaper reports; these observations form a file on each country. Trained raters assess this evidence and then, depending on the data in the file, they check off items on checklists. Points are assigned and a total score for each country is obtained. Most often, the raters group the total scores to form an ordinal typology—a classification variable—that has only a few categories, for example, high, middle, or low. Ideally, the rating agency assesses the reliability and validity of the scale or typology prior to releasing it to researchers and the general public. Measures such as these can serve as covariates and classifications in multilevel models.

Results

The first set of analyses explore the effects of the covariates on the two variance components ($\hat{\sigma}_r^2$ and $\hat{\sigma}_c^2$) and on HDI rank. A model that includes only the regional intercepts is estimated first. This intercepts-only model provides a benchmark for comparing subsequent models by producing the maximum variance in the HDI that is between regions ($\hat{\sigma}_r^2$), the maximum variance between countries within the regions ($\hat{\sigma}_c^2$), and the maximum values of the $-2 \times$ residual log-likelihood ($-2LL_R$) and goodness-of-fit statistics. Compared with the values of the baseline model, the model composed of all of the indicators of the covariates—civilization, slavery, freedom, conflict, corruption, and national debt—produces the largest reductions in the $-2LL_R$, the BIC, and the random effects; it is the best descriptive model in this set.

To uncover which of the covariates has the most important effect on the HDI, the analysis deletes sequentially from the full model each of the covariates one at a time. The deletion from the model of an important covariate will adversely affect the fit of the model as indicated by the increased values of $\hat{\sigma}_r^2$ and $\hat{\sigma}_c^2$ and by poorer values of the fit statistics, all compared with those values for the full model. The most important covariate is the one whose deletion from the full model results in the most unfavorable values of these indicators of fit. For this system of variables and their measures, a country's civilization zone has the most important descriptive effect on the HDI.

For the data of this chapter, two nesting variables—full political freedom (1) or its absence (0) and higher levels of the new slavery (e.g., debt bondage, forced labor, forced prostitution, and chattel slavery)—each reduce the variance between the regions so that the nested variance $\hat{\sigma}^2_{r(t)}$ is no longer statistically significant. These two factors thus account for the regional variation in the HDI rank for this system of variables and their measures. When this chapter replicates the main analysis using the underlying scores for the HDI, these two factors work in conjunction with civil disorder to account for the variance between the regions—civilization zone, national debt, and corruption also have important effects.

Policy Implications

Civilization zones matter but they do not irrevocably determine development outcomes. To enhance human development, the UN's social and economic commissions could encourage their countries to address the factors whose effects this research has documented: Countries could enhance social opportunity by working toward the elimination of debt bondage and other forms of slavery (especially via the education of children and adult females and males), nurture political freedom by moving toward fully democratic political systems, and enhance civil society by stopping the violence and social unrest that weakens solidarity and protective security. These changes would enable the operation of a mutually reinforcing system of beneficial relationships among political democracy, emancipative employment, and human development. This beneficial circular feedback process could be enhanced by the mitigation of such negative correlates of rank on the HDI as national debt, which constrains economic facilities, and corruption, which weakens transparency guarantees. Provision of these changes could provide national contexts that would encourage free human agency and strengthen social bonds, which in turn would further stimulate social and economic development.

A Globalized Conflect

The Social Problem

This chapter considers some hypothesized causes of contemporary violence against European Jews. Israeli Jews and Palestinian Muslims continue to kill each other in the Holy Land. Apparently, this conflict interacts with antisemitic and anti-Western propaganda to worsen contemporary anti-Jewish violence in Europe and elsewhere in our globalized world. Because of the unanticipated consequences of this violence, it is not only a problem for Jewish communities: It may desensitize the perpetrators, thus making them more susceptible to recruitment for participation in even more henious actions. It may stigmatize the Muslin communities, making their minority status in European and other countries even more tenuous. It may weaken the cohesion of the country in which the violence takes place, thus making consensual democratic politics more difficult to achieve. Chapter 8 aims to uncover some of the reasons why this violence has erupted, identify the perpetrators (often young Muslim males, neo-Nazis, and right-wing thugs), and develop a theory for explaining why ordinary Europeans, the majority of whom are Christians, do not vigorously protest this anti-Jewish violence.

Theory

Previous theoretical explanations that simply link antisemitic attitudes to anti-Jewish violence are very blunt. This chapter seeks a more nuanced explanation for the counts of anti-Jewish violence in 10 European countries during the period of the second intifada against Israel (circa 2000–2005). It considers four generative factors: a societal predisposition indicated by the Jewish and Muslim population sizes of a country, a key contextual variable; a stimulus or triggering variable, the perpetrators' perceptions of the events in the Middle East that may mobilize them to attack Jews; and two facilitating factors that characterize ordinary Europeans, their bystander indifference to the violence, and their cognitive ambivalence. The multilevel modeling suggests which of these factors are most crucial, and which are less relevant.

Study Design

To facilitate an intuitive understanding of the relationships that the multilevel models will subsequently quantify, this chapter first reports findings based on the close inspection of a series of tables and from qualitative reports about the perpetrators. The tables explore relationships among the violent events; the population typology; and indicators of attitudes bearing on antisemitism, Israel, and the Palestinian Authority. Ideally, explorations such as these should be conducted on a data set that is separate from the one used for statistical modeling. Because alternative data sets were not available, these explorations use counts of violence for three years (2001, 2002, and 2003). The main quantitative analysis does not use the counts for 2002; the replication uses data from 2001 to 2005. Because the counts of violence are higher for the countries with many Jews and many Muslims, and because there is little difference in violence between the other population categories, the initial analysis uses the dichotomized index; the replication uses all three categories.

Modeling the longitudinal counts of violence for each country, the chapter develops Poisson regression models that nest each country by its Jewish and

Muslim population category, the key contextual variable. As is customary in multilevel modeling, the chapter first reports a baseline model that designates the countries as random and estimates the variance of the random effect $d_{j(i)}$ between the countries j nested in their population category (i) and the variance of the residual random effect e_{ijk} that also takes into account the variability of the observations across the two points in time k. In this initial model on the natural logarithm scale, the variance of $\hat{d}_{j(i)} = 1.34$ and the variance of $\hat{e}_{ijk} = 3.81$; a lot of variability is unexplained. The analysis aims to find good-fitting, parsimonious models that reduce these variances to smaller and statistically insignificant values.

Measures

The analytic data set includes subjective measures of attitudes and objective measures from the census and agencies concerned about extremism. The construction of this data set illustrates how analysts can use the World Wide Web as a source of data that can form the empirical base for theory building and statistical modeling. Because the respondent-level surveys were not available, the initial analyses in this chapter are based on data from downloaded reports of aggregated survey results for the 10 countries the Anti-Defamation League (ADL) surveyed in 2002 (Time 1) and 2004 (Time 2). These surveys probed European attitudes about Jewish people and the conflict between Israel and the Palestinian Authority. For these 10 countries in a specific year, the chapter uses downloaded yearly counts of anti-Jewish violence taken from the website of the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism at Tel Aviv University. For the period of the second intifada (autumn 2000 to roughly 2005), these summary counts include major attacks like shootings, knifings, bombings, and arson; and major violent incidents like physical aggression without the use of a weapon and vandalism. The chapter also uses downloaded population counts from the European Union's census website and estimated Jewish and Muslim population percentages from the website of the World Jewish Congress and the Central Intelligence Agency's online fact book. All of these data were entered in an Excel spreadsheet and, on the basis of the population data, a typology was created that classifies the countries as having many Jews (100,000 or more) and many Muslims (500,000 or more); few Jews (less than 100,000) and many Muslims; or few Jews and few Muslims (less than 500,000). By uploading the spreadsheet into SAS, the data were ready for statistical analysis.

Results

The social structural model quantifies a country's predisposition toward anti-Jewish violence. It includes the time-period indicator and two fixed cross-sectional indicators—the population classification and a binary (0,1) indictor for Belgium, which has a high violence count given its population classification. The violence increased from Time 1 to Time 2 and much of the violence occurred in the three countries that have many Jews and many Muslims, namely Germany, the UK, and France.

The mobilization model combines this structural propensity toward violence against Jews with a triggering item that captures favorable attitudes toward the Palestinians, as indicated by the proportion of the country who agree that the Palestinian Authority truly wants a peace agreement with Israel. As the mutual violence between Israelis and Palestinians intensifies, this variable may capture the effect of the precipitating events in the Middle East as perceived through the lens of anti-Israel propaganda often emanating from the media of Arab countries. These perceptions of events provide the stimulus that combines with the structural predisposition to facilitate the flare-ups of anti-Jewish violence. This model fits the data better than the structural model and better than models that include traditional antisemitic stereotypes or disagreement with Israel's stance on peace.

But what factors allow this violence to take place? Models emphasizing the unresponsiveness of bystanders and the ambivalence of ordinary Europeans due to their conflicting attitudes toward both sides in the Middle East conflict (valuing Palestinian lives and valuing Jewish lives) fit the data almost as closely as the mobilization model. In sum, the chapter argues that a country's sizes of its Jewish and Muslim populations predispose it toward its count of violence. The events in the Middle East mobilize Muslim youths and others (skinheads and neo-Nazis thugs) to perpetrate the violence. As the replication with 5 years of data suggests, ordinary Europeans do not protest this violence vigorously because their ambivalence stemming from conflicting cognitive forces leads them to withdraw. Moreover, as bystanders to the violence through media reports, they become indifferent due to the mechanism of diffusion of responsibility. Taken together, these findings clarify why certain countries have the higher counts of violence.

Policy Implications

Obviously, peace between Israel and the Palestinians would likely stabilize or even reduce the high rates of anti-Jewish violence in Europe and elsewhere in our globalized world. Given that this peace is elusive, what else can be done? The results of the modeling suggest these recommendations: Jewish, Muslim, and Christian communities should work together to foster tolerance and mutual respect, and to discourage abusive behavior against Jewish people, Muslims, and other minorities. The scurrilous antisemitic propaganda originating from Middle Eastern countries and elsewhere should be blocked and countered. Moreover, the news reports of the mainstream media should strive for objective reporting of the events in the Mid-east, correcting their well-documented biases that are anti-Israel. Ordinary Europeans should realize that violence against Jewish people may predict more widespread violence against the larger society.

Will Claims Workers Dislike a Fraud Detector?

The Social Problem

The consequences of work-place automation for workers has been, and will continue to be, a pressing area of inquiry because such innovations as expert systems, robots, and artificial intelligence, along with the globalization of work, can lead to layoffs, unemployment, deskilling, and worker rebellion. Presenting a case study of the effects of an expert system, Chapter 9 elucidates the extent to which claims workers perceived a computerized fraud detector (CFRD) as a potential threat to the substantive complexity of their work and perhaps to their jobs. CFRD is an expert system that processes automobile insurance claims and assigns a suspicion score to questionable claims. Management hoped that it would reduce the prohibitive costs of automobile insurance fraud, a major form of white collar crime in the USA. Instead, the pilot implementation of CFRD in three claims offices triggered a rebellion: the claims workers expressed strong sentiments against computerized fraud detection. This chapter explains why this happened.

Theory

Following the writings of Karl Marx and Frederick Engels, sociologists (e.g., Blauner 1964) and humanists (e.g., Marcus 1974) have studied how management's control of the workplace can produce alienated workers. More recently, Melvin Kohn and his many colleagues use Weberian and Marxist conceptions to guide their survey research on the effects of work that is substantively complex. Because such patterns of work require thought and independent judgment—workers are not closely supervised and the flow of work is not routine—it encourages intellectual growth and autonomy. Contrarily, patterns of work that are externally imposed, repetitive, and routine stifle creativity and self-direction. Smith (2008b, 48–49) provides a brief overview of Kohn's extensive research program and relevant references.

Study Design

The initial study design closely matched three claims offices that received the CFRD treatment with three claims offices that did not receive the CFRD treatment. First-hand observations of the claims offices and focused interviews with claims workers informed the construction of a survey questionnaire, copies of which were distributed to all of the workers in the six offices. These questionnaires tapped the workers' attitudes about computerized fraud detection, knowledge about networks of people (claimants, lawyers, physicians) who commit insurance fraud, receptiveness to innovation, rank in the offices, and so forth. With indexes of the workers' receptivity to innovation and rank in the office controlled, the null hypothesis (H_0) predicted no difference between target and comparison offices on an index composed of negative responses to three survey questions about computerized fraud detection.

However, the field observations uncovered that in two target offices and in one comparison office a second, more comprehensive administrative computer system—Millennium 2000 (M2K)—also was being installed. Thus, the initial design was changed in mid-stream resulting in this new design typology: offices have both CFRD and M2K being installed, only CFRD, only M2K, only one system, and no new computer systems. Multilevel modeling would be applied to reduce any biases stemming from lack of balance and clustered data.

Measures

The explanatory structure of Chapter 9 defines worker attitudes as level-1 variables and their claims office as a level-2 variable. The claims offices are thought to be a random sample from the insurance company's universe of claims offices. These "random" offices are nested by the typology of the number of computer installations that the office is experiencing; this typology is the pivotal contextual variable. A factor analysis of answers to the questions of the survey uncovered two orthogonal constructs: the employee's receptivity to innovation and the rank of the employee's job. These covariates are the level-1 control variables.

Results

Primarily because of the disruption to their workspace engendered by the simultaneous introduction of two new computer systems, employees in offices with both new systems were strongly against computerized fraud detection, whereas employees in offices with only one new system differed only slightly in attitude from employees in offices with no new systems. Because of the threats CFRD posed to their occupational self-direction, identity as fraud-detection experts, and perhaps to their jobs, the higher status employees opposed computerized fraud detection. Moreover, employees who in general resist innovations disapproved of the fraud detector.

This chapter distinguishes implied causal relationships from those that are merely associational (Holland 1986, 945–958). Whereas the office typologies gauge the number and kinds of interventions and may connote causal effects, the scores of an employee on the two covariates are just fixed attributes of the employee. Because these static attributes are not manipulated variables, they are not conceptualized as causal (Rubin 1974; Sobel 1995, 17–26; Sobel 1996, 361–370). To clarify the causal assumptions, various endnotes to this chapter apply

Pearl's *do*(x) operator, which signifies setting $X = x$ in an ideal, closely matched experiment (Pearl [2000] 2009, second edition, 22–24). The original study design specified that CFRD would be the only manipulated variable; all of the other variables would be set to constant values. Since CFRD would be installed in three target offices and not in the three matched comparison offices, X would be set to $x =$ CFRD and no other variable would be manipulated. But in some offices, there happened to be another variable that was being manipulated, Z was inadvertently set to $z =$ the other new computer system. Thus, the broken study design of this evaluation of the fraud detector is denoted *do*(x, z) and not *do*(x). This change necessitated the redesign of the study so that each type of treatment group would have the same treatment; that is, the stable-unit-treatment-value assumption (SUTVA) would not be violated.

Policy Implications

The results of this study suggest that management should introduce new computer systems one at a time and not simultaneously. Moreover, these systems should at least preserve if not increase the substantive complexity of work. If these systems can in fact reduce the number of workers required to perform the needed tasks, then management should retrain the displaced workers and use attrition and not layoffs to reduce the labor force.

. . .

Discussions of causality contrast the study of the *causes of effects* (Heckman 2005a, 2) with the study of the *effects of a cause* (Holland 1986, 945; Morgan and Winship 2007, 280–282). The next three chapters focus on estimating the implied causes of effects; that is, the relative effects of multiple factors on a response variable (Coleman 1964, 116, 189–240). The chapters in Part 3 focus on the effects of a cause; that is, the effects of an intervention on one or more outcome variables (Rubin 1974).

Chapter 7
Global Human Development

Without ignoring the importance of economic growth, we must look well beyond it... Development has to be more concerned with enhancing the lives we lead and the freedoms we enjoy.

—Amartya Sen (1999, 14)

In the recent past, strategies for ameliorating the problems of poor countries had polarized. Some strategists emphasized rapid economic development, arguing that without it all that can be redistributed is poverty. Other strategists advocated the equitable distribution of economic benefits, arguing that rapid economic growth may diminish the quality of life: it may change traditional cultures, create inequalities of wealth, and pollute the environment. Toward resolving this conundrum and spurring development, the United Nations hosted the Millennium Summit in September 2000 at which 189 countries adopted the Millennium Declaration that established these eight development goals to be achieved b1y 2015 or earlier: (1) eradicate extreme poverty and hunger; (2) achieve universal primary education; (3) promote gender equality and empower women; (4) reduce child mortality; (5) improve maternal health; (6) combat HIV/AIDS, malaria, and other diseases; (7) ensure environmental sustainability; and (8) develop a global partnership for development (UNDP 2003). Achievement of these goals would vastly improve the human development of the nations of the world.

Human development, as defined by Haq (1995, 13–28) and Sen (1999, 13–24, 2001, 506–513), is the enrichment of the choices people will have for leading a decent, secure life. Thus defined, human development implies more than economic growth; it has social components as well. Guided by this multidimensional perspective, the United Nations Development Program (UNDP) created its human development index (HDI) that combines in one score a measure of a country's level of economic development, indicated by its income per capita, and its level of social development, indicated by measures of its literacy and longevity (Mehrotra 1997a, 21–24, 1997b).[1] Although the UN data are plentiful, few academic studies have statistically analyzed the determinants of rank on the HDI and on its components for countries and global regions.[2]

Aiming to help facilitate the achievement of the Millennium development goals, this chapter asks: By how much, and why, do regions of the world exhibit disparities in human development? This chapter offers answers to this two-part question

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_7, © Springer Science+Business Media B.V. 2011

by analyzing a multilevel database of variables on countries and on the global regions to which the countries belong. It conceptualizes the HDI as measuring the substantive freedoms a country provides its citizens (Sen 1999). Substantive freedoms are capacities for health, education, and economic well-being that people may claim simply because they are human beings.

This chapter describes the effects of the following instrumental factors that may facilitate or inhibit these substantive freedoms (Sen 1999, 10): *Social opportunity*, assessed by the country's absence of contemporary slavery—debt bondage, forced labor, prostitution, chattel slavery, and perhaps bride price—all of which imply the absence of emancipative employment. *Political freedom*, indicated by the country's level of democracy. *Economic facilities* indicated by whether or not the country is highly indebted and poor. *Protective security*, indicated here by its level of internal conflict and social unrest. *Transparency guarantees*, measured by perceptions of corruption. Because *civilization zones*, as defined by Huntington (1968), are based on religious beliefs, ethnicities, and cultures, and these may affect development, this chapter probes how the different zones bear on the HDI.

To account for why the regions of the world exhibit different levels of human development, this chapter introduces typologies into multilevel models. Following UN definitions, it groups countries (the level-1 units) into regions (the level-2 units) and assesses the fixed factors that predict a country's position on the HDI and that reduce the country-to-country variance within these regions. Then, consistent with George and Bennett's advocacy of "middle range" theory (2005, 235–239), it introduces qualitative typologies that may account for the region-to-region variability. It distinguishes associational (i.e., non-causal) from putative causal effects as follows: it classifies the regional random effects by typologies creating homogeneous region × typology subgroups in which the estimates of the level-2 random effects may become statistically insignificant and thus accounted for. It applies Bonferroni adjustments (hereafter, Bon.) to tighten the significance levels of any atheoretical multiple comparisons. It assesses the effects of the key variables on the ranking of the countries and replicates these findings using their underlying scores, calculating effect sizes when the latter are analyzed.

Literature Review and Hypotheses

Although this chapter cites many previous contributions, it primarily draws on Huntington's *Clash of Civilizations* and Sen's *Development as Freedom* for its theoretical framework. These studies exemplify different conceptual modes of explanation: intrinsic cultures versus potentially manipulable instrumental factors.

Civilizations

Huntington's ([1996] 1997, 26–27) map of the world depicts his characterization of the dominant civilization of each country as Western, Latin American, African (i.e., non-Muslim sub-Saharan), Islamic, Sinic (Chinese), Hindu, Orthodox, Buddhist, or Japanese. This chapter refers to these characterizations as *civilization zones* because not everyone in a country necessarily shares the same civilization, religion, and culture.[3]

These zones may be distinct geographically, but, depending on the measure used, all civilizations share some concern for human rights (Sen 1999, 227–248). The data of this present study suggest the following: On an index of observed human rights abuses the West has the lowest score and the Sinic the highest score. Index scores for total trafficking—the involuntary smuggling of people between countries—which are based on Kevin Bales's measures (2002, 84–85), suggest that the Hindu category has the highest unfavorable score, followed by the Orthodox and Japan. Moreover, certain affluent zones (Japan, the West) are net importers of people for inappropriate purposes whereas the poorer zones (African, Buddhist, Hindu, Islamic, Latin, Orthodox, and Sinic) are net exporters. The Freedom House measure of civil liberties implies that the Hindu zone has the third most favorable score after the West and Japan. The ratio of males to females favors males in only Islamic, Hindu, and Sinic zones.

Each zone thus exhibits strengths and weaknesses regarding human rights. Because human rights and human development are mutually reinforcing, their measures are correlated.[4] Consequently, civilization zones, which are largely defined by religion, may influence the social aspects of development. Given the linkages between religion and economic development (Barro and McCleary 2003), these zones may also influence the economic aspect of development:

> *Hypothesis 1.* A country's civilization zone influences its level on the HDI, a key indicator of its substantive freedoms. Japan and the West may exhibit higher levels; the other zones, lower levels.

Instrumental Factors and Substantive Freedoms

Sen (1999) postulates an individual actor who wants to express her agency by making informed decisions and actions. She thus desires freedom, which is her intrinsic right. Development is the enhancement of the person's substantive freedoms, which are basic capacities to live the life that one has reason to value. Because the desire for freedom is intrinsic to all humans, all people have the right to at least enabling levels of capacities for social and economic development indicated by literacy, health care, and economic resources beyond poverty. Each of these aspects of human development can be assumed to vary on a scale from 0 to 1; their average is the total substantive freedom capacity that a society provides its citizens.

The individual with her level of resources wants to exercise her agency but is constrained by certain societal blockages that limit her pursuit of freedom and, in the aggregate, her country's human development score. Development involves the removal of these blockages so that the agency of its citizens is enhanced. Thus,

> *Hypotheses* 2a.–2e. The following instrumental factors can enhance a country's levels of substantive freedoms:

2a. Social opportunities (access to education, health facilities, and the free labor market, which imply the absence of contemporary slavery)
2b. Political freedoms (protected consultation—political and civil rights)
2c. Economic facilities (a society unencumbered with debt that provides opportunities for participation in employment, trade, and production)
2d. Protective security (a lack of civil conflict and unrest)
2e. Transparency guarantees (elite honesty and lack of corruption)

Conceptually, these instrumental factors can be assumed to vary on a scale from 0 (restricts freedom) to 1 (enhances freedom): their average score gauges the level of instrumental factors supportive of freedom that the society provides its citizens. The cross-tabulation of these measures locates a country according to its average substantive freedom score (assessed by the HDI) and its average score on the instrumental factors. Some societies (Norway, Iceland, Sweden, Australia, and the Netherlands) are developed, that is, free, and have scores near 1 on both the instrumental factors and on substantive freedoms. Some societies (Burundi, Mali, Burkina Faso, Niger, and Sierra Leone) are underdeveloped and unfree, having scores near zero on both measures.

Developing societies may take different paths toward the zone of free, developed societies (Sen 1999, 41–46). Some societies follow a process of development based on social supports, providing their citizens initially with higher levels of substantive freedoms (health care, education, and an economic safety net) and lower levels of instrumental factors (primarily due to weak market-based economies). Sri Lanka, pre-reform China, Costa Rica, and Kerala exemplify this path—they experienced rapid increases in social development without much economic growth. Some other societies follow a process of development based primarily on the enhancement of economic growth and less on political freedom (and also less on enhanced substantive freedoms). South Korea and Taiwan exemplify this path—they experienced high economic growth, which drove their growth in social development.

Figure 7.1 depicts these hypothetical relationships—the *x*-axis plots a country's average score for instrumental factors supportive of freedom; the *y*-axis, its score for substantive freedoms. Near the intersection of high scores on both dimensions is the zone of developed countries whose citizens enjoy freedom and agency. Near the intersection of low scores (0,0) is the zone of the underdeveloped and unfree countries. The graph depicts two idealized paths toward the zone of free, developed countries. The support-led process of development begins by increasing literacy, health, and economic supports that alleviate poverty. Based on these resources, the process of enhancing instrumental factors begins, eventually reaching the development zone of free societies. Japan exemplifies this road to development. The process led by economic growth begins with a minimal level of substantive freedom capabilities, but market mechanisms or a planned economy stimulate economic growth. However, the lack of political freedoms may constrain the upward surge in substantive freedoms; the average of the instrumental factors must be near 1 in order for a country to reach the zone of freedom. Brazil, post-reform China, and Russia follow this path but have not yet reached the zone of free, developed countries. Although this process may unfold over a period of years, a cross-sectional analysis of HDI scores later in this chapter confirms the correlations between instrumental factors supportive of freedom and substantive freedoms.

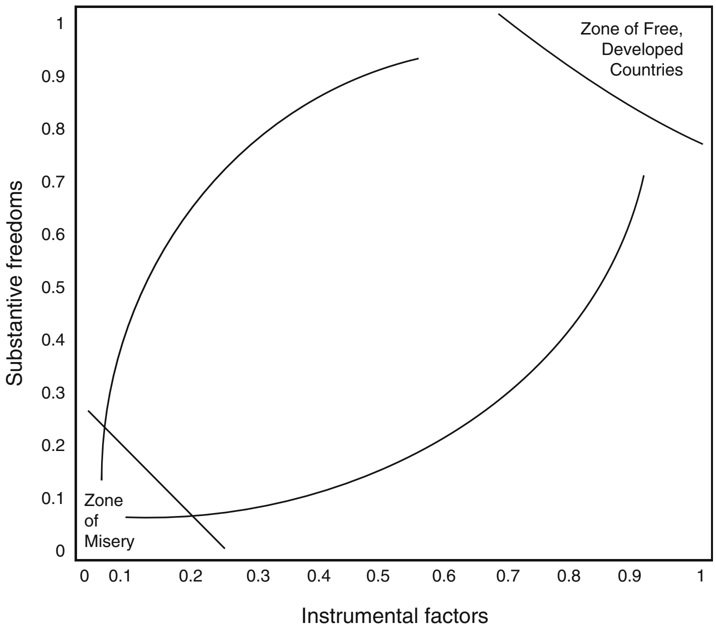
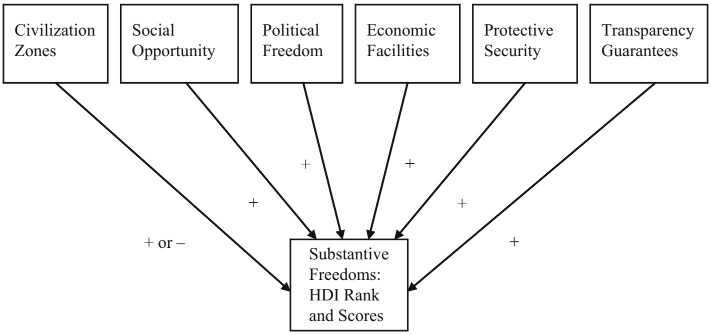


Fig. 7.1 Paths toward human development

Research Methods

The Multilevel Study Design

The above theory suggests a study design in which countries are the primary unit of analysis, the HDI position of a country is the response variable, and the civilization zone and the instrumental factors of the country are the covariates; that is, multiple factors influence a response, as this diagram depicts:



A key research goal is to describe which of these factors have the largest direct effects on the HDI, an indicator of the substantive freedoms a country provides its citizens. However, because these data are clustered within global regions and the observations among countries within a region are more similar to each other than to the observations on countries in distant regions, a multilevel model is needed that can produce valid estimates of the effects, their standard errors, and confidence intervals. Thus, the two levels of the data are countries and their properties (level-1), which are contained within the regions of the world (level-2).[5] To achieve this initial research goal, the chapter applies descriptive multilevel linear models (MLMs) that provide appropriate estimates of the effects of the covariates and their standard errors and also of the two variance components used in statistical inference: respectively, σ_r^2, which is the variance in human development that is between the regions, and σ_c^2, which is the variance in human development of countries within the regions.

A second key research goal is to explain why the different regions exhibit different levels on the HDI. To achieve this goal, this chapter views the variance components as response variables whose values are influenced by the fixed effects of the covariates and by the classification (i.e., nesting) of the regions by typologies; some of which may account for the regional random effects by reducing the variance between regions to insignificance. In this way the chapter conducts both descriptive and explanatory analyses of the determinants of human development, based on these data and measures.

Measures

In all of the models, the measures of the covariates are either prior in time to, or on equal footing with, the 1999 HDI. The UNDP derives a country's HDI value from aggregated objective observations on individual citizens thus forming an analytical property of the countries (Lazarsfeld and Menzel [1961] 1972, 227–228). The countries' levels of the empirical indicators of their civilization zone and instrumental factors—slavery, democracy, debt, conflict, and corruption—are based on experts' ratings that form nominal and ordinal typologies. In the explanatory models that account for the variance in the HDI that is between the regions, some of these typologies, when used to classify the regions, do in fact reduce to insignificance the differences between the regions.

Substantive Freedoms

The basic capacities for substantive freedoms that a country provides its citizens are gauged by its rank on the 1999 HDI, which is the main empirical response variable; future studies will focus on longitudinal change. This index has three equally weighted dimensions (Haq 1995, 46–72): A *long and healthy life* as indicated by life expectancy at birth; *knowledge* as assessed by a combination of adult literacy (weighted 2/3 by the UNDP) and the combined gross primary, secondary, and tertiary enrollment ratio (weighted 1/3 by the UNDP); and a *decent standard of living* as indicated by real gross domestic product (GDP) per capita (PPP, US$)—purchasing power parity (PPP) per capita in dollars. Income serves as a proxy for all the other aspects of human development not covered by a long and healthy life and knowledge.

This formula can express a country's performance on each of these dimensions as a value between 0 and 1:

$$\text{Index dimension} = \frac{(\text{Actual value} - \text{Minimum value})}{(\text{Maximum value} - \text{Minimum value})}$$

The UNDP posits these maximum and minimum values, respectively: life expectancy at birth in years (85, 25); adult literacy rate (100%, 0%); combined gross enrollment ratio (100%, 0%); and GDP per capita, PPP US$ ($40,000, $100)—since 1999, the UNDP takes the natural logarithm of this economic component. The use of the logarithm of income implies that a decent level of human development does not require unlimited income.

Human development is indicated by the simple average of the scores for each dimension and thus ranges from 0 to 1. Countries with high human development in 1999 had an average score of about 0.91, those with medium human development had an average score of about 0.68, and those with low human development had an average score of about 0.44. On the basis of these scores the UNDP ranks the countries by assigning equal-interval ordinal numbers: the country with the highest score is assigned the number 1 and the country with the lowest score is assigned 162, there are no ties. SAS treats the HDI ranking as if it were a normally distributed, continuous variable.[6] Because the response variable is a ranking and the predictors are categorical, the resulting statistics are similar to those in the Spearman (r_s) family of statistics (Smith 1985b, 1986).

Regions

The United Nations has created economic and social commissions that provide technical assistance, training, and capacity building to their constituent regions and to their member states. These commissions are coordinated to the following geographical areas, their constituent regions are in parentheses: Africa (Eastern Africa, Middle Africa, Northern Africa, Southern Africa, and Western Africa); Asia and the Pacific (Eastern Asia, South-Central Asia, South-Eastern Asia, and Oceania); Europe (Eastern Europe, Northern Europe, Southern Europe, and Western Europe); Latin America and the Caribbean (Central America, Caribbean, and South America); and Western Asia (Western Asia). Grouped within these regions, and also within North America, are the countries this study analyzes. Table 7.1 lists the 18 regions, the 138 countries that have complete data, the regions into which these countries are grouped, the reliability of the sample mean for each region, and each country's attributed civilization zone.[7]

Table 7.1 138 countries in 18 regions, reliabilities ($\hat{\lambda}$) and civilization zones in parentheses[a]

Eastern Africa ($\hat{\lambda} = 0.92$): Burundi (A), Djibouti (I), Eritrea (I), Ethiopia (A), Kenya (A), Madagascar (A), Malawi (A), Mauritius (A), Mozambique (A), Rwanda (A), Tanzania (A), Uganda (A), Zambia (A), Zimbabwe (A)

Middle Africa ($\hat{\lambda} = 0.87$): Angola (A), Cameroon (A)**,** Central African Republic (A), Chad (I), Congo (A), Democratic Republic of Congo (A), Equatorial Guinea (A), Gabon (A)

Northern Africa ($\hat{\lambda} = 0.81$): Algeria (I), Egypt (I), Morocco (I), Sudan (I), Tunisia (I)

Southern Africa ($\hat{\lambda} = 0.81$): Botswana (A)**,** Lesotho (A), Namibia (A), South Africa (A)**,** Swaziland (A)

Western Africa ($\hat{\lambda} = 0.93$): Benin (A), Burkina Faso (A), Cape Verde (W), Cote d'Ivoire (A), Gambia (I), Ghana (A), Guinea (I), Guinea-Bissau (I), Mali (I), Mauritania (I), Niger (I), Nigeria (A), Senegal (I), Sierra Leone (A), Togo (A)

North America ($\hat{\lambda} = 0.63$): Canada (W), USA (W)

Central America ($\hat{\lambda} = 0.81$): Costa Rica (L), El Salvador (L), Guatemala (L), Mexico (M), Panama (L)

Caribbean ($\hat{\lambda} = 0.81$): Barbados (L), Dominican Republic (L), Haiti (W), Jamaica (W), Trinidad and Tobago (L)

South America ($\hat{\lambda} = 0.91$): Argentina (L), Bolivia (L), Brazil (L), Chile (L), Colombia (L), Ecuador (L), Guyana (H), Paraguay (L), Peru (L), Suriname (W), Uruguay (L), Venezuela (L)

Eastern Asia ($\hat{\lambda} = 0.77$): China (S), Japan (J), Republic of Korea (S), Mongolia (B)

South-Central Asia ($\hat{\lambda} = 0.90$): Bangladesh (I), Bhutan (B), India (H), Iran (I), Kazakhstan (I), Kyrgyzstan (I), Nepal (H), Pakistan (I), Sri Lanka (B), Tajikistan (I), Uzbekistan (I)

South-Eastern Asia ($\hat{\lambda} = 0.88$): Brunei Darussalam (I), Cambodia (B), Indonesia (I), Lao People's Democratic Republic (B), Malaysia (I), Philippines (W), Singapore (S), Thailand (B), Viet Nam (S)

Western Asia ($\hat{\lambda} = 0.92$): Armenia (O), Azerbaijan (I), Bahrain (I), Georgia (O), Israel (W), Jordan (I), Kuwait (I), Lebanon (I), Oman (I), Qatar (I), Saudi Arabia (I), Syrian Arab Republic (I), Turkey (I), Yemen (I)

Eastern Europe ($\hat{\lambda} = 0.88$): Belarus (O), Bulgaria (O), Czech Republic (W), Hungary (W), Poland (W), Romania (O), Russian Federation (O), Slovakia (W), Ukraine (O)

Northern Europe ($\hat{\lambda} = 0.77$): Denmark (W), Estonia (W), Sweden (W), UK (W)

Southern Europe ($\hat{\lambda} = 0.86$): Albania (I), Croatia (W), Greece (O), Italy (W), Portugal (W), Slovenia (W), Spain (W)

Western Europe ($\hat{\lambda} = 0.86$): Austria (W), Belgium (W), France (W), Germany (W), Luxembourg (W), Netherlands (W), Switzerland (W)

Oceania ($\hat{\lambda} = 0.63$): Australia (W), Papua New Guinea (W)

[a] Civilization zones based on Huntington (1996, 26–27): A = African, B = Buddhist, H = Hindu, I = Islamic, J = Japan, L = Latin, O = Orthodox, S = Sinic, W = Western (Haiti = W). The simple average of the regional reliabilities is $\bar{\lambda} = 0.83$. For Africa, $\bar{\lambda} = 0.87$; the Americas, $\bar{\lambda} = 0.79$; Asia, $\bar{\lambda} = 0.87$; Europe, $\bar{\lambda} = 0.84$, and Oceania, $\bar{\lambda} = 0.63$. If Mexico is classified as North American then the $\bar{\lambda} = 0.72$. If data for New Zealand were available then the reliability for Oceania would also be $\bar{\lambda} = 0.72$, sufficiently high.

Civilization zones

By carefully inspecting Huntington's world map ([1996] 1997, 26–27) and also distributions of religious affiliations, a nominal variable was created that classifies each country according to its dominant civilization zone á la Huntington, but without some of the nuances of his thoughts about Haiti (136–137), torn countries (139–154), the eastern boundary of Western civilization (157–163), and cleft countries (163–168)—these simplifications facilitate the statistical modeling. The reliability of this classification was checked by ascertaining its congruence with Beckfield's reading of Huntington's map; 87 of the 90 countries that were jointly and independently classified exhibited agreement.[8] A society's civilization zone influences its world polity ties (Beckfield 2003, 417). Latin American, African, Islamic, Sinic, Hindu, and Buddhist societies have significantly fewer ties to international nongovernmental organizations, and these differences have grown since 1960.

These civilization zones also exhibit lower levels on the HDI than the West and Japan. In the unconditional (i.e., no control variables) model, there are significant differences in rank between civilization zones ($\hat{\sigma}^2_{HC} = 1,610.8$, $z = 1.9$, $p < .029$), as well as country-to-country differences within a zone ($\hat{\sigma}^2_c = 2,556.8$, $z = 8.71$, $p < .0001$). Japanese ($p < .002$) and Western zones ($p < .0005$) have significantly enhanced levels on the HDI (i.e., lower rank scores)—thereby supporting the first hypothesis—whereas African ($p < .0001$), Hindu ($p < .019$), and Islamic ($t = 1.92$, $p = 0.057$) zones have significantly poorer levels (i.e., higher rank scores).

Restricted social opportunity: slavery

Patterson (1982, 13) defines slavery as "the permanent, violent domination of natally alienated and generally dishonored persons." Similarly, Bales (1999, 6) suggests that slavery is the total control of one person by another for the purpose of economic exploitation; most often the slaveholder uses violence or its threat to obtain compliance. More formally (Bales 2004, 4): "[Slavery is] a social and economic relationship marked by the loss of free will where a person is forced through violence or the threat of violence to give up the ability to sell his or her labor power." Contemporary slavery, as Bales conceptualizes and measures it (2002), includes debt bondage, in which a person becomes collateral for a loan (Shastri 1990), chattel slavery in which masters assert ownership over slaves, and the trafficking of people from one place to another for purposes of physical or sexual exploitation: unfree marriages, forced labor, slavery, and prostitution (Bales 2000, 73–134; Clark 2003). Slavery violates the 1948 Universal Declaration of Human Rights; the Supplementary Convention of 1956 that aims to abolish servile status (debt bondage, serfdom, unfree marriage, and exploitation of children for their labor); the Economic, Social, and Cultural Covenant of 1966 on the freedom to choose work; and the Rome Final Act of 1998 that aims to abolish the trafficking of people. Slaves have very restricted social opportunities and countries with many slaves have weak emancipative values.

Bales developed his ordinal scales of slavery by collating documentary evidence and his own observations about trafficking and the number of slaves in a country. From these data, he created—and this study uses—a four-category ordinal classification of a country's level of contemporary slavery; its correlations with validating variables are consistent with expectations.[9] Countries Bales characterized as having higher levels of slavery have these correlates: unfavorable ratios of female to male primary education; human rights violations (higher levels of life integrity violations, abuse of human rights, and trafficking of humans); restricted political freedoms (lower levels of gender empowerment, civil rights, and political rights); and poverty (lower levels of literacy, enrollment in schools, life expectancy, and GDP per capita). All of these bivariate Spearman rank correlations (hereafter, r_s) are statistically significant ($p < .0002$). Bales (1999, 9) tallies at least 27 million slaves in the world, and "perhaps 15 to 20 million, is represented by bonded labor in India, Pakistan, Bangladesh, and Nepal."

Slavery obviously reduces human development because it prevents children and adults from being educated, reduces the per capita income of the poor, and shortens the life span; but the relative sizes of its impacts compared with those of the other covariates are not obvious. Consistent with hypothesis 2a, countries with higher levels of contemporary slavery have lower levels of human development ($F = 23.9$; $p < .0001$). The unconditional least squares means are: none or very little slavery = 1, then HDI rank is 43.6; some slavery = 2, then, 89.1; high slavery = 3, then, 111.7; and highest slavery = 4, then 110. Countries with slavery = 1 have significantly better HDI values than countries with any of the other categories ($p < .0001$, Bon.); countries with slavery = 2 have slightly better values than those with scores of 3 or 4 (unadjusted $p = 0.018$; $p = 0.11$, Bon.); and there is no significant difference in human development between countries with slavery = 3 and those with slavery = 4 ($p = 1$, Bon.). These categories thus may be dichotomized as emancipative employment (1 = categories 1 or 2) versus slavery (0 = categories 3 or 4).

Political freedom: democracy

Democratic regimes, as defined conceptually by Tilly (2007, [1995] 1997, 205; 2000, 4–5), have high levels of protected consultation: such regimes sustain broad and equal citizenship, provide binding consultation with their citizens concerning governmental activities and personnel, and protect their citizens from arbitrary governmental agents and actions. Increased protective consultation defines increased democratization. Political rights tap breadth of participation, equality of participation, and consultation; civil liberties tap protection from arbitrary governmental actions and agents (Tilly 2004). Broad and equal political rights enable citizens to participate freely in political processes that culminate in binding consultation—free, competitive elections. This implies that adults have the right to vote and to compete for political offices. Civil liberties allow freedom of discussion, assembly, and demonstration; free and independent media; an independent judiciary; the rule of law in civil and criminal matters; trade unions and collective bargaining; secure property rights; freedom of religion and personal freedoms; freedom from corruption and from an intrusive government; and freedom from exploitation. Dahl's (1989, 221–222) institutions of polyarchy and the democratic process comprise all of the above aspects of political freedom.

This chapter gauges democracy by using the 1999 Freedom House composite index that combines their collinear ($r_s = 0.92$) seven-category political rights and civil rights scales—it groups countries as fully democratic (score = 1), partly free (score = 2), or unfree (score = 3).[10] This trichotomy strongly correlates with validating indicators of democracy that the *Human Development Report 2002* tabulates (Table A1.1, 38–41). These include the democratic polity index (10, democratic to −10, authoritarian), freedom of the press, voice and accountability, political stability, rule of law, and governmental effectiveness. All of the bivariate r_s are statistically significant ($p < .0001$).[11]

Although democracy and human development reciprocally interact (Paxton 2002), and economic development sustains democracy (Lipset [1959] 1981, 27–63; Barro 1999; Welzel et al. 2003, 367),[12] this chapter assesses how a country's political system influences its human development (Haq 1995, 67–75; Hoff and Stiglitz 2001, 427–428; Heller 1999). If democracy is strong, then poor people may exercise their voice to claim basic capacities for health care, education, and economic support, which are components of the HDI. If democracy is weak, then elites may transfer resources from the poor to the rich, or to the military, thus limiting funds for health and education and limiting the real GDP per capital (Bollen and Jackman 1985, 438–439). Moreover, dictatorships compared with democracies have higher levels of infant mortality (Zweifel et al. 2001) and female mortality (Prezeworski et al. 2000, 256–257), thereby reducing scores on the longevity component of the index.

Thus, as hypothesis 2b suggests, countries that exhibit higher levels of political freedom will also exhibit enhanced human development (F value = 44.7, $p < .0001$), but this relationship is not linear—countries classified as fully democratic have considerably enhanced human development ($p < .0001$, Bon.) compared with the partly free and the unfree countries, with little difference between the latter two ($p = 0.71$). The unconditional least-squares means are respectively 48, 103.6, and 106.5, which suggest this dichotomy: democracy (1 = fully democratic political system) versus not fully democratic (0). Both components of the Freedom House measure are associated with human development. Regarding political rights and HDI rank, the crucial distinction is between countries with a score of 1 (most free) versus all other countries with scores 2 through 7 ($p < .0001$, Bon.). Regarding civil rights, the crucial distinctions are between countries with scores of 1 or 2 versus those with scores 3 through 7 ($p < .0001$, Bon.).[13]

Weak economic facilities: national debt

Circa 1999, the World Bank and the International Monetary Fund identified highly indebted poor countries (HIPCs) that could participate in their debt relief program. This study contrasts the less debt-encumbered countries with these 38 HIPCs: Angola, Benin, Bolivia, Burkina Faso, Cameroon, Central African Republic, Chad, Congo (Republic), Côte d´ Ivoire, Equatorial Guinea, Ethiopia, Ghana, Guinea-Bissau, Guyana, Honduras, Kenya, Laos, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Nicaragua, Niger, Rwanda, São Tomé and Príncipe, Senegal, Sierra Leone, Somalia, Sudan, Tanzania, Togo, Uganda, Vietnam, Yemen, and Zambia.[14]

HIPCs are weak states: their world-system positions place them in the periphery or semiperiphery and not in the core—York, Rosa, and Dietz (2003, 289–290) report a correlation of 0.53 ($N = 110$, $p < .001$) between their measures of total national debt and weak world-system position. HIPCs are more likely to exhibit higher scores for perceived corruption and graft, slavery, conflict, and infant mortality; and lower scores for governmental effectiveness, political rights, civil liberties, and GDP; these bivariate r_s are statistically significant ($p < .0001$). Foreign investment concentration is uncorrelated with this measure of national debt ($r_s = 0.11$, $p = 0.30$). HIPCs may have lower human development because funds, which could be used to improve health, education, and economic facilities, are used instead to pay off the national debt and interest. Thus, consistent with hypothesis 2c, low national debt significantly improves the HDI rank by -65.3 ($p < .0001$, Bon.) from the mean of 135.9; Eastern Africa ($p < .0001$), Middle Africa ($p < .003$), and Western Africa ($p < .0001$) have higher debt than other regions.

Poor protective security: internal conflict and social unrest

Internal conflict and social unrest (hereafter, conflict) imply a lack of protective security. Conflict reduces human development because it may lead to deaths, which reduce longevity; to disruptions of schooling, which reduce literacy and learning; and to disruptions of work, which constrain economic development. Bales assessed a country's level of conflict in years just prior to 1999 as no serious internal conflict and unrest = 0, low = 1, and high = 2. Countries with higher scores are more likely to exhibit abuse of power by police, unlawful killings, and human rights abuses, and less likely on UN measures to exhibit political stability, lack of violence, democracy, law and order, and the rule of law (UNDP 2002, Table A1.1, 38–41); all of these bivariate r_s are statistically significant ($p < .0001$). Consistent with hypothesis 2d, at high levels of conflict, the mean HDI rank is 111.8; low conflict improves it by -10.53; and no serious conflict improves it by -64.7.[15] The difference in the mean rank between countries with no conflict and countries with either of the other two categories is statistically significant ($p < .0001$, Bon.), suggesting this dichotomy: no internal chaos (1) versus some conflict and unrest (0).

Weak transparency guarantees: corruption

Corruption is the use of public and organizational resources for private gain through bribery, favoritism, and fraud (Myrdal 1971, 200–210); it implies a lack of transparency guarantees. As did Lipset and Lenz (2000, 13) and Welzel et al. (2003, 357), this study uses the corruption perception index (CPI) of Transparency

International, which synthesizes numerous surveys of expert and public views of the extent of corruption in the various countries of the world. The scores for perceived corruption range from 1 (Most Corrupt, assigned to Iraq and Myanmar) to 10 (Least Corrupt, assigned to Denmark).[16] Countries with scores indicative of high corruption also have high scores for graft, abuse of power by police, unlawful killings, and human rights abuse; and low scores for government effectiveness, the rule of law, and law and order—these r_s are statistically significant ($p < .0001$).

Corruption adversely affects human development via its negative effects on economic and social development and on effective democracy (UNDP 2002, 63–67; Thomas 2001, 164; Welzel et al. 2003, 357). Consistent with hypothesis 2e, each unit decrease in corruption improves the HDI rank of a country by -17.45 ($p < .0001$) from the intercept value of 152.3. The explanatory analyses nest the regions by this ordinal attribute: low = 1 to high = 4; the significant differences are between the low corruption category and the other three ($p_{12} = 0.02$; $p_{13} = 0.01$; $p_{14} = 0.08$, Bon.). Thus, integrity (1 = low corruption) versus the three other levels (0) is an appropriate dichotomy.

Statistical Methods

The analyses apply multilevel modeling to quantify the effects of the cultural zones and instrumental factors on substantive freedoms (Littell et al. 1996; 2006, 55–91). Because the annual reports of the UNDP emphasize the rank-order of countries, this chapter first develops descriptive and explanatory models that probe the determinants of HDI rank. Assessing the robustness of these findings, it then replicates these analyses using the underlying scores for the index and for its components. In these replications, it quantifies the effect size of a fixed covariate by dividing its regression coefficient by the standard deviation of the response variable.

Weights, Bonferroni Adjustments, BIC, and R^2

The SAS runs adjust for differences in population size by weighting each country by the square root of its population (SAS 1990, 927). This mild transformation reduces the spread (Tukey 1977, 543) and is a compromise between weighting the data by population sizes and not weighting the data (Firebaugh 2003, 126–127). SAS then minimizes the weighted residual sum of squares $\sum_i w_i (y_i - \hat{y}_i)^2$, where w_i is the square root of the population of country i, y_i is the observed value of the response variable, and $\hat{y}_i$ is the predicted value. Thus, the squared difference between actual and predicted HDI level for a gigantic country like China is weighted more heavily than that for a miniscule country like Haiti, rather than allowing these countries to have equal weights in the calculation of the residual sum of squares. The weight has no effect on the degrees of freedom or on the number of observations, but it does influence the calculation of means and the results of multiple comparison tests.

The effects on the response variable of some of the categories of the covariates were not specified in advance of the analysis. Consequently, to tighten the tests of significance, the statistical analyses apply Bonferroni adjustments (Bon.) for exploratory post hoc multiple comparisons. For pair-wise comparisons among a set of means, these adjustments make it more difficult to reject the null hypothesis of no difference; that is, to find an effect. Such adjustments reduce the likelihood of erroneous rejections of the null hypothesis (Type-1 errors or false positives) but increase the likelihood of erroneous acceptances of the null hypothesis (Type-2 errors or false negatives).

For selecting which of a number of related models provides the most parsimonious fit to the data, Schwarz's Bayesian Information Criterion (BIC) statistic (sometimes referred to as SBC) provides a useful criterion. Holding constant the quantity −2log likelihood, the model that uses fewer parameters to provide a close fit is preferred to models that require more parameters to provide a close fit. When the smaller value of the BIC indicates the better fitting model (i.e., "Smaller is Better") as in these models, the formula for the BIC is $-2l + d(\log n + 1)$, where l denotes the maximum value of the (possibly restricted residual) log likelihood, d denotes the dimension of the model (the number of estimated covariance parameters and the rank of the design matrix of fixed effects), and n denotes the number of observations (SAS Institute Inc. 1997, 587; Schwarz 1978).

For quantifying explained variance, R^2 analogs are defined at each level as the difference between the variance components for the baseline (i.e., intercepts only) model and the variance component for the current model divided by the variance component for the baseline model (Kreft and DeLeeuw 1998, 116–119).

For testing the significance of the between-region variance, the log likelihood of the model that includes $\hat{\sigma}_r^2$ can be compared with the log likelihood of the almost identical model that does not include $\hat{\sigma}_r^2$. If the χ^2 test with 1 degree of freedom rejects the null hypothesis of no difference at $\alpha \leq 0.05$, then $\hat{\sigma}_r^2$ is statistically significant. If p is noticeably larger than 0.05, say $p > 0.10$, then $\hat{\sigma}_r^2$ is not statistically significant.

The Variance Components Model

In the following weighted unconditional model (i.e., no control variables), which groups 161 countries in the 18 regions, there is unexplained region-to-region variance in the HDI ranking ($\hat{\sigma}_r^2 = 1{,}826.4$, $z = 2.8$, $p = 0.003$) and unexplained country-to-country variance within a region ($\hat{\sigma}_c^2 = 2{,}132.5, z = 8.46, p < .0001$). Following Littell et al. ([1996] 2006, 59–61), the statistical model that produced these estimates is as follows: Let y_{ij} denote the rank (or score) of the j^{th} country of the i^{th} region. Then,

$$y_{ij} = \ \mu_r + a_i + e_{ij}, \quad i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, c_i \tag{1}$$

where

$$a_i \sim iid\ N(0, \sigma_r^2)$$
$$e_{ij} \sim iid\ N(0, \sigma_c^2).$$

In words, Eq. (1) states that the HDI rank of country j in region i is equal to the overall mean rank on the HDI, which is μ_r (this is the fixed part of the model) plus two random effects: a_i and e_{ij}. There are $r = 18$ regions and c_i countries in the i^{th} region. The random effect a_i is assumed to be composed of independent, identically distributed normal errors that have a mean of zero and constant variance $\hat{\sigma}_r^2$. This variance assesses the region-to-region variability; here, it is the variance of the true HDI means of the regions around the grand mean μ_r. The random effect e_{ij} is assumed to be composed of independent, identically distributed normal errors that have a mean of zero and constant variance σ_c^2, it is the country-to-country variance within a region. A similar model guides the analysis of HDI scores.

For the rank data, the Shapiro–Wilk test does not reject the null hypothesis that the residuals are normally distributed ($p <$ W $= 0.80$, which is much greater than $p = 0.05$, the critical probability); the stem and leaf plot and normal probability plot also suggest this. The intraclass correlation ρ (rho) (Bryk and Raudenbush 1992, 63), which represents here the proportion of the variance in the response variable that is between regions, is $\hat{\rho} = \hat{\sigma}_r^2 / (\hat{\sigma}_r^2 + \hat{\sigma}_c^2) = 1{,}826.4/(1{,}826.5 + 2{,}132.5) = 0.46$; it is substantial.

The estimate of the overall mean HDI rank (i.e., the intercept) is 75.9. Its reliability is defined as $\hat{\lambda}_{.} = \text{Reliability}\,(\bar{Y}) = \hat{\sigma}_r^2 / [\hat{\sigma}_r^2 + (\hat{\sigma}_c^2/n)] = \ 1{,}826.4/$ $[1{,}826.5 + (2{,}132.5/161)] = 0.99$. It is very high, as is the average of the regional reliabilities, $\bar{\hat{\lambda}} = 0.83$. Regions that have significantly poorer levels (i.e., higher rank scores) include Eastern Africa ($p < .0001$), Middle Africa ($p < .0001$), Northern Africa ($p < .025$), Western Africa ($p < .0001$), and South-Central Asia ($p < .002$). Regions that have significantly enhanced levels (i.e., lower rank scores) include North America ($p < .0001$), Northern Europe ($p < .0001$), Southern Europe ($p < .0012$), and Western Europe ($p < .0001$). The analyses aim to account for this variation by developing conditional hierarchical models that have additional fixed covariates and that classify the random effects of region by the categories of the typologies.[17]

Descriptive and Explanatory Models

In the descriptive models, the fixed effects of a country's covariates predict its position on the HDI and account for some of the region-to-region variance. Because such models as these relate various properties of countries to another property, its HDI position, they are associational (i.e., descriptive) and not causal (Holland 1986). The explanatory models can be thought of as potentially causal because there is some "doing"—the regions are classified by a property of its constituent elements (Lazarsfeld and Menzel [1961] 1972, 233), and the classification may cause the variance between regions to disappear. For example, a country has the property "regime type." It may have the value "democracy" or "not a democracy" on this property. In the descriptive models, this property of a country is merely one predictive covariate of its position on the HDI along with the other covariates. However, in a model that, for example, tests the explanatory power of democracy, the regions are classified (i.e., nested) by the regime typology while holding constant its fixed effect and those of the other covariates; in SAS, the nested variable, here the regions, cannot be a covariate. This nesting of regions creates region × typology subgroups that may be more homogeneous than those of the un-nested regions. If the variance between the regions across these subgroups symbolized as $\hat{\sigma}^2_{r(n)}$ (n indicates the nesting typology) becomes insignificant, then that nesting variable may cause the regional differences. Conceptually, the estimate of the causal effect $\hat{\delta} = \hat{\sigma}^2_r - \hat{\sigma}^2_{r(n)}$. Given that $\hat{\sigma}^2_r$ is statistically significant, the significance of the effect of nesting is tested by the likelihood-ratio test that compares the significance of the covariance parameter of the nested model $\hat{\sigma}^2_{r(n)}$ to that of the null model in which it is absent. If $\hat{\sigma}^2_{r(n)}$ is not statistically significant (i.e., zero), then the causal effect of the nesting equals the amount of $\hat{\sigma}^2_r$.

Descriptive models

To quantify the variance components and the fixed effects of the covariates, descriptive multilevel regression models with random effects are estimated first: Let y_{ij} denote the position on the HDI of the j^{th} country of the i^{th} region; μ_r denote the intercept level of the HDI; $X_{\mathrm{k}ij}$ denote a covariate k; and β_{ij} denotes its regression coefficient. Then,

$$y_{ij} = \mu_r + \Sigma\beta_{\mathrm{ij}}X_{\mathrm{k}ij} + a_i + e_{ij},\ i = 1, 2, \ldots, \mathrm{r},\ j = 1, 2, \ldots, c_i,\ k = 1, 2, \ldots, n \quad (2)$$

where

$$\begin{aligned} a_i &\sim iid N\left(0,\ \sigma^2_r\right) \\ e_{ij} &\sim iid N\left(0,\ \sigma^2_c\right). \end{aligned}$$

After Proc Mixed calculates the estimates of the variance parameters of this system, the region-to-region variability in HDI quantified by $\hat{\sigma}^2_r$ is of special interest.[18] If it is still statistically significant after all of the covariates have had their effects, then the reasons why the different regions have different levels of human development are not fully known; other models, perhaps those that classify the regions, are needed.

Explanatory models

Conceptually, the explanatory models may uncover candidate causal relationships as follows: Regions have random effects on a country's position on the HDI as quantified by the region-to-region variance and the variance among countries within the regions.[19] Holding constant the other covariates, a typology that encompasses an instrumental factor (designated IFTYPE) is used to classify the regions' random effects—this "doing" of the classification is consistent with Pearl's *do*(x) operator (2000, 85). The IFTYPE is more diffuse than the regional variable—countries in different regions may have the same value on IFTYPE and countries in the same region may have different values. The typologies that IFTYPE encompasses can be the eight civilization zones, the four (or two) values of slavery, the three (or two) values of political freedom, the three (or two) values of conflict, the two values of national debt, or the four (or two) values of corruption.

Here the IFTYPE can be conceptualized as either prior to both region and a country's position on the HDI or as intervening between region and position on the HDI. Regarding the first conceptualization, because the IFTYPE classification nests the region within its categories, it can be thought to be at a higher structural level than region and position on the HDI and thus prior to both of these variables; it may *explain* their covariance parameter $\hat{\sigma}_r^2$ (Lazarsfeld 1955b, *xi;* Lieberson 1997, 375–378):

$$\mathrm{R} \leftarrow \mathrm{IFTYPE} \rightarrow \mathrm{HDI}.$$

Regarding the second conceptualization, because regions of the world can be viewed as ultimate exogenous variables, the IFTYPE classification is then an intervening variable between region and a country's position on the HDI; it may interpret their covariance parameter $\hat{\sigma}_r^2$:

$$\mathrm{R} \rightarrow \mathrm{IFTYPE} \rightarrow \mathrm{HDI}.$$

In either case, Proc Mixed can create region × IFTYPE subgroups and calculate the estimate of the regional random effect for each subgroup, testing its significance (Littell et al. [1996] 2006, 75–76). If this nesting of the regions in the categories of IFTYPE eliminates the significance of the regional random effects, then that classification is a candidate causal factor since it accounts for (i.e., eliminates) the relationship between regions and a county's position on the HDI. Because of the ambiguity about which ordering of these variables is best, this chapter uses "accounts for" rather than "explains" or "interprets" when a test factor in IFTYPE reduces to insignificance a level-2 variance. But it refers to models that nest the random effects of region by IFTYPE as "explanatory models."

The explanatory models aim to determine if the between-region variance in a county's position on the HDI can be attributed to the different values of the IFTYPE. IFTYPE is considered to be a fixed effect as are the other covariates; the regions nested within an IFTYPE are considered as random effects and, given their nesting, cannot appear as a covariate. Region has a covariance parameter (σ_r^2)

with the HDI. If the control [i.e., nesting = $do(x)$] eliminates the significance of this parameter, then the initial relationship is either spurious or interpreted and the nesting typology accounts for the original relationship. Let y_{ijk} denote the value of the HDI of the k^{th} country of the j^{th} region of the i^{th} category of IFTYPE; $X_{\text{m}ijk}$ denotes a covariate m (m not equal to IFTYPE), and β_{ijk} denotes its regression coefficient. Then,

$$\begin{gathered} y_{ijk} = \mu + \beta_{ijk}\text{IFTYPE}_i + \Sigma\ \beta_{ijk}X_{\text{m}ijk} + a_{j(i)} + e_{ijk} \\ i = 1, 2, \ldots, \text{t}, \quad j = 1, 2, \ldots, r, \quad k = 1, 2, \ldots, n_{ij} \end{gathered} \tag{3}$$

where

$$\begin{aligned} a_{j(i)} &\sim iidN\left(0,\ \sigma^2_{r(IFTYPE)}\right) \\ e_{ijk} &\sim iidN\left(0, \sigma^2_c\right). \end{aligned}$$

In the subsequent explanatory models an estimate of $\sigma^2_{r(IFTYPE)}$will be depicted as $\hat{\sigma}^2_{r(n)}$to indicate the nesting of the random regions by a typological test factor.

Results

All of the subsequent models are based on the same 138 countries, use the square root of a country's population as a weight, and relate various fixed properties of a country to the response variables.[20] The preferred models are selected by considering the statistical significances of the variance components and the fixed effects, the sizes of the intraclass correlation coefficient $\hat{\rho}$(rho), and the sizes and changes in the BIC and the R^2 analogs.

Descriptive Models of HDI Rank

Table 7.2 presents the random and fixed effects for eight descriptive models.[21] Model 1 includes only the intercept, which here is the grand mean HDI rank; it equals 75.6. The intraclass $\hat{\rho} = 0.46$, the BIC = 1,349, and the values of the R^2 analogs (0 and 0) provide baselines for comparing the models—models that produce smaller values of $\hat{\rho}$ and BIC and larger values of R^2 usually are preferred to other models. Model 2, which includes the main effects of all of the covariates, is the best of these eight descriptive models. It has the second smallest $\hat{\rho}$, smallest BIC, largest level-2 R^2, largest reduction and percent reduction in BIC, and the smallest variance that is between regions, $\hat{\sigma}^2_r = 120$ (z = 1.2, p = 0.107, one-tailed test).

Table 7.2 Descriptive models of human development rank, random and fixed effects

Models	Model 1, Intercept	Model 2, Full descriptive model	Model 3, without civilizations	Model 4, without slavery	Model 5, without freedom	Model 6, without HIPC	Model 7, without conflict	Model 8, without corruption
Variance components								
σ_r^2	1,878.4	120.0	468.5	159.4	104.3	213.2	133.4	190.5
z	2.8	1.2	2.5	1.4	1.2	1.5	1.3	1.4
p	0.003	0.107	0.007	0.081	0.122	0.062	0.104	0.079
σ_c^2	2,214.7	1,119.3	1,161.3	1,162.5	1,168.2	1,229.9	1,269.8	1,164.8
z	7.8	6.7	7.4	6.9	6.8	6.9	6.8	6.7
p	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Intraclass ρ	0.46	0.10	0.29	0.12	0.08	0.15	0.10	0.14
Fit statistics								
Level-2 R^2 Analog	0.00	0.94	0.75	0.92	0.94	0.89	0.93	0.90
Level-1 R^2 Analog	0.00	0.49	0.48	0.48	0.47	0.44	0.43	0.47
BIC	1,349	1,119	1,198	1,144	1,134	1,141	1,146	1,132
Change in BIC	0	−229.9	−150.3	−204.6	−214.6	−207.1	−203.0	−216.8
% Change in BIC	0	−17.0%	−11.1%	−15.2%	−15.9%	−15.4%	−15.1%	−16.1%
Intercept	75.6	127.1	148.8	122.9	119.9	113.7	129.9	110.1
t	7.2	12.0	15.7	11.3	11.8	10.3	11.6	11.3
p	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Type 3 tests of fixed effects								
Civilization zone								
F value	–	5.5	–	6.0	7.1	4.5	4.7	6.2
Pr > *F*	–	< .0001	–	< .0001	< .0001	< .0001	< .0001	< .0001
New slavery								
F value	–	3.5	6.2	–	3.8	3.6	2.0[b]	4.9
Pr > *F*	–	0.0176	< .0006	–	0.0121	0.0153	0.1242	0.0033
Political freedom								
F value	–	2.8[a]	6.2	3.3	–	2.1[a]	3.9	3.4
Pr > *F*	–	0.0689	0.0028	0.0423	–	0.1272	0.0234	0.0369
Not a HIPC								
F value	–	19.4	16.2	19.9	18.7	–	15.8	21.9
Pr > *F*	–	< .0001	0.0001	< .0001	< .0001	–	0.0001	< .0001
Conflict								
F value	–	9.1	6.5	6.8	10.5	7.5	–	17.2
Pr > *F*	–	0.0002	0.0022	0.0017	< .0001	0.0009	–	< .0001
Corruption								
F value	–	11.1	18.4	15.5	12.6	13.5	26.9	–
Pr > *F*	–	0.0012	< .0001	0.0001	0.0006	0.0004	< .0001	–

[a] When political freedom is dichotomized as fully democratic versus the other two categories its effect is statistically significant: Model 2, $F = 5.4, p = .02$; Model 6, $F = 4, p = .049$.

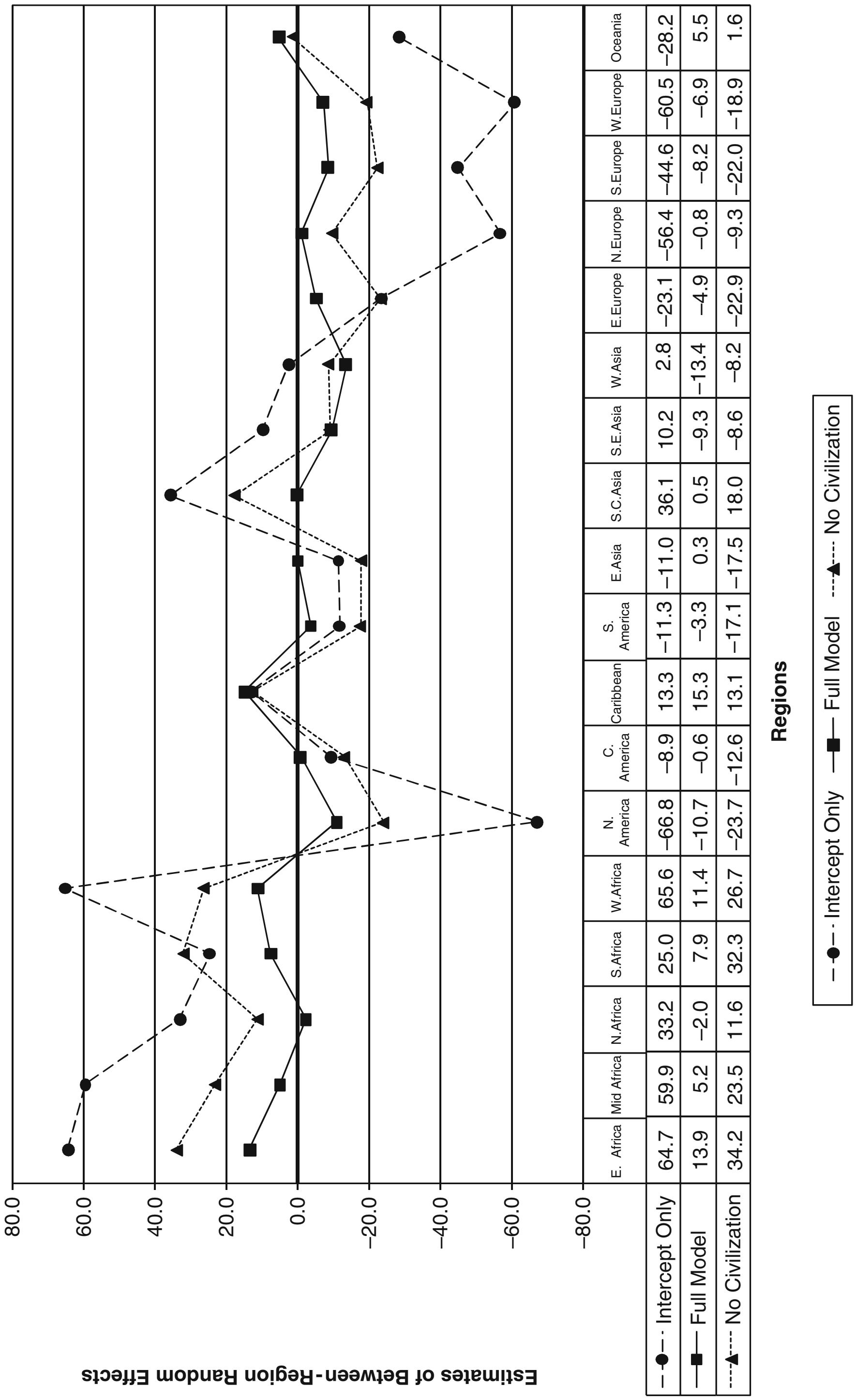
[b] When slavery is dichotomized as None (1) or Very Little (2) versus High (3) or Very High (4) then the effect is statistically significant: Model 7, $F = 5.7, p = 0.02$.

However, the likelihood-ratio test, which compares the difference in log likelihood between the model that specifies region as random and the independent-errors model that lacks σ_r^2, rejects the null hypothesis of no difference—the variance between regions, although small, still is important ($df = 1$, $\chi^2 = 5.5$, $p = 0.019$): Eastern Africa's random effect estimate of 13.9 ($p = 0.038$) and Western Asia's random effect estimate of -13.4 ($p = 0.027$) are statistically significant; the other 16 estimates are not significant. This model produces the largest reduction in the variance that is between countries within regions, from $\hat{\sigma}_c^2 = 2{,}214.7$ to $\hat{\sigma}_c^2 = 1{,}119.3$, a percentage reduction of 49.5. The Type 3 test of the statistical significance of the independent contributions of each fixed covariate indicates that all but the trichotomous indicator of political freedom significantly influences rank on the HDI. The effect of political freedom becomes statistically significant ($F = 5.38$, $p = 0.022$) when the dichotomy fully democratic (1) versus not (0) is used, and the model is reestimated.

By deleting in succession each covariate and reestimating the system without the effect of the deleted variable, Models 3 through 8 ascertain which of the covariates in the full model most pivotally accounts for a country's HDI rank. The deletion of an important predictor will noticeably increase the BIC from that of the full model; produce smaller decreases and percent change in BIC from the benchmark values of Model 1; reduce the estimates of the R^2 analogs; increase the significance of $\hat{\sigma}_r^2$; shift the lower bound of the confidence interval away from zero; increase the intraclass correlation $\hat{\rho}$; and increase the volatility of the random effects. Using these criteria the statistics for Model 3 indicate that civilization zone is more important than the other variables in predicting the HDI rank for this system.

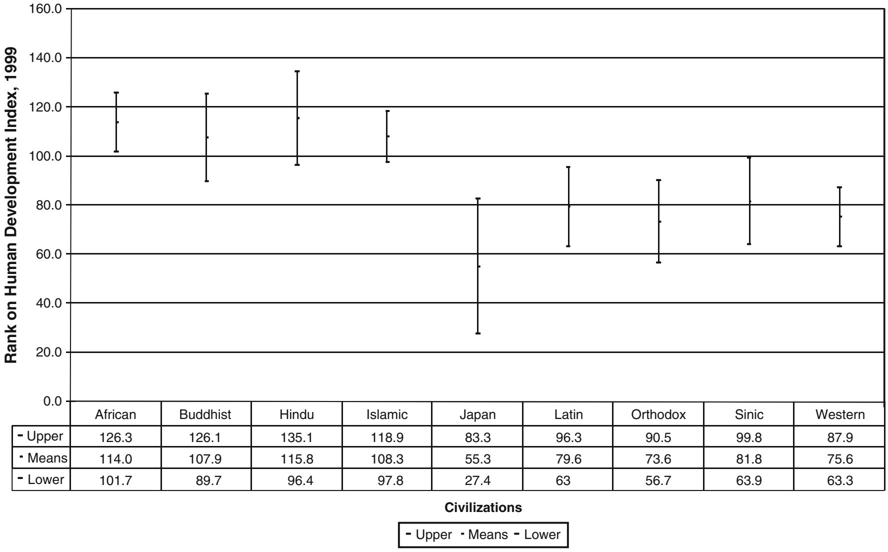
Illustrating the flattening of the random effects due to the covariates, Fig. 7.2 plots the random effects by region for three models: the baseline unconditional model with no fixed covariates (Model 1), the full conditional model (Model 2), and the conditional model that lacks the control for civilization (Model 3). When there are no covariates, the random effects vary from a positive peak value of 65.6 to a nadir of -66.6 HDI rank units. When the model includes all of the covariates, the average of the random effects flatten hugging the zero axis, the positive peak is only 15.3 HDI units; the nadir, -13.4. When civilization is dropped from the model, the variance increases; the positive peak is 34.2; the nadir, -23.7 HDI units.

Depicting the differences among civilization zones in their HDI rank, Fig. 7.3 presents their least-squares means from Model 2. There are three clusters: consistent with hypothesis 1, Japan has the most favorable rank; the African, Buddhist, Hindu, and Islamic have unfavorable scores; and the Latin, Orthodox, Sinic, and Western have middling scores. The Bonferroni tests indicate the following differences and similarities. The African category has impoverished (that is, elevated) rank scores compared with Japan (+58.6, $p = 0.007$), Latin (+34.4, $p = 0.04$), Orthodox (+40.4, $p = 0.004$), and the West (+38.4, $p = 0.001$). The Buddhist category has impoverished rank scores compared with Japan (+52.5, $p = 0.028$) and the West (+32.3, $p = 0.065$). The Hindu category has impoverished scores compared with Japan (+60.4, $p = 0.018$), Orthodox (+42.2, $p = 0.032$), and the West (+40.2, $p = 0.006$). Similarly, the Islamic category has impoverished scores



	E. Africa	Mid Africa	N.Africa	S.Africa	W.Africa	N. America	C. America	Caribbean	S. America	E.Asia	S.C.Asia	S.E.Asia	W.Asia	E.Europe	N.Europe	S.Europe	W.Europe	Oceania
● Intercept Only	64.7	59.9	33.2	25.0	65.6	−66.8	−8.9	13.3	−11.3	−11.0	36.1	10.2	2.8	−23.1	−56.4	−44.6	−60.5	−28.2
■ Full Model	13.9	5.2	−2.0	7.9	11.4	−10.7	−0.6	15.3	−3.3	0.3	0.5	−9.3	−13.4	−4.9	−0.8	−8.2	−6.9	5.5
▲ No Civilization	34.2	23.5	11.6	32.3	26.7	−23.7	−12.6	13.1	−17.1	−17.5	18.0	−8.6	−8.2	−22.9	−9.3	−22.0	−18.9	1.6

Fig. 7.2 The controls for the covariates flatten the distributions of random effects



	African	Buddhist	Hindu	Islamic	Japan	Latin	Orthodox	Sinic	Western
- Upper	126.3	126.1	135.1	118.9	83.3	96.3	90.5	99.8	87.9
· Means	114.0	107.9	115.8	108.3	55.3	79.6	73.6	81.8	75.6
- Lower	101.7	89.7	96.4	97.8	27.4	63	56.7	63.9	63.3

Fig. 7.3 Least-squares means of human development rank scores by civilization zones

compared with Japan (+52.98, $p = 0.017$), Orthodox (+34.7, $p = 0.008$), and the West (+32.7, $p = 0.002$). Japan has more favorable scores than Latin, Orthodox, Sinic, and Western categories, but these differences are not statistically significant ($p = 1.0$). The Latin scores are similar to those of the Orthodox, Sinic, and Western categories ($p = 1.0$). The Orthodox scores are similar to those of the Sinic and Western categories ($p = 1.0$). Finally, the Sinic and Western categories have similar scores ($p = 1.0$). Apparently, civilization zones matter when predicting the HDI rank, but other factors account for the regional differences.

Explanatory Models of HDI Rank

These analyses aim to uncover which of the covariates, when used as a typology that nests the random regions—thereby aiming to create homogeneous subgroups of countries—will, when compared with the full predictive model, further flatten the estimated regional random effects and the estimated regional means. This further flattening is indicated primarily by smaller and less statistically significant values of $\hat{\sigma}^2_{r(n)}$ compared with $\hat{\sigma}^2_r$. Other diagnostic indicators are a likelihood-ratio test that indicates no statistical difference between the model and an analogous independent-errors model that lacks the regional covariance parameter, smaller values of $\hat{\rho}$ and of BIC, larger reductions and percent reductions in the BICs, and favorable R^2s. These analyses also aim to uncover which variables reduce $\hat{\sigma}^2_c$, the variance between countries within regions.

Table 7.3 Explanatory models of human development rank, random and fixed effects

Models	Model 1, Intercept only	Model 2, Full descriptive model	Model 3, Region classified by civilizations	Model 4, Region classified by slavery	Model 5, Region classified by freedom index	Model 6, Region classified by full democracy	Model 7, Region classified by not a HIPC	Model 8, Region classified by conflict	Model 9, Region classified by corruption
Variance components									
σ_r^2 and $\sigma_{r(n)}^2$	1,878.4	120.0	98.3	126.5	123.5	65.3	120.3	196.1	220.1
z	2.8	1.2	1.5	1.0	1.8	0.9	1.4	1.7	1.9
p	0.003	0.107	0.066	0.171	0.036	0.176	0.088	0.044	0.063
Likelihood ratio χ^2	146	5.52	2.69	1.45	7.21	2.60	5.45	5.26	11.15
p	$< .0001$	0.019	0.017	0.228	0.007	0.107	0.020	0.022	0.0008
σ_c^2	2,214.7	1,119.3	1,116.0	1,062.7	1,028.0	1,184.4	1,095.7	940.3	855.3
z	7.8	6.7	6.8	4.4	6.4	6.3	6.5	5.5	5.3
p	$< .0001$	<0.0001	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Intraclass ρ	0.46	0.10	0.08	0.11	0.11	0.05	0.10	0.17	0.20
Fit statistics									
Level-2 R^2 Analog	0.00	0.94	0.95	0.93	0.93	0.97	0.94	0.90	0.88
Level-1 R^2 Analog	0.00	0.49	0.50	0.52	0.54	0.47	0.51	0.58	0.61
BIC	1,349	1,119	1,120	1,125	1,119	1,123	1,141	1,121	1,101
Change in BIC	0	−229.90	−228.60	−224.00	−230.00	−225.90	−207.10	−227.90	−248.00
% Change in BIC	0	−17.0%	−17.0%	−16.6%	−17.1%	−16.8%	−15.4%	−16.9%	−18.4%
Intercept	75.6	127.1	121.7	128.5	132.1	127.2	130.5	122.5	114.5
t	7.2	12.0	11.2	11.5	11.5	11.8	11.8	10.9	10.6
p	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$

(continued)

Table 7.3 (continued)

Models	Model 1, Intercept only	Model 2, Full descriptive model	Model 3, Region classified by civilizations	Model 4, Region classified by slavery	Model 5, Region classified by freedom index	Model 6, Region classified by full democracy	Model 7, Region classified by not a HIPC	Model 8, Region classified by conflict	Model 9, Region classified by corruption
Type 3 tests of significance:									
Civilization zone									
F value	–	5.5	7.5	6.7	7.7	8.2	6.2	7.1	5.5
$Pr > F$	–	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
New slavery									
F value	–	3.5	3.9	2.0[a]	5.7	3.9	4.1	5.7	2.8
$Pr > F$	–	0.0176	0.0119	0.1234	0.0013	0.0109	0.0084	0.0014	0.0472
Political freedom									
F value	–	2.8[b]	3.1	2.2[b]	1.8[b]	2.1[b]	2.8[b]	2.3[b]	3.6
$Pr > F$	–	0.0689	0.0487	0.118	0.1878	0.1273	0.0682	0.1115	0.0325
Not a HIPC									
F value	–	19.4	17.3	21.8	15.6	19.8	12.5	20.2	28.4
$Pr > F$	–	$< .0001$	0.0001	$< .0001$	0.0002	$< .0001$	0.0018	$< .0001$	$< .0001$
Conflict									
F value	–	9.1	8.5	8.8	9.8	9.2	8.2	6.0	10.3
$Pr > F$	–	0.0002	0.0004	0.0004	0.0002	0.0002	0.0005	0.0054	.0001
Corruption									
F value	–	11.1	11.1	11.6	12.7	11.5	11.6	6.0	4.2
$Pr > F$	–	0.0012	0.0013	0.0010	0.0006	0.001	0.0010	0.0168	0.0111

[a] When slavery is dichotomized as None (1) or Very Little (2) versus High (3) or Very High (4) then the effect is statistically significant, $F = 5.7$, $p = 0.02$.

[b] When political freedom is dichotomized as fully democratic versus the other two categories its effects are statistically significant as follows: Model 2, $F = 5.4$, $p = 0.02$; Model 4, $F = 4.4$, $p = 0.04$; Model 5, $F = 4.2$, $p = 0.05$; Model 6, $F = 4.2$, $p = 0.0497$; Model 7, $F = 5.5$, $p = 0.02$; and Model 8, $F = 4.5$; $p = 0.04$.

Table 7.3 presents the seven alternative explanatory models (Models 3 through 9) that classify the random effect of region; to facilitate comparisons, it also reports again Model 1, which includes only the intercept, and Model 2, the full descriptive model.

Civilization zones

These zones have important fixed effects but when they classify the regional random effects they do not explain the variance in rank that is between the regions, see Model 3. The $\hat{\sigma}^2_{r(n)} = 98.3$ is smaller than $\hat{\sigma}^2_r$ for the full descriptive model, but it does approach statistical significance, the $p = 0.066$. However, when the log likelihood of Model 3 is compared with that of the analogous independent-errors model (i.e., the model that lacks the level-2 variance component) the likelihood-ratio test rejects the null hypothesis of no difference, thereby indicating the significance of $\hat{\sigma}^2_{r(n)}(\text{df} = 1, \chi^2 = 5.67, p = 0.017)$.

Proc Mixed parameterizes the estimates of the random effects so that their mean is zero. It tests the hypothesis that there is no difference between an estimate and this mean. Using this test, different civilization zones do not explain all the variability that is between regions: Western African countries that are classified as Islamic have a significantly higher (i.e., worse) rank (+16.17, $p = 0.030$) than the mean of zero; and Western Asian countries that are classified as Islamic have a significantly lower (i.e., better) rank (-13.8, $p = 0.031$). However, this model does explain a lot: for the other 35 region × zone combinations the null hypothesis of no difference between the random-effect estimate and the mean of zero is not rejected. When compared to Model 2, the nesting of region by civilization zone does not add much explanatory power: the log likelihoods differ by only 0.2 and the BIC and changes in BIC are about the same. Slavery and fully democratic regimes account for more of the regional variance in HDI rank.[22]

Contemporary slavery

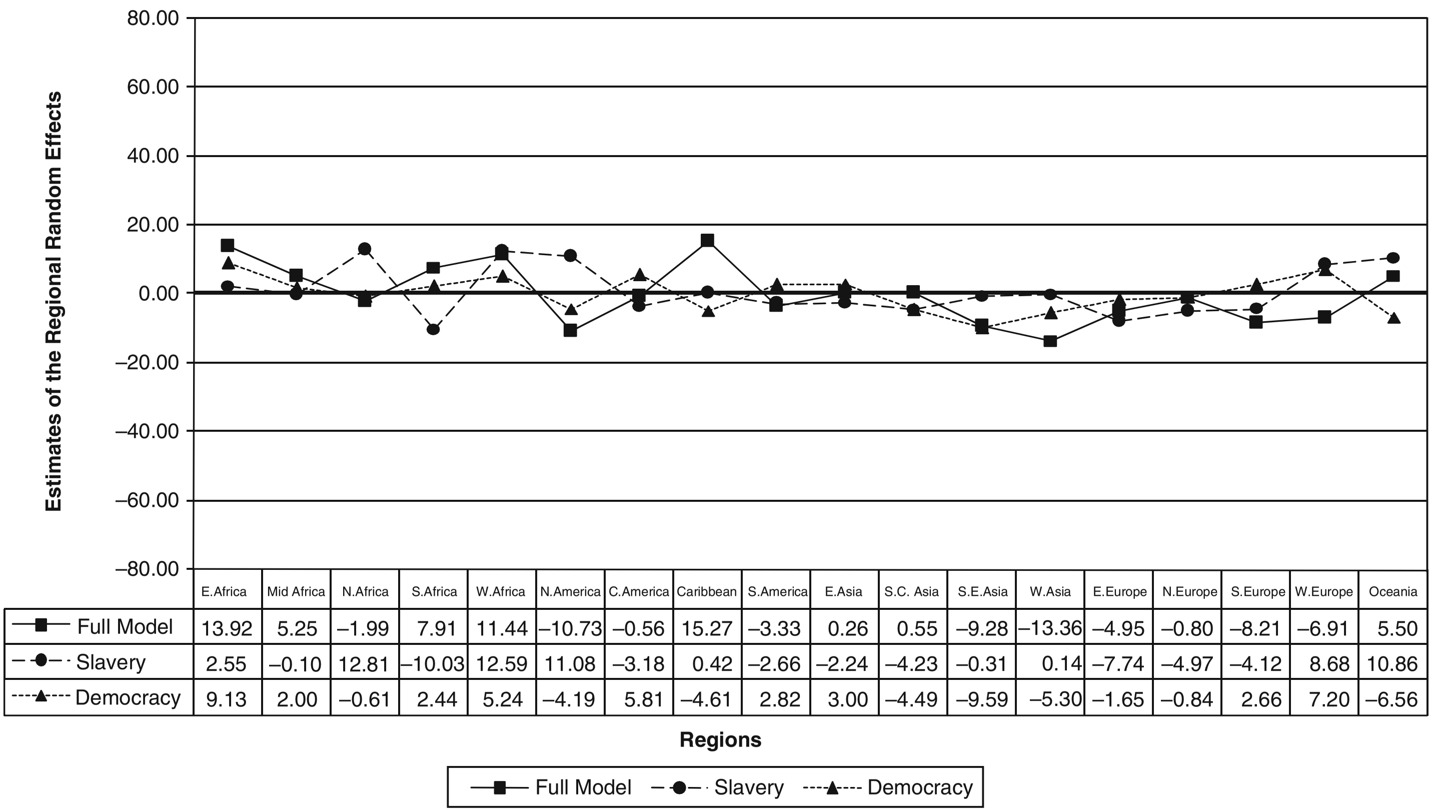
When all of the covariates are retained, and when the four categories of slavery are used to classify the regions, the resulting Model 4 is the best of the models thus far in terms of reducing the significance of the variance that is between regions. Although its value of $\hat{\sigma}^2_{r(n)} = 126.5$ is higher than the estimates of $\hat{\sigma}^2_r$ or $\hat{\sigma}^2_{r(n)}$ for other key models—the full predictive model, and the classification models for civilization zone, democracy models, and debt—the null hypothesis that implies no significant variance between the regions is not rejected, the probability = 0.171. Moreover, when the likelihood-ratio test compares the $-2 \times$ residual log likelihood of this model (1,116.9) with that (1,118.4) for the analogous independent-errors model, the difference in χ^2 of 1.5 (DF $= 1$, $p = 0.23$) indicates that the null hypothesis of no difference between these models is not rejected. In fact, in all 45 trials the null hypothesis of no difference between the random-effect estimate for a unique region × slavery combination and the mean of zero is not rejected; the differences lack statistical significance. These statistics imply that for these data slavery universally accounts for much of the regional variance in the HDI ranking. However, compared with the other models both the level-2 $R^2 = 0.93$ and the intraclass correlation $\hat{\rho} = 0.11$ are middling, which suggests that $\hat{\sigma}^2_{r(n)}$ is still rather high relative to the total of the covariance parameters. The value of $\hat{\sigma}^2_c = 1{,}062.7$ is lower than those for the other competitive models assessed thus far. The Type 3 tests of the significance of the fixed effects on the HDI indicate that the contributions of civilization zone ($p < .0001$), national debt ($p < .0001$), conflict ($p = .0004$), and corruption ($p = 0.0010$) are very statistically significant. But, to obtain significance, the measures of freedom ($p = 0.039$) and slavery ($p = 0.019$) need to be dichotomized. For the various levels of slavery the least-squares HDI means are as follows: slavery 1 = 84.3; 2 = 83; 3 = 93; and 4 = 99.6.

Democracy

When the categories of freedom are dichotomized as fully democratic (i.e., fully free = 1) versus not (partly free plus unfree = 0) and then used to nest the regions, Model 6, the resulting model, accounts for the variance that is between the regions. (Model 5 that nests the regions by the trichotomous measure of freedom does not account for this variance.) The $\hat{\sigma}^2_{r(n)} = 65.3$, the two R^2 analogs (0.97 and 0.47), and the intraclass $\hat{\rho} = 0.05$ are the most favorable of these models. The null hypothesis of no significant between-region variance is not rejected, $p = 0.174$, which is about the same as the $p = 0.176$ for slavery in Model 4. In all 30 unique region × democracy combinations, the null hypothesis of no difference between the random-effects estimates and the mean of zero is not rejected. The Type 3 tests indicate that all of the fixed effects of the covariates are statistically significant except for democracy, which must be dichotomized to attain significance ($p = 0.0497$). The likelihood-ratio test for the significance of the between-region variance parameter ($df = 1$, $\chi^2 = 2.69$, $p = 0.107$) slightly favors the slavery model ($df = 1$, $\chi^2 = 1.45$, $p = 0.228$). But the BIC statistics very slightly favor the democracy classification model (BIC = 1,123) over the slavery classification model (BIC = 1,125), as do the ranges of the random effects estimates (18.72 to 22.85). Figure 7.4 depicts the SAS estimates of these regional random effects, comparing their reduced variability with estimates for the full descriptive model.[23]

Highly indebted poor countries

Model 7 reports the results when the regions are nested by the HIPC typology. For these data the resulting explanation of the variability that is between regions ($\hat{\sigma}^2_{r(n)} = 120.3$) is not as favorable as those explanations provided by the full descriptive model and the classification models for civilization zone, slavery, and freedom. However, the least-squares means clearly indicate that countries referred



	E.Africa	Mid Africa	N.Africa	S.Africa	W.Africa	N.America	C.America	Caribbean	S.America	E.Asia	S.C. Asia	S.E.Asia	W.Asia	E.Europe	N.Europe	S.Europe	W.Europe	Oceania
■ Full Model	13.92	5.25	–1.99	7.91	11.44	–10.73	–0.56	15.27	–3.33	0.26	0.55	–9.28	–13.36	–4.95	–0.80	–8.21	–6.91	5.50
● Slavery	2.55	–0.10	12.81	–10.03	12.59	11.08	–3.18	0.42	–2.66	–2.24	–4.23	–0.31	0.14	–7.74	–4.97	–4.12	8.68	10.86
▲ Democracy	9.13	2.00	–0.61	2.44	5.24	–4.19	5.81	–4.61	2.82	3.00	–4.49	–9.59	–5.30	–1.65	–0.84	2.66	7.20	–6.56

Fig. 7.4 Compared with the full descriptive model, the nesting of the regions by slavery or by full democracy (1 or 0) further flattens the distributions of random effects estimates

to as low debt have much more favorable human development rank than highly indebted poor countries (estimate = -25.5, $p = 0.002$, Bon.). This model does provide a better explanation of the between-region variance than either the conflict or corruption classification models; in only one of the 25 debt × region combinations (Western Asia) is the null hypothesis of no difference between the random-effect estimate and the mean of zero rejected (-15, $p < 0.015$).

Conflict

In Model 8, which classifies the regions by the level of conflict, the variance that is between the regions remains high $(\hat{\sigma}^2_{r(n)} = 196.1)$ as does the intraclass correlation $\hat{\rho} = 0.17$. The variance that is among countries within regions is relatively low $(\hat{\sigma}^2_c = 940.3)$ and the level-1 R^2 also is favorable. The BIC statistic is about the same as those for Models 2 through 5. The null hypothesis of no difference between the random-effects estimate of a unique region × conflict combination and the mean of zero is rejected in three of the 41 comparisons. When the level of conflict is very low, Western Africa has a worse HDI rank (+19, $p = 0.04$) but South-Eastern Asia (-18.94, $p < 0.03$) and Western Asia (-17.75, $p = 0.04$) have better HDI positions. The least-squares means indicate that countries with the least conflict have a better rank, 77.1, compared with the other two categories, 98.4 and 100.83, respectively ($p < 0.02$, Bon.).

Corruption

When in Model 9 the regions are classified by the four levels of corruption, the variability between regions is larger $(\hat{\sigma}^2_{r(n)} = 220.1)$ and more statistically significant ($p = 0.03$) than that of any of other substantive models. In five of the 47 region × corruption combinations, the null hypothesis of no difference between the random-effects estimate and the mean of zero is not rejected, so there is variance between regions that remains to be explained. But the corruption typology offers the best explanation of the variability among countries within regions, $\hat{\sigma}^2_c = 855.3$ is the smallest value. The BIC = 1,101 suggests that this model fits better than the other models, as do the statistics for change and percentage change in BIC. The Type 3 tests of the fixed effects indicate that all of the covariates have statistically significant effects. The least-squares means indicate that countries with the lowest level of corruption have the better rank, 71.2, whereas the countries with the other levels of corruption have significantly worse rank, 97.7, 100.2, and 96.6, with no significant difference among them. When the latter three categories are grouped together, their mean rank is 98.9, which is significantly higher than 71.2 ($p = 0.002$, Bon.).

These analyses of the countries' rank on the HDI generally are consistent with the sensitizing hypotheses: Civilization zones correlate with human development—Japan and, less so, the West have the most favorable rank; all of the instrumental factors have strong fixed effects; the typologies of emancipative employment (i.e., low slavery) and democracy tend to explain the regional differences; highly indebted poor countries have very significant negative effects, and differences in corruption noticeably influence the variance among countries within the regions. Because these findings are based on the countries' rank on the HDI, they may be fragile. Consequently, to replicate key analyses, the underlying scores are modeled next.

Tests and Replications Using Index Scores

This section first relates an overall index of instrumental factors supportive of freedom to the HDI and to its underlying scores, combined and disaggregated. Then it assesses the unique effects of the binary indicators of the instrumental factors on the measures of human development. Finally, it replicates the explanatory analyses using the underlying HDI scores and finds that emancipative employment and full political democracy coupled with the lack of internal chaos account for the regional differences in human development scores.

The index of instrumental factors

The index of instrumental factors supportive of freedom is the sum of the following binary indicators: emancipative employment (1 = no or low slavery) versus slavery (0); fully democratic political system (1) versus not fully democratic (0); low national debt (1) versus HIPC (0); no internal chaos (1) versus conflict and unrest (0); and integrity (1 = low corruption) versus corruption (0). Their sum creates a reliable index ($\alpha = 0.69$ weighted, $\alpha = 0.68$ unweighted). As Sen's theory predicts, this index of the instrumental factors is strongly positively correlated with a country's substantive freedoms, assessed by its HDI rank (reversed so 1 is the lowest value), HDI score, and the scores for the component indexes, either weighted by the square root of the countries populations or not weighted, see Table 7.4.

Effects on the HDI ranks and scores, and on component scores

Table 7.5 documents that civilization zones and each of the binary indicators of the instrumental factors have associational effects on the various measures of the substantive freedoms: that is, on HDI rank, scores for the HDI, and scores for longevity, literacy, and GDP per capita. When the size of the effect of each instrumental factor is assessed by dividing its unstandardized fixed effect on the zero-to-one scale by the standard deviation of the response variable, low national debt (i.e., countries that are not HIPCs) consistently has the strongest effect on the

Table 7.4 The Pearson correlations between the index of instrumental freedoms and the indicators of substantive freedoms are positive and statistically significant

Indicators of substantive freedom	Index of instrumental factors (weighted data)	Index of instrumental factors (data not weighted)
HDI rank	0.84	0.82
	$p < .0001$	$p < .0001$
HDI score	0.78	0.80
	$p < .0001$	$p < .0001$
Longevity score	0.68	0.69
	$p < .0001$	$p < .0001$
Literacy score	0.69	0.72
	$p < .0001$	$p < .0001$
GDP per capita	0.84	0.81
	$p < .0001$	$p < .0001$

The square root of a country's population is the weight.

Table 7.5 Effects of indicators of instrumental freedoms on indicators of human development, controlling for civilization zones, proc mixed estimates

Response variable	HDI ranking	HDI scores	Longevity scores	Literacy scores	GDP per capita scores
Measures					
σ_r^2	149.34[a]	0.0031[a]	0.0097[a]	0.0038[a]	0.0008
z	1.3	1.5	2.1	1.6	0.9
p	0.21	0.15	0.03	0.12	0.38
Likelihood-ratio test p	0.017	0.001	<.0001	0.001	0.20
Bonferroni p	0.085	0.005	0.0005	0.005	1
σ_c^2	1,065.4	0.020	0.016	0.049	0.029
z	6.7	6.9	7.09	7.33	7.29
p	< .0001	< .0001	< .0001	< .0001	< .0001
BIC	1,137	−212.7	−220.0	−108.8	−179.1
Intercept β_0	74.3	0.55	0.54	0.60	0.47
t	8.8	14.8	13.5	11.1	11.9
p	< .0001	< .0001	< .0001	< .0001	< .0001
Civilization zones					
F value	5.5	3.97	2.8	4.1	3.6
$Pr > F$	< .0001	0.0004	0.008	0.0003	0.001
Bonferroni p	0.0005	0.002	0.040	0.0015	0.005
Instrumental factors					
Emancipative employment					
β	12.1	0.058	0.052	0.079	0.045
t	3.3	3.62	3.45	3.20	2.44
p	0.0015	0.0005	0.001	0.002	0.016
Bonferroni p	0.0075	0.0025	0.005	0.010	0.080
Effect size	0.12	0.15	0.14	0.18	0.11

(continued)

Table 7.5 (continued)

Response variable	HDI ranking	HDI scores	Longevity scores	Literacy scores	GDP per capita scores
Full democracy					
β	10.00	0.042	0.048	0.04	0.05
t	2.1	2.09	2.53	1.2	2.0
p	0.036	0.039	0.013	0.221	0.045
Bonferroni p	0.180	0.195	0.065	1.000	0.225
Effect size	0.10	0.11	0.12	0.09	0.12
Not a HIPC					
β	24.6	0.119	0.093	0.134	0.137
t	5.1	5.7	4.7	4.2	5.8
p	<.0001	<.0001	<.0001	<.0001	<.0001
Bonferroni p	0.0005	0.0005	0.0005	0.0005	0.0005
Effect size	0.25	0.31	0.24	0.31	0.34
No internal chaos					
β	19.7	0.055	0.043	0.048	0.076
t	5.3	3.4	2.9	1.9	4.0
p	<.0001	0.001	0.004	0.057	.0001
Bonferroni p	0.0005	0.005	0.020	0.285	0.0005
Effect size	0.20	0.15	0.11	0.11	0.19
Integrity					
β	21.7	0.064	0.043	0.046	0.124
t	3.5	2.4	1.7	1.2	4.3
p	0.001	0.019	0.101	0.248	< .0001
Bonferroni p	0.001	0.095	0.505	1.000	0.0005
Effect size	0.22	0.17	0.11	0.11	0.31
Standard deviation of response variable	97.9	0.379	0.382	0.428	0.399

[a] The likelihood-ratio test suggests that this variability between regions is statistically significant. The Bonferroni p-values calculated by Proc Multtest weaken the statistical significance of the parameters. In this table the effect size equals the unstandardized regression coefficient divided by the standard deviation of the response variable.

measures of substantive freedoms, compared with the other instrumental factors. The likelihood-ratio test of the significance of $\hat{\sigma}_r^2$ indicates that this set of covariates explains the variability between the regions only when the response variable is the country's score for GDP per capita. To explain $\hat{\sigma}_r^2$ for the total HDI scores, it is necessary to classify the random effect by typological variables, as follows.

Replication of the explanatory analysis using HDI scores

Table 7.6 replicates the earlier explanatory analysis of the HDI ranking of Table 7.3 by using a country's total HDI score as the response variable and the dichotomized measures of the instrumental factors as covariates. The earlier analysis suggested that full democracy and the absence of slavery account for the variance between

Table 7.6 Explanatory models of human development scores, random and fixed effects

Models	Model 1, Intercept only model HDI scores	Model 2, Full descriptive model HDI scores	Model 3, Region classified by civilizations and no chaos	Model 4, Region classified by slavery and no chaos	Model 5, Region classified by full democracy and no chaos	Model 6, Region classified by not HIPC and no chaos	Model 7, Region classified by no chaos	Model 8, Region classified by integrity and no chaos
Measures								
σ_r^2 and $\sigma_{r(n)}^2$	0.026	0.003	0.003	0.002	0.0007	0.003	0.003	0.003
z	2.8	1.5	1.8	1.4	0.7	1.5	1.3	1.5
p	0.005	0.15	0.08	0.17	0.51	0.14	0.19	0.15
Likelihood ratio test p	$< .0001$	0.001	0.01	0.12	0.37	0.03	0.05	0.04
σ_c^2	0.0324	0.020	0.019	0.019	0.024	0.020	0.020	0.020
z	7.8	6.9	5.9	4.7	6.1	5.9	5.9	6.1
p	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Intraclass ρ	0.45	0.13	0.13	0.11	0.03	0.11	0.12	0.12
BIC	−177.7	−212.7	−206.4	−202.3	−201.4	−205.2	−205.1	−205.3
Change in BIC	–	−35.0	−28.7	−24.6	−23.7	−27.5	−27.4	−27.60
% Change in BIC	–	19.7%	16.2%	13.8%	13.3%	15.5%	15.4%	15.5%
Intercept	0.71	0.55	0.56	0.54	0.55	0.52	0.55	0.55
t	18.0	14.8	14.7	14.2	15.0	12.8	20.7	13.9
p	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Type 3 test of significance								
Civilization zones								
F value	–	4.0	6.7	7.0	10.0	6.2	5.1	5.2
$Pr > F$	–	0.0004	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$	$< .0001$
Effects of:								
Emancipatory employment								
β	–	0.06	0.07	0.07	0.06	0.06	0.06	0.06
t	–	3.62	4.1	3.2	3.7	3.6	3.7	3.8

p	–	0.001	< .0001	0.002	0.000	0.001	0.000	0.000
Effect size	–	0.15	0.19	0.19	0.17	0.17	0.17	0.17
Democracy								
β	–	0.04	0.04	0.04	0.04	0.04	0.04	0.04
t	–	2.09	1.58	1.97	1.7	1.82	1.94	1.74
p	–	0.04	0.12	0.05	0.10	0.07	0.06	0.09
Effect size	–	0.11	0.09	0.12	0.11	0.11	0.11	0.10
Not a HIPC								
β	–	0.12	0.11	0.12	0.13	0.14	0.12	0.12
t	–	5.67	5.26	5.42	6.26	4.98	5.72	5.75
p	–	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Effect size	–	0.31	0.29	0.32	0.35	0.37	0.32	0.32
No internal chaos								
β	–	0.06	0.05	0.06	0.05	0.07	0.06	0.06
t	–	3.42	2.31	2.79	2.88	2.75	2.48	2.33
p	–	0.00	0.03	0.01	0.01	0.01	0.02	0.03
Effect size	–	0.15	0.14	0.15	0.15	0.17	0.17	0.15
Elite integrity								
$\beta\beta$	–	0.06	0.07	0.07	0.07	0.06	0.06	0.08
t	–	2.38	2.36	2.49	2.76	2.07	1.95	2.59
p	–	0.02	0.02	0.02	0.01	0.04	0.05	0.01
Effect size	–	0.17	0.18	0.18	0.19	0.15	0.15	0.22
Standard deviation of response variable	0.379	0.379	0.379	0.379	0.379	0.379	0.379	0.379

Here the effect size equals the unstandardized regression coefficient divided by the standard deviation of the response variable.

regions. However, when the total HDI score is the response variable and democracy or slavery is used singularly to nest the random regions, there is unexplained variability in $\hat{\sigma}^2_{r(n)}$. To account for this variance, a joint classification with no internal chaos (1) versus some chaos (0) is necessary. As expected, the classification by civilization zone (with or without the joint classification by absence of internal chaos) does not account for the regional random effect.

When the dichotomous typological test variable is fully democratic (1) versus not (0), the likelihood-ratio test indicates that $\hat{\sigma}^2_{r(n)}$ is statistically significant ($p = 0.0165$); in three of the 30 unique region × democracy combinations, the null hypothesis of no difference between the random-effect estimate and the mean of zero is rejected. However, when democracy and no internal chaos jointly classify the random regions, these attributes of a country strongly account for the variance between the regions. The likelihood-ratio test indicates that $\hat{\sigma}^2_{r(n)}$ is not statistically significant ($p = 0.37$): in the 44 unique region × democracy × no internal chaos combinations not one difference between the random-effect estimate and the mean is statistically significant.

When the typological test variable is the four categories of slavery, the likelihood-ratio test indicates that $\sigma^2_{r(n)}$ is statistically significant ($p = 0.0185$); in two of 45 region × slavery combinations the difference between the random-effect estimate and the mean is statistically significant. When slavery and no internal chaos jointly classify the random regions, these attributes do account for the variability between the regions, but less strongly than democracy in combination with no internal chaos. The likelihood-ratio test indicates that $\sigma^2_{r(n)}$ is not statistically significant ($p = 0.115$); in the 62 region × slavery × no internal chaos combinations only two random-effect estimates differ from the mean by statistically significant values.

Of the five instrumental covariates, countries that are HIPCs consistently exhibit the largest negative effects on human development rank and scores.

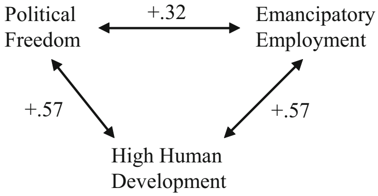
Discussion

Summary

This chapter clarified why the various regions of the world have different levels of human development. To do so, it measured instrumental factors and civilization zones and assessed how these attributes of countries influence the substantive freedoms the countries provide for their citizens. The latter were measured by the 1999 human development index and by its component scores for longevity, literacy, and gross domestic product per capita; subsequent research can build on this base by studying longitudinal change. The cross-tabulation of the form of measurement, HDI ranks or scores, with the goal of the analysis, quantification of effects or causal inference, organizes the four parallel analyses. When the response variable is the ranking of the 138 countries, and the regions are not nested by typologies, then the civilization zones have a larger associational effect on the HDI ranking than any of the indicators of the instrumental factors. Highly indebted poor countries (HIPCs) consistently exhibit low levels of human development, as do countries that are scarred by corruption and internal conflict and unrest. However, when the regions are nested by typologies based on the instrumental factors, fully democratic polities and emancipative employment (i.e., the absence of contemporary slavery) account for the differences in HDI rank between the regions. This basic pattern of results holds when the underlying summary HDI scores of the countries are modeled. But now, in order for democracy and emancipative employment to account for the regional variance in HDI scores, each of these attributes must nest the regions jointly with the indicator of no internal chaos.

Interpretation of Explanatory Effects

Why do full freedom and emancipatory employment account for the regional differences in the HDI? These three correlated constructs form a mutually reinforcing beneficent cycle of dynamic cumulative causation (Myrdal [1944] 1964, 1065–1070), as depicted below (the correlations are Spearman's $r_{s,}$ $p < .0001$):[24]



Countries enjoying full political freedom are more likely to have emancipatory employment (e.g., entrepreneurship and low slavery), and also higher human development. However, these correlations also imply that countries that lack full political freedom are also more likely to be characterized by the new slavery and low human development, forming a vicious cycle of relationships (Myrdal [1944] 1964, 75–78).

Beneficent and vicious cycles are described by Doris Kearns Goodwin (2005, 77–78), when she compares the North and slaveholding South in the USA circa 1835. Citing the historian Kenneth Stampp ([1959] 1991, 201), she reports:

> The North of this period "teemed with bustling, restless men and women who believed passionately in 'progress' and equated it with growth and change; the air was filled with the excitement of intellectual ferment and with the schemes of entrepreneurs; and the land was honey-combed with societies aiming at nothing less than the total reform of mankind.

Citing descriptions of the South by William and Francis Seward, who journeyed to Virginia from New York, Goodwin forms this picture (2005, 77):

> The poverty, neglect, and stagnation Seward surveyed seemed to pervade both the landscape and its inhabitants. Slavery trapped a large portion of the Southern population, preventing upward mobility. Illiteracy rates were high, access to education difficult.

> While a small planter aristocracy grew rich from holdings in land and slaves, the static Southern economy did not support the creation of a sizable middle class.

Goodwin (2005, 78) documents the cruelty of slavery using this quotation from Seward's (1877, 271) autobiography, in which he describes the treatment of slave children prior to their being auctioned:

> Ten naked little boys, between six and twelve years old, tied together, two and two, by their wrists, were all fastened to a long rope, and followed by a tall, gaunt white man, who, with his long lash, whipped up the sad and weary little procession, drove it to the horse-trough to drink, and thence to a shed, where they lay down on the ground and sobbed and moaned themselves to sleep.

Southern poverty, slavery, and limited democracy formed a vicious cycle whose effects still linger today (Burd-Sharps et al. 2008, 2009).[25]

The Role of Civilization Zones

The three dimensions of the HDI (longevity, literacy, and economic welfare) form a unitary measure: the same predictors—civilization zones and instrumental factors—have similar effects on each of the dimensions. Moreover, when a summary index of instrumental factors supportive of human development is correlated with the various response measures—HDI rank, HDI scores, and the scores for the components—the correlations are very high, thus further substantiating Sen's linkage between instrumental factors and substantive freedoms that hypothesis 2 foretold. Consistent with hypothesis 1, across all of the hierarchical models Huntington's civilization zones have important descriptive effects on human development. Because of these important effects and the popularity of explanations rooted in differences among civilizations and their cultures, the role of civilizations in social science explanations calls for further examination.

Because the civilization of a country as defined by Huntington is derived from cultural, religious, ethnic, and linguistic patterns, it is tempting to think that civilization zone is the key determinant of a country's instrumental freedoms, and its citizens' levels of rational thought and ability to reason. With the country as the unit of analysis, these variables are correlated, suggesting that their interrelationships are complex. To simplify, let us assume that slavery is the response variable, civilization zones are the key exogenous attributes, and political freedom, corruption, conflict, and debt are intervening control variables. When Proc Mixed is used to calculate the regression coefficients, civilization zones ($p < .0001$) and corruption ($p = 0.0007$) have significant fixed effects on slavery, whereas political freedom ($p = 0.275$), conflict ($p = 0.657$) and national debt ($p = 0.481$) do not. When the model is reestimated with civilization zones and corruption as the predictors of slavery, and the Bonferroni adjustments are made to test the differences in the predicted least-squares means between the various civilizations, the Hindu category has a significantly higher mean score on slavery (here primarily debt bondage) compared with these categories: African ($p = 0.0002$), Islamic ($p = 0.0001$), Latin ($p < .0001$), Orthodox ($p < .0001$), Sinic (0.026), and Western ($p = 0.0006$).

These significant differences between the Hindu category and all of the other categories disappear when the ratio of female primary school enrollment as a percentage of the male primary enrollment for years 1995 through 1997 is added as an additional control. The difference between the slavery mean for the Hindu category and the means of these other categories become not significant statistically as follows: African ($p = 0.896$), Islamic ($p = 0.884$), Latin ($p = 0.747$), Orthodox ($p = 0.813$), Sinic (p = 0.62), and Western ($p = 1.0$). Thus, women's primary school education relative to men's interprets the effect of the Hindu category on slavery; the civilization zone is not directly responsible. Women's primary and subsequent higher level education may have this beneficial effect because it increases literacy and ability to reason, enables literate women to help their husbands gain literacy, provides opportunities for better jobs that will enable families to move out of poverty, facilitates the exercise of political and economic rights, reduces excessive fertility, and increases demand for utilization of health services (Mehrotra 1997b, 102–105; Krishnan 1997, 213–229).

A country's civilization zone is an expressive indicator; whereas its score for the ratio of women's primary education to that of men is an underlying "trait" of the country, a more predictive efficient factor (Lazarsfeld 1959, 49–60). Because some of the efficient factors reside at a lower level of analysis, the linkages at the macrolevel between a country's civilization zone and its values on such properties as the HDI or slavery are not immutable. James Coleman's (1990, 6–23) macro to micro to macro explanatory paradigm clarifies this point (also see Nagel 1961, 481–485 and Berg-Schlosser 2003, 381). The initial observation links the Hindu zone with slavery, an important determinant of human development, but a control at the more microlevel for women's primary education relative to men's eliminates the direct effect of that civilization zone on slavery. Thus, social mechanisms at the microlevel can short circuit or modify macrorelationships.

Coleman's reductive paradigm is not the only viable approach toward explaining the relationship between two macrovariables. The multilevel data analysis clarified the relationship between region (x) and HDI (y) by nesting the regions by a test factor (t), a classification typology at the macrolevel (Lazarsfeld 1955b, *xi*). If with this control the original relationship disappears, then the test factor accounts for the initial relationship. Alternatively, consistent with Coleman's paradigm, the regression of slavery (y) on civilization zone (x) was interpreted by an intervening test factor (t), the ratio of women's to men's primary education; this test factor was not at a higher level of aggregation than the other variables.

Policy Implications

Civilization zones may matter intrinsically, but they do not irrevocably determine development outcomes; manipulable instrumental factors are important. A country in a unique civilization zone may have ameliorative mechanisms in place that countervail against suppressing causes like poverty and lack of women's (and men's) education, thereby improving its overall human development score. Even in Gujarat, India, where Hindu nationalism is strong, such social mechanisms are already in place. Inspired by ecumenical Gandhian thought, the Self-Employed Women's Association (SEWA) is a trade union of very poor, self-employed women workers—illiterate and economically vulnerable women who are street vendors, head loaders, block printers, street sweepers, and home-based workers (Vaux and Lund 2003, 265–270). SEWA provides them with a bank for credit and savings, health care, childcare, insurance, legal services, education, and housing, all of which work against debt bondage and poverty. In Muslim Pakistan, a country with high rates of debt bondage, reform has been less successful (Bales 1999, 149–194), whereas in Muslim Bangladesh, the Grameen Bank and the Bangladesh Rural Advancement Committee (BRAC) have reduced the debt bondage and poverty of women through microcredit and cooperative networks.

To further enhance human development, the UN's social and economic commissions could encourage their member countries to improve the strength of the instrumental factors that this research has documented; these are broadly consistent with the UN's Millennium development goals (UNDP 2003; Sachs 2005, Table 1, 3). Specifically, countries could enhance: social opportunity by eliminating debt bondage and other forms of slavery (especially via the education of children and adult females and males); political freedom by moving toward fully democratic political systems; transparency of economic transactions by minimizing corruption; economic facilities by mitigating national debt; and protective security by controlling civil disorder (Fukuda-Parr 2003). Provision of these changes may engender beneficial cycles that would strengthen civil society and encourage free human agency, which in turn would stimulate social and economic development and reduce poverty.

Poverty is most salient in developing countries but it also scars the highly developed (UNDP 2010, 94–99, 161, 221). In Western Europe many Muslims live in segregated communities marked by severe unemployment, horrendous crime, terrible housing, and inferior schools. These social and economic conditions may bear on the functioning of the system of variables that is the focus of the next chapter. It studies theoretically, empirically, and longitudinally several candidate causes of contemporary violence against Jewish people and property in ten European countries during the period of the second intifada against Israel.

Acknowledgements The author thanks Helen Fein, for collegial advice; Kevin Bales of Free the Slaves, for providing several of the measures; Irina Koltoniuc, for helping to prepare the analytic data base; Philip Gibbs of the SAS Institute, for clarifying nuances of Proc Mixed; and Misha Salganik, for critiquing the notation of the nested random effects. The author derived this chapter from his article in *Quality & Quantity* (vol. 43, issue 1, 337–361; © 2009 Springer).

Endnotes

[1] Welzel, Inglehart, and Klingemann (2003, 345–346) define the concept of human development as comprising socioeconomic development, emancipative cultural values, and effective democracy whereas this study conceptualizes health, education, and economic welfare as the key components of human development and relates political freedom to different levels of the HDI.

[2] The extensive use of the HDI and parallel indexes that assess gender-related development, women's empowerment, and poverty could advance the social sciences: Sociologist may find the HDI useful because its three combined components provide an overall measure of system performance; it bears on aspects of the AGIL scheme (Parsons et al. 1953). Neoinstitutionalists could use this index to assess how different institutional arrangements influence development (Meyer et al. 1997). Socioeconomists may find the HDI compatible because its broad definition works against the narrow use of satisfaction of material wants as the generic measure of development (Basu 2001, 71; Etzioni 1988). For communitarians (Etzioni 2001, 232–245) and operational philosophers (Rapoport 1950, 238–239; 1953, 93–102) this index provides minimal apolitical standards—literacy, a long and healthy life, and lack of poverty—that the United Nations certifies for making cross-cultural judgments. Political scientists, economists, and dependency theorists may find this index of interest because it provides an overall measure of development, but one that allows the separate analysis of its components (Adelman 2001, 117; Basu 2001, 72; Kentor and Boswell 2003; Inglehart 2003). Economic historians have used the HDI in their investigations of human welfare in the past (Crafts 2001, 323–326) and have proposed and tested improvements in it (Leandro Prados del la Escosura, personal communication). Burd-Sharps et al. (2008, 2009) and Lewis and Burd-Sharps (2010) have developed and applied human developments measures for the United States and component states.

[3] Sen (2002, 30–33) has critiqued Huntington's classification: Many countries do not have unitary cultures—India is classified as Hindu, but 150–160 million Muslims live there. The religious traditions of Japan's 124 million people are Shintoist (124 million) and Buddhist (93 million); some people are both Shintoist and Buddhist. African countries have different forms of past colonization, tribal organizations, languages, and indigenous religions. People have complex group affiliations and not just one salient civilization identity—Protestants and Catholics in Northern Ireland share a common Western zone but conflict mars their relationships. Partitioning people on the basis of their civilization or culture may contribute to conflicts in the world—not all Muslims dislike the United States. The concept of civilization zone (rather than civilization) mitigates some of Sen's critique, as does Galtung's (1992, 32–33) classification of countries based on geopolitics and power that is similar to Huntington's.

[4] The HDI and the indicators of human rights have these significant Spearman correlations ($p < .0001$): for human rights abuse, $r_s = 0.53$; for total trafficking in humans, $r_s = 0.35$; for restricted civil liberties, $r_s = 0.59$; and for trafficking from, $r_s = 0.49$.

[5] The variables and their effects are thus at the macrolevels of region, civilization zone, and country and do not necessarily characterize the relationships and attitudes of individual citizens (King 1997; Meier 2001, 30; Landes 1998, 516–517). For example, Huntington classifies the United States as a Western civilization (1997, 20–29) but this global property (Lazarsfeld and Menzel [1961] 1972, 228–229) does not imply that all people in the United States are Western and that this identity is most salient.

[6] Counter to intuition, which expects a uniform distribution, the Kolmogorov–Smirnov test does not reject the null hypothesis that the unweighted HDI scores are distributed normally (the $\hat{D} = 0.06$ and $p = 0.20$). Moreover, the ratio of the skewness of the distribution to its standard error is small ($-0.0113/0.191 = -0.06$); both of these statistics support the assumption of normality. However, the ratio of the kurtosis to its standard error is more extreme than the critical value of -1.96 (-1.191 / $0.380 = -3.13$), which indicates that the tails of the distribution are shorter than those of a normal distribution.

[7] This study analyzes the variability of the 18 regions and not the variability of the five commission areas because the latter are not exhaustive, the regions are more homogeneous than the areas, and the statistical model would be rather complicated—three levels with a fourth classificatory typology. For further information about the activities of these commissions and the definitions of regions see the *Yearbook of the United Nations* 2000 (2002, 923–958) and the *Statistical Yearbook* (2001).

[8] Beckfield classified Jamaica (Protestant, 61.3%; Roman Catholic, 4%) and Haiti (Protestant, 16%; Roman Catholic, 80%) as Latin. On the basis of an anthropologist's insights, this study classifies these countries as Western. Huntington's map indicates that the various islands of the Philippines (Roman Catholic, 83%; Protestant, 9%; Moslem, 4%; Buddhist, 3%) may be Sinic, Buddhist, or Western. Beckfield chose to classify this hybrid country as Sinic, whereas due to its Catholic, Spanish and American heritages this study classifies it as Western. Papua New Guinea (Christian, 66%; Indigenous beliefs, 34%) is classified as Western. The use of these distinctions does not imply agreement or disagreement with Huntington's depictions.

[9] The region-to-region variability of slavery is $\hat{\sigma}_s^2 = 0.36$ ($z = 2.28$, $p = 0.011$); its country-to-country variability within regions is $\sigma_c^2 = 3.25$ ($z = 8.14$, $p < .0001$). The regional mean level of slavery is 2.02 on the scale of 1 to 4; the regions of Western Africa ($p < .0015$), South-Central Asia ($p < .0001$), and South-Eastern Asia ($p < .002$) have higher levels; North America ($p < .035$) lower levels.

[10] York, Rosa, and Dietz (2003, 290) created (0, 1) indicator variables for political rights and for civil rights and found no effects on the ecological footprint of a country. Their groupings are free = 1 and 2, partly free = 3 through 5, and not free = 6 and 7.

[11] The region-to-region variability in political freedom ($\hat{\sigma}_f^2 = .29$, $z = 2.35$, $p = .0095$) is statistically significant as is the country-to-country variability within regions ($\sigma_c^2 = 1.78, z = 8.53, p < .0001$). The mean level of political freedom is 2.2. North America ($p < .024$), Northern Europe ($p < .017$), Southern Europe ($p < .04$), and Western Europe ($p < .009$) have higher levels of freedom. Northern Africa ($p < .002$), Eastern Africa ($p < .045$), Middle Africa ($p < .0041$), Eastern Asia ($p < .039$), and Western Asia ($p < .009$) have lower levels.

[12] Linz (2000, 37) states that a free enterprise, liberal economic infrastructure does not necessarily lead to the development of a liberal political democracy. Prezeworski et al. (2000, 178–179) find that economic development does not tend to create democracies, but wealthy societies that are democratic tend to be stable.

[13] To count democracies Dahl (1998, 198) suggests regrouping the Freedom House measures as follows, group countries ranked 1, or most free, on political rights with countries ranked 1, 2, or 3 on civil liberties. His grouping enhances the effects of democracy on the HDI.

[14] This list, which was downloaded from the Jubilee 2000 web site, is very similar to that reported by Easterly (2002, 128). His recent list of HIPCs also includes Burundi, the war-torn Congo (Democratic Republic), and Guinea. For the period 1979–1997 because of missing data he analyzed only 28 out of 37 of these countries.

[15] The region-to-region variability in internal conflict is not large ($\hat{\sigma}_{ic}^2 = 0.158$, $z = 2.1$, $p = 0.018$) but significant, as is the much larger country-to-country variability within regions ($\sigma_c^2 = 1.6$, $z = 8.6$, $p < .0001$). The mean level is 0.75. Middle Africa ($p < .003$), Northern Africa ($p < .031$), and South-Central Asia ($p < .005$) have higher levels. North America ($p < .03$), Northern Europe ($p < .02$), and Western Europe ($p < .01$) have lower levels.

[16] The region-to-region variability in perceived corruption ($\hat{\sigma}_{pc}^2 = 3.5$, $z = 2.73$, $p = 0.0032$) is significant as is the country-to-country variability within regions ($\sigma_c^2 = 5.65$, $z = 8.54$, $p < .0001$). The mean is 4.45. North America ($p < .0001$), Northern Europe ($p < .0001$), Western Europe ($p < .0001$), and Oceania ($p < .0014$) have less corruption than the mean. Eastern Africa ($p < .006$), Middle Africa ($p < .0035$), Western Africa ($p < .001$), South-Central Asia ($p < .015$), and South-Eastern Asia ($p < .006$) exhibit more than the mean.

[17] Strictly, this model requires a random sample of regions from a population of regions, and a random sample of countries within each region. However, this chapter follows an approach that is often applied in analysis of variance. Imagine these data being arrayed by region and by country within region and for each of these observations there is a mean HDI score for a particular year. Now imagine that each of that year's HDI scores represents the result of random sampling from a distribution of hypothetical replications for each region-country unit. Although this study focuses on the HDI scores for all regions and countries in 1999, the year of the human development report could have been selected at random from the distribution of reports from 1990 through 2005. Each report replicates earlier reports. For assumptions of this kind see King, Keohane, and Verba (1994, 56–59, 76–82) and Moore and McCabe (1989, 714–15, 719).

[18] Proc Mixed will compute the confidence limits for the random effects producing either the asymmetric Satterthwaite limits or the symmetric Wald limits (SAS Institute 1997a, 585). The Parms/Nobound statement allows it to calculate the symmetric Wald confidence limits, which this study prefers. If this statement is not used and if the random effect is bounded by zero it calculates the Satterthwaite limits. For these data the upper bound of the Satterthwaite limits are very high and the lower bound is near zero. For the Wald limits the lower bound often is negative and must be truncated to zero and the upper bound is much lower than that of the Satterthwaite limits.

[19] For discussion of how terms like σ_r^2 can be conceptualized as a covariance in a compound symmetric structure see Littell et al. ([1996] 2006, 174–177).

[20] Lack of information about the level of the new slavery is primarily responsible for the reduction in the number of cases from 179 to 138. The 30 countries that have problematic slavery codes of zero are: Bahamas, Belize, Comoros, Cuba, Cyprus, Dominica, East Timor, Fiji, Finland, Grenada, Honduras, Iceland, Ireland, North Korea, Latvia, Lithuania, Maldives, Malta, Micronesia, New Zealand, Nicaragua, Norway, Saint Kitts and Nevis, Saint Lucia, Saint Vincent and the Grenadines, Samoa (Western), Seychelles, Solomon Islands, Taiwan, Vanuatu. It is not clear whether the zero code implies little or no slavery or insufficient information. This question should be resolved prior to applying multiple imputation of missing data. Many of these zero-coded countries also lack region codes. Antigua and Barbuda, Hong Kong (China), and Sao Tome and Principe lack region codes and perhaps other information as well. The following 14 countries generally lack substantive information: Afghanistan, Bosnia, Iraq, Liberia, Libyan Arab Jamahiriya, Macedonia, Moldova Republic, Myanmar (Burma), Somalia, Turkmenistan, and United Arab Emirates.

[21] The Shapiro–Wilk test, stem and leaf plot, and normal probability plot are consistent with the assumption that the residuals of these models are normally distributed. The residuals suggest that Mauritius, Singapore, Barbados, Israel, and Japan often have better human development scores than predicted by the models; Laos, Mongolia, Albania, Yemen, and especially Papua New Guinea have worse scores.

[22] Since region and civilization are both macrovariables, a plausible alternative model would cross-classify the countries by these two variables (Littell et al. [1996] 2006, 542–549). However, when this model is estimated the region × civilization random effect is zero. Moreover, when this interaction is classified by either freedom or slavery, the random effect is not significant. These null effects support the choice of models of this study.

[23] When the ratio of female primary school enrollment as a percentage of the male primary enrollment (for years 1995 through 1997) taken from the World Bank Index is added as an additional control, it strengthen slavery's explanation of the variability that is between regions; the $\hat{\sigma}^2_{r(n)} = 100.7$, $z = 0.87$, and the $p = 0.192$. The latter probability is higher than that of any of the other models. This variable weakens the explanatory power of the dichotomous democracy typology; the $\hat{\sigma}^2_{r(n)} = 265$, $z = 1.50$, and the $p = 0.067$. However, when women's primary education is used as an additional covariate it severely increases the missingness of the data. Also, this variable may be confounded with the knowledge dimension of the HDI. Subsequent research will address these issues.

[24] The partial Spearman correlations are (clockwise from Full Political Freedom): 0.13, 0.48, and 0.40.

[25] Nathan Nunn (2008) finds that African countries that suffered the most from the slave trade in the past suffer the most from reduced economic development in the present; namely, the greater the number of slaves exported from a country, the worse the current economic performance.

Chapter 8
A Globalized Conflict

'CST' [Community Security Trust, a Jewish organization] argues that there is a consistent link between the increases in the number of antisemitic incidents in Britain and the heightening of tensions in the Middle East.

—European Union Monitoring Centre (2004b, 199)

Antisemitic acts [in France] are ascribed to youth from neighbourhoods sensitive to the [Mid-east] conflict, principally of North African descent.

—European Union Monitoring Center (2004b, 98)

The Middle East has lit the match that kindled the recent fire, but what fuels the fire are the meeting points between the interests of radical Islamists and various factions of the European left and the extreme right.

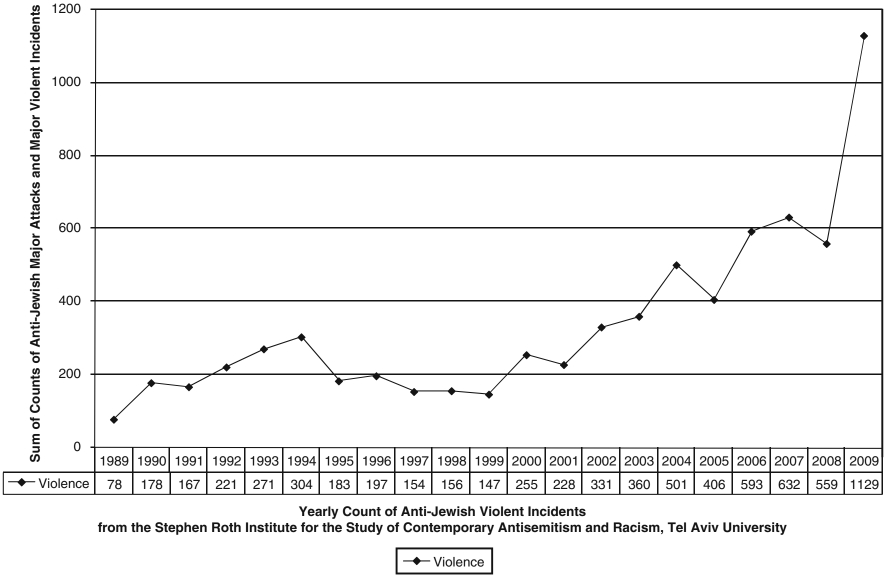
—Stephen Roth Institute (2004)

Operation Cast Lead in January 2009 triggered a wave of antisemitic manifestations. . . .This trend subsided in February and March, but even during the months that followed this peak of antisemitic incidents, the baseline remained higher than before the war. In fact, there has been a rising trend since the early 1990s, even in years when there was no significant Middle East trigger. Thus, the origins of the 2009 escalation in antisemitic expressions must lie deeper.

—Stephen Roth Institute (2010, 2)

During the period 1989 through 2009, the incidence of anti-Jewish violent events increased worldwide from the relatively low counts in 1989 to the extremely high counts in 2009. This violence is a major social problem not only because of the pain inflicted on the Jewish people and their communities, but also because the perpetrators become desensitized to their actions making them less humane and thus more likely to target non-Jews. Moreover, the apathetic response of the non-Jewish communities toward this violence weakens societal cohesion, which undermines the ability of the larger society to respond effectively to the anti-Jewish violence, the threat of terrorist attacks, and ordinary crime.

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_8, © Springer Science+Business Media B.V. 2011



	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
◆ Violence	78	178	167	221	271	304	183	197	154	156	147	255	228	331	360	501	406	593	632	559	1129

Yearly Count of Anti-Jewish Violent Incidents
from the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism, Tel Aviv University

Fig. 8.1 A major social problem: Anti-Jewish violence worldwide, 1989–2009

For the period 1989 through 2009, Fig. 8.1 depicts the sum of the counts of major attacks (e.g., shootings, knifings, bombings, and arson) and major violent incidents (e.g.,vandalism and physical aggression) directed against Jewish people worldwide, as tabulated by the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism, Tel Aviv University (2010). This trend line spans four periods of intense Israeli–Palestinian conflict: the Palestinian's first intifada (i.e., uprising) against Israel (1987–1993), the second intifada (2000–2004), the second Lebanon war (7/12–9/8/2006), and the Gaza war (12/27/2008–1/18/2009). As the trend line suggests, anti-Jewish violence worldwide ebbed and flowed in part with the levels of the violence in the occupied territories, Israel, and Lebanon. The first intifada began in 1987 and ended with the beginning of the Oslo Peace Process in 1993. During this period, violent anti-Jewish events increased from about 78 in 1989 to 271 in 1993, peaking in 1994 (304 events). Thereafter, the counts declined until the beginning of the second intifada in 2000 (255 events) and increased each year until 2004 (501 events), when the uprising abated. The counts declined in 2005 (406 events) and then, in part as a consequence of the war in Lebanon in the summer of 2006, the counts increased to higher values in 2006 (593 events) and 2007 (632 events). After declining in 2008 (559 events), violence increased to extremely high counts (1,129), which were triggered primarily by Israel's Operation Cast Lead in the Gaza strip. Apparently, with globalization, these Mid-east conflicts have contributed to anti-Jewish violence throughout the world. Globalization implies the transnational interconnections of peoples and ideologies brought about by economic interdependence, migrations, the mass media, nongovernmental agencies, religious organizations, and social networks (Watson 2004, 143–153).

This chapter focuses primarily on a consequence of globalization: it models anti-Jewish violence in 10 European countries for the period of the second intifada. This uprising began toward the end of September 2000, with the killing of an Israeli soldier on the 27th and Ariel Sharon's provocative visit to the mosque compound on the Temple Mount in Jerusalem the next day. Sharon had received prior permission from Yasser Arafat and other Palestinian officials for him and his Likud party followers to visit the area. This visit triggered the wave of violence that resulted in the deaths of at least 3,223 Palestinians and 950 Israelis through 2004. Although this uprising did not result in peace, the violence became less intense with the death of Arafat in November 2004, Mahmoud Abbas's ascension to the leadership of the Palestinian Authority, the truce of Sharm al-Sheikh that Abbas and Sharon declared in February 2005, and Israel's withdrawal from their Gaza settlements.[1] Unfortunately, violence still continues today in the Middle East, which no doubt exacerbates anti-Jewish violence in Europe and worldwide.

An earlier study explored how the following factors contributed to this violence against European Jews (Smith 2004): Countries with populations comprising numerous Jews and numerous Muslims exhibited elevated counts of violence. This finding, although based on correlations between aggregated variables, is consistent with recent interpretations of the interpersonal violence against Jews in Europe as being perpetrated in large part by young Muslim males of North African origin. Apparently, the youths' perception of events in the occupied territories and Israel combine with social support from their communities and with antisemitic propaganda emanating from Arab countries and local sources to mobilize them to commit anti-Jewish violence.

This earlier study further pointed to an inconsistency. On the one hand, countries that had experienced deportations of their Jewish populations to the death camps during the Holocaust had lower counts of anti-Jewish violence. On the other hand, countries that exhibited high awareness of the Holocaust (knowledge and remembrance) had slightly higher counts of violence.[2] This inconsistency suggests an ambivalence among ordinary Europeans that may inhibit them from taking ameliorative actions. As Martin Luther King stated: "Man's inhumanity to man is not only perpetrated by the vitriolic actions of those who are bad. It is also perpetrated by the vitiating inaction of those who are good."[3] Building upon King's insights, this chapter reviews relevant prior theorizing and research; explores anti-Jewish violent incidents in ten European countries; identifies the likely perpetrators; and develops, tests, compares, and replicates multilevel models that account for the violence during the time period of the second intifada.

Prior Theory and Research

Antisemitism refers to negative attitudes, prejudice, and hostility toward Jewish people simply because of their identity as Jews; a non-Jew mistakenly perceived to be Jewish could be subjected to an antisemitic attack (EU Monitoring

Center 2004a, 4–5). This chapter refers to such contemporary dispositions as "antisemitic," distinguishing this from the "anti-Semitic" beliefs of the 1930s and 1940s. In anti-Semitism, an imagined Semitic race was responsible for the stereotypes. In antisemitism, the stereotypes are rooted in anti-Jewish cultures.

Theoretical modeling of empirical data rarely focuses on the sources of contemporary European violence against Jews. Because of this absence of directly relevant research, this chapter draws on more general theories. In the context of the globalization of the second intifada, the guiding theoretical constructs stem from these four hypothetical determinants of the violence: the relative sizes of a country's Jewish and Muslim populations; how interpretations of the events in the Middle East mobilize the perpetrators; the unresponsiveness of bystanders; and the ambivalence of ordinary Europeans.

Population Characteristics

Previous studies by Spilerman (1970, 1971, 1976), which linked population density and urban violence, suggest that a country's population sizes of Jews and Muslims may engender its level of anti-Jewish violence. To explain the riots that marked urban ghettoes in the USA during the 1960s, Spilerman applied a number of statistical models—the Poisson, gamma, and negative binomial—to the time series of the counts of these violent protests. Guided by how well the various models fit the data, he concluded that the communities were heterogeneous in their propensity toward disorder. He rejected models that assumed an identical probability of experiencing a disorder (simple Poisson models), positive or negative reinforcement of riot propensity, and contagion from one community to another. To test his mathematical theorizing, he applied regression analysis and found that when he statistically controlled for the sizes of the Negro communities, other community characteristics (such as poverty and political representation) had little or no effect on riot propensities. He viewed population size as antecedent to other community characteristics and to the riots and thus viewed their initial correlations as spurious.

Mazur (1972) critiqued Spilerman's analysis primarily by applying Poisson models within each year of data and then by conducting other analyses from which he concluded that the causes of riots are complex and not well understood. Spilerman (1972) rebutted Mazur's critique but he conceded that that when communities were grouped by population sizes and simple Poisson models applied within these groupings, then those models fit the data.[4] Based on the Spilerman–Mazur debate, this chapter conjectures that the relative sizes of a country's Jewish and Muslim populations influence that country's propensity toward anti-Jewish violence, and that the Poisson distribution can provide an initial model of the counts of the violence. Contemporary methods of multilevel modeling of Poisson sampling distributions, which combine mathematical and statistical models, are thus relevant methods for this study.[5]

Mobilization

Previous studies of the use of resources, propaganda, and conditions of life bear on the mobilization of the perpetrators of anti-Jewish violence.

Mobilization of resources

The resource mobilization theoretical perspective focuses on the strategic thinking of the leaders of a social movement and on the discontents and grievances of the protestors (Tilly 1978, Zald and McCarthy 2002). The second intifada was not merely the spontaneous protests of oppressed Muslims over their grievances; for it can also be viewed as a somewhat planned series of protests that Palestinian leaders designed to pressure Israel toward changing its policies regarding the occupied territories and the Israeli settlements there. The leaders of the insurgency and their followers used resources like these: the mass media to characterize the Israeli army as repressive, antisemitic propaganda to demonize Jews and Israelis, attacks on Jews in Europe to reduce their support for Israel, and suicide bombers to demoralize the Israelis (Pape 2005, Brym and Bader 2006).

Propaganda

Contemporary anti-Jewish propaganda attacks the Jewish people, Zionism, and Israel; see Table 8.1, which organizes statements about Jews reported in "Antisemitism Worldwide 2003/4, General Analysis" (Stephen Roth Institute 2004). Such propaganda may reach ordinary Muslims in Europe via television programs emanating from the Arab countries, the internet, exhortations from radical clerics, and the writings of Muslim intellectuals; see for example Samarah (2001).[6] Although people may claim that they are against Zionism and against the State of Israel and not personally antisemitic, the indicators may suggest otherwise. For each main topic (Jews, Zionism, and Israel) there are three repeated themes: power, evilness, and abolishment.[7] Such propaganda may mobilize perpetrators to commit violent acts against Jews, immobilize ordinary Europeans to act against the perpetrators, and demoralize the Jewish community.[8]

Antisemitic propaganda can influence Jews to weaken their support of Israel and to accept the negative stereotypes about their group; the latter bears on "self-hate" and "identity threat." Lewin ([1942] 1999, 330) referred to the acceptance of negative stereotypes about one's ethno-religious group as self-hate and suggested ways for Jewish people to avoid succumbing to the antisemitism that was rampant in Europe and common in the USA during the 1930s.[9] Pride in the Zionist movement was pivotal; it helped inoculate Jews against this antisemitism ([1948] 1997, 140–141). The social psychologist Claude Steele refers to the acceptance of negative stereotypes about one's group as "identity threat" (1997). Such threats

Table 8.1 Anti-Jewish themes gleaned from the Roth Institute's "Anti-Semitism Worldwide, 2003/4, General Analysis"

Attitudes about Jewish people	
Jews are powerful	
	They control the world with their money.
	They are the secret rulers of the world.
	They influence leaders of governments
	They control the American government and media.
Jews are evil	
	They are the "root of all evil".
	They are bloodthirsty.
Abolish remembrance of the holocaust	
	Jews lie about the gas chambers and the holocaust.
	Anne Frank's diary is a fake.
Attitudes about Zionism	
Zionism seeks world domination	
	Zionism is a synonym for Jewish world power.
	Zionism is a centuries old trend that aims at world domination.
	A Zionist leadership is ruling the US.
Zionism is evil	
	Zionism and Nazism are the same.
	Zionism is worse than Nazism.
	Zionist governments are inhuman and genocidal.
Abolish the Jewish state	
	Ending the Jewish state would solve the Middle East conflict.
	A Jewish state is an oxymoron.
Attitudes about Israel	
The army uses its military power ruthlessly	
	It is perpetrating a real genocide of the Palestinian people.
	It massacres Palestinian children.
	It sells organs removed from dead Palestinians.
	It uses crude Nazi methods.
	Gaza is similar to the Warsaw ghetto.
Prime Minister Ariel Sharon is evil	
	He should be prosecuted for crimes against humanity.
	He is a pig excreting on Arabs.
	He is clearly a Nazi.
The Israeli People are evil	
	They are the incarnation of Nazi mentality and ideology.
	They are doing to the Palestinians what the Nazis did to them.
	They are oppressing the Palestinians.
Abolish Israel	
	Israel has no right to exist as a Jewish state.
	It is an artificial state that aims to destroy the Arab states.
	It is a Nazi country that attacks defenseless Palestinians.
	It is a threat to world peace.
	Its corrupt political culture is spreading to the United States.
	It has influenced the U.S. to abandon civil rights and to become militaristic.
	Americans should divest themselves of investment in Israel.

may lead to "disidentification," the uncoupling of the person from the identities that are threatened. For example, contemporary European Jews who are experiencing either negative propaganda about Israel or anti-Jewish violence might downplay their identification with Israel, stressing that they are Jews but not Israelis, and muting their protest about the violence.

Conditions of life

During the violent protests of the 1960s in the USA, some scholars formulated the "riffraff" thesis that placed the blame for these urban riots on the small percentage of the black population (1 or 2%) who participated—the riffraff—and on criminals. However, empirical research revealed that about 18% of the rioters lived in the riot area, that they represented a wide cross-section of young black males, and that many of them had no prior arrests (Fogelson and Hill 1968). To quell these protests, the researchers suggested that the institutional causes be identified and eliminated.

About 15% of European Muslims are unemployed (http://Euro-Muslim.info) and many live in congested, slum-like conditions. Well-meaning people sometimes blame these deleterious environments for the violence of the young men against Jews, whom they perceive as supporting Israel, their enemy, or against other Muslims, whom they perceive as violating strict codes of conduct. However, a myopic focus on those perpetrators may obscure other relevant factors that should be dealt with: the support that the Muslim community expresses for the violence; the importation of antisemitic propaganda from Arab countries via television; the low regard European public opinion has for its Muslim residents; anti-Israel bias to the news reports about conflicts in the Middle East; the actions of skin-heads who are on the extreme right; and the inability of the citizens, police, and courts to control the perpetrators and limit the violence.

Bystander Unresponsiveness to Crimes

Bystanders cannot be relied on to prevent attacks on other people, as this sad happening illustrates: On the night of March 13, 1964, in a middle-class neighborhood in Queens, New York, Catherine "Kitty" Genovese was stabbed three times over the course of a half an hour. Thirty-eight people watched the stabbings from their homes. No one directly intervened to help her or to apprehend the killer. Only one person called the police, albeit belatedly. The onlookers' reluctance to help puzzled the police and the public. Even the onlookers, none of whom disliked the victim, were bewildered by their passivity.

To solve this puzzle of bystander reluctance to help, Darley and Latané (1968, 1970; Darley 2000) conducted a series of laboratory experiments that tested their hypothesis that a person's taking responsibility for aiding a victim becomes attenuated when the number of bystanders is increased from one to many—there is diffusion of responsibility. The bystanders may inhibit each other from taking helpful actions because they may not define the situation as an emergency; and, if they do perceive a crisis, they may assume that someone else will provide the needed aid. Personal animosity toward the victim is not a factor. This absence of animosity provides a foundation for this chapter's bystander unresponsive model, in which positive sentiments toward Jews are a key component.

In the case of anti-Jewish violence, unresponsive bystanders are countless and helpful interventions few. One example is an event that prompted the Simon Wiesenthal Center (*Response* 2004, 4) to issue a warning to Jews against travel to Belgium:

> The Center condemned the stabbing of a Jewish student by a gang of 15 young Arabs outside of a Yeshiva in a suburb of Antwerp. 'This latest daytime attack was almost inevitable, considering that Belgian authorities have done virtually nothing to counter the rise of antisemitism in their country.' charged Rabbi Cooper. 'The Jewish community is increasingly vulnerable against the unending virulent anti-Israel propaganda and over-the-top media bias against the Jewish state and her supporters. Authorities have failed to take the necessary steps to insure the safety and security of Jewish institutions. In fact, the Jewish school that the young victim attends has been targeted before.'

Although data presented later in this chapter indicate that Belgium has higher than expected rates of anti-Jewish violence, its citizens are not marked by high levels of thinking in terms of conventional antisemitic stereotypes.

Ambivalence

To understand ambivalent attitudes toward Jews, studies of cross-pressures and cognitive dissonance are helpful.

Cross-pressures

When identities, attitudes, and interests are consistent, they will reinforce a person's choice of actions that are associated with these dispositions. When identities, attitudes, and interests are inconsistent, the resulting cross-pressures may pull the person in different directions, leading her to temporize or to withdraw from the situation (Lazarsfeld et al. [1944] 1948). Ordinary Europeans personally may disapprove of violence against Jews and other groups, but active involvement to ameliorate the violence may not be in their immediate self-interest. The cross-pressure between attitude and interest may mute their response; the anti-Jewish violence is a peripheral concern.

Cognitive dissonance

Cognitive dissonance arises when two relevant cognitive elements are inconsistent: the obverse of one element follows from the other (Festinger 1957, 13). Dissonance is uncomfortable, motivating the person to reduce or eliminate the inconsistency thereby achieving cognitive consonance. Consistency can be achieved by changing one of the dissonant cognitive elements, adding new cognitive elements, or reducing the importance of the dissonant elements.[10] The perpetrators of the anti-Jewish violence may feel that their actions are consistent with their own attitudes, the psychological environments their social networks provide, and antisemitic propaganda. Feeling no cognitive dissonance, they act violently without inhibition. Only when their cognitive and social supports atrophy will they desist.

Many ordinary Europeans do not exhibit antisemitic attitudes. In theory, therefore, the attacks against their Jewish co-citizens ought to engender an uncomfortable sense of dissonance. A seemingly obvious way for them to reduce this dissonance would be to protest the anti-Jewish violence of the perpetrators, thereby making their own behavior consonant with their own attitudes. But ordinary Europeans are not visibly protesting the violence that is anti-Jewish. This chapter aims to explain why.

Exploring the Data

To gain an understanding of the data upon which the multivariate analyses rest, this section qualitatively explores the interrelationships among Jewish and Muslim population distributions, the counts of violence, antisemitic attitudes, and attitudes about the conflict between Israel and the Palestinians. The subsequent quantitative analyses will then formalize and advance the intuitions gained from these qualitative explorations:

During the period of the second intifada, whenever the conflict in the Middle East intensified, the violence in Europe surged (Porat, 2004, 1):

> In terms of a timeline, we are looking at three waves of violent events—a period of some 6 weeks in October–November 2000, coinciding with the outbreak of the second intifada; then a lull in the acts of violence and their renewal in September–October 2001, coinciding with the Durban Conference and in the events of September 11; and then another lull up until Passover 2002 and the start of the Israeli army's Operation Defensive Shield, signaling another round that continued until the main stages of the elections in France in July 2002.

Because monthly data were not available, and yearly data for 2004 and 2005 were not yet available when the main analysis was conducted, this section examines the yearly counts for the early period of the intifada, for the years 2001, 2002, and 2003. These data were downloaded from the web site of the Stephen Roth Institute. This Institute monitors contemporary manifestations of antisemitism and racism around the world as well as the activities of extremists and hate groups (right-wing, left-wing, and Islamist). Its database includes event descriptions that researchers have gleaned from one or more sources such as newspaper or journal articles, electronic publications, leaflets and pamphlets, advertisements, posters, cartoons, video cassettes, books and monographs, and official releases.

A country's yearly count of violence that this chapter models is the sum of the Institute's tabulation of major attacks (including shootings, knifings, bombings, and arson) and the much more frequent major violent incidents (physical aggression without the use of a weapon and vandalism).[11] These data were used rather than counts from police records or from each country's monitoring agencies because the Institute has applied its criteria, definitions, and methods for counting events consistently from year to year across the countries. As expected, the Institute's counts of events vary with the countries' sizes of their Jewish and Muslim populations and with attitudes; however, these counts consistently underestimate actual anti-Jewish incidents.

Jewish and Muslim Populations

Population estimates of European Jews and Muslims are usually stated imprecisely as the percentage of a country's population (World Jewish Congress 2004; Central Intelligence Agency 2003). Here, these rough percentages were multiplied by accurate population figures from the census of the European Union (2004) and, on this basis, the countries were cross-classified according to their numbers of Jews and Muslims as follows: Countries with 100,000 or more Jews have "many" Jews; those with less than 100,000 have "few" Jews. Countries with 500,000 or more Muslims have "many" Muslims; those with less than 500,000 have "few" Muslims. Austria, Belgium, Denmark, and Switzerland have few Jews and few Muslims; Italy, the Netherlands, and Spain have few Jews and many Muslims. Because the counts of violence are similar for these two categories and they exhibit no difference in effect, in the multilevel models of data on the earlier period of the intifada these categories were grouped together as having few Jews, regardless of the number of Muslims. France, Germany, and the United Kingdom (hereafter UK) have many Jews and many Muslims and thus can be expected to have the highest propensity toward high counts of anti-Jewish violence. When this chapter replicates the main analysis using data for five years (2001 through 2005), the separate effects of the three population categories are assessed.

The ten European countries under study have only about 1.2 million Jews compared with 12.8 million Muslims, for an overall ratio of about 11 Muslims per Jew. Immigration has enhanced this Islamization of Europe and, as Muslims gain citizenship, makes them a potential political force. If Muslims are in fact the main perpetrators of anti-Jewish violent acts, then, given the sufficient numbers of Jews and Muslims in all ten countries, concern about the violence should be voiced not only by the Jewish and Muslim communities, but also by ordinary citizens and governmental agencies.

For the 3-year period from 2001 through 2003, Table 8.2 reports the classification of the countries, their population, the estimated number of Jews in these countries prior to the Holocaust (Fein 1979, 52–53 and other sources), the estimated current numbers of Jews and of Muslims, the ratio of Muslims to Jews, the counts of anti-Jewish violence (major attacks plus major violent incidents), and the rate of these events per 10,000 Jews for the 3-year period (not annualized). Countries with many Jews and many Muslims have the highest counts of anti-Jewish violence,

Table 8.2 Jewish and Muslim populations and anti-Jewish violence

Classification of country	Country population 2003	Jewish population 1935–1941	Jewish population 2004	Muslim population 2004	Ratio of Muslims to one Jew (2004)	Count of Anti-Jewish Violence 2001–2003	Rate of Violence per 10,000 Jews 2001–2003
Few Jews (0) and Few Muslims (0)	Austria 8,174,762	185,000	8,000	326,990	41	7	8.8
	Belgium 10,355,800	100,000	35,000	350,000	10	39	11.1
	Denmark 5,383,500	8,000	8,000	107,670	13	12	15.0
	Switzerland 7,317,900	18,000	18,000	182,948	10	14	7.8
Sub Total	31,231,962	311,000	69,000	967,608	14	72	10.4
Few Jews (0) and Many Muslims (1)	Italy 57,321,000	55,000	35,000	700,000	20	9	2.6
	Netherlands 16,192,600	140,000	30,000	750,000	25	8	2.7
	Spain 40,683,000	3,000	40,000	1,220,490	31	11	2.8
Sub Total	114,196,600	198,000	105,000	2,670,490	25	28	2.7
Many Jews (1) and Many Muslims (1)	France 59,630,100	350,000	650,000	4,472,258	7	154	2.4
	Germany 82,536,700	504,000	100,000	3,200,000	32	69	6.9
	UK 59,328,900	300,000	280,000	1,500,000	5	133	4.8
Sub Total	201,495,700	1,154,000	1,030,000	9,172,258	9	356	3.5
Total	346,924,262	1,663,000	1,204,000	12,810,356	11	456	3.8

Notes: UK = United Kingdom. Estimates of Muslim populations are from the CIA World fact book and various other sources. Counts of anti-Jewish violence are from the website of the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism. Countries with 100,000 or more Jews have “Many Jews”, Countries with 500,000 or more Muslims have “Many Muslims”.

averaging about 119 events per country. In countries with few Jews and many Muslims, the average count per country is much lower, only 9.3. The average count for countries with few Jews and few Muslims is about 18. Belgium is an outlier, contributing 39 violent acts. When Belgium is excluded, the average for the other three countries is about 11. Even with Belgium included, the average per country count of violence for the seven countries with few Jews (regardless of the number of Muslims) is much less than the average for the three countries with many Jews and many Muslims, 14.3 compared with 118.7.

Belgium may have higher violence counts because the Jewish population is concentrated in two large cities, Brussels and Antwerp; the Antwerp Jews live in densely populated Jewish areas, and their religious Orthodoxy may make them highly visible as potential targets. Right-wing Flemish nationalism is strong—in 2004 the Vlaams Blok won 24% of the Flanders vote (Minder, 7/21/04); and the left disliked the government of Ariel Sharon. These salient political and social forces may immobilize the government, the police, and ordinary Belgians, thereby allowing the violence to continue.

The rates of violence per 10,000 Jews provide a different perspective. In countries with many Jews and many Muslims the rates are not the highest—2.4 for France, 6.9 for Germany, 4.8 for the United Kingdom, for an average of 3.5—France has the highest count of anti-Jewish violence (154) but the lowest rate per 10,000 Jews (2.4). When Jews are few and Muslims many, these rates are lowest, about 2.7 per 10,000 Jews. In countries with few Jews and few Muslims, the rates are higher than in any other grouping—8.8 for Austria, 11.1 for Belgium, 15.0 for Denmark, 7.8 for Switzerland, and 10.4 over all. In these countries where Jewish populations are relatively small, a few incidents can increase the rates dramatically; this chapter thus models the raw counts.

Antisemitic Attitudes

The measures of a country's antisemitic attitudes are taken from the reports of the Anti-Defamation League (hereafter, ADL) about their cross-national surveys (2002a, 2002b, 2004, 2005), which were conducted through telephone interviews by the staff of Taylor Nelson Sofres (TNS). They surveyed the UK, France, Germany, Belgium, and Denmark between May 16 and June 4, 2002; the Netherlands, Austria, Italy, Spain, and Switzerland between September 9 and September 29, 2002; and all ten countries between March 16 and April 8, 2004. Additionally, they surveyed those ten countries and also Poland and Hungary between April 11 and May 6, 2005.[12] Prior to the reporting of the distributions of the variables, the 500 completed surveys for each country at each time point underwent minor weighting to national population data using governmental information on age and gender. The margin of error for each country is ± 4.4% at the 95% level of confidence.

The 2002 and 2004 surveys asked respondents in all ten countries four standard questions that tap antisemitic stereotyping—these items are based on the early work of Glock and Stark (1966, 101–161)—and a fifth question about the Holocaust, which is highly correlated with the four items (ADL 2004, 4, 21). The following three antisemitic themes organize these questions: Jewish in-group loyalty to other Jews, Jewish corruption, and dislike of Holocaust remembrance.

"I am now going to read out a series of statements, some of them you will agree with and some of them you will not. Please say which ones you think are probably true and which ones you think are probably false":

Jewish in-group loyalty to other Jews:
Jews don't care what happens to anyone but their own kind ("Own Kind").
Jews are more loyal to Israel than to this country ("Loyal to Israel").

Jewish corruption:
Jews are more willing than others to use shady practices to get what they want ("Shady").
Jews have too much power in the business world ("Too Much Power").

Dislike of Holocaust remembrance:
Jews still talk too much about what happened to them in the Holocaust ("Too Much Shoah Talk").

Table 8.3 presents Spearman rank correlations (r_s) of the responses to these questions and their relationships with the combined counts of violence for 2001 and 2003. The unit of analysis is the country at two points in time so the number of data base cases is 20. The five indicators of antisemitism are highly intercorrelated (all correlations are statistically significant at the $p = 0.05$ level or more), suggesting that the item about Jews talking too much about the Holocaust could be used along

Table 8.3 At the country level of analysis, the five indicators of antisemitic stereotyping are highly correlated with each other but are negatively correlated with the counts of violence for 2001 and 2003

	Own kind	Loyal to Israel	Shady	Too much power	Too much Shoah talk	Violence 2001 and 2003
Own kind	1	0.76 $p = .0001$	0.81 $p < .0001$	0.62 $p = 0.0037$	0.76 $p < .0001$	−0.29 $p = 0.217$
Loyal to Israel		1	0.55 $p = 0.012$	0.53 $p = 0.017$	0.68 $p = 0.001$	−0.44 $p = 0.053$
Shady			1	0.71 $p = 0.0004$	0.76 $p < .0001$	−0.34 $p = 0.138$
Too much power				1	0.63 $p = 0.003$	−0.14 $p = 0.566$
Too much Shoah talk					1	−0.20 $p = 0.407$
Violence 2001 & 2003						1

The number of cases is 20: 10 countries at two points in time. The correlations are Spearman rho (r_s) rank correlations. The probability (p) is the probability that a correlation that size could happen by chance. The antisemitism items are from the 2002 and 2004 surveys sponsored by the ADL.

with the others as a new indicator of antisemitic attitudes. All five items have negative correlations with the counts of violence, as does their reliable index (Cronbach's $\alpha = 0.89$)—a scale or index with an α of 0.70 or greater is usually thought to be sufficiently reliable. The antisemitism index's surprising negative correlation of -0.394 ($p = 0.085$) with violence suggests that: Countries exhibiting *higher* levels of antisemitic attitudes exhibit *lower* counts of anti-Jewish violence. Or, countries with *lower* levels of antisemitic attitudes exhibit *higher* counts of anti-Jewish violence.[13] However, as explicated later in this chapter, if a country's social structural predisposition toward violence and its level of support for the Palestinians are controlled, then the effect of its antisemitism on its count of violence is not statistically significant.

By classifying the countries according to the sizes of their Jewish and Muslim populations, Table 8.4 explores these relationships further. Across all five indicators of antisemitic attitudes, countries grouped as having many Muslims and many Jews have the lowest average score for these antisemitic attitudes but the highest counts of anti-Jewish violence. As measured here, antisemitic attitudes of ordinary Europeans are not the driving force of the violence during the period of the second intifada; the more pivotal contributing factors are attitudes about the participants in that conflict.[14]

Israel and the Palestinians

The 2002 and 2004 ADL surveys asked a number of questions about attitudes toward Israel and its conflict with the Palestine Authority. The following four indicators of disaffection from Israel form an additive index. Because this index has face validity but weak reliability ($\alpha = 0.26$), the multivariate analyses use the first item alone, and an index composed of the last two of the four items, which does have high reliability:

Palestinian Authority desires peace:
Please tell me whether you agree a lot, agree a little, disagree a little or disagree a lot with the following statement: "The Palestinian Authority truly wants to reach a peace agreement with Israel." (Agreement = "PAProPeace.")

Sympathizes with Palestinians:
Thinking specifically about the current conflict between Israel and the Palestinians, are your sympathies more with the Israelis or more with the Palestinians? (More with the Palestinians = "ProPalestine.")

Disagrees with Israel's stance on peace:
Please tell me whether you agree a lot, agree a little, disagree a little or disagree a lot with the following statement: "Israel truly wants to reach a peace agreement with the Palestinians." (Disagreement = "IsraelAntiPeace.")
Thinking generally about Israel, would you say that your views are very favorable, fairly favorable, neither favorable nor unfavorable, fairly unfavorable, or very unfavorable? (Very or fairly unfavorable = "ViewsIsraelUnfavorably.")

Table 8.4 Jewish and Muslim population classification, antisemitic attitudes, and anti-Jewish violence

Classification of country	Country	Jews don't care what happens to anyone but their own kind	Jews are more loyal to Israel than to this country	Jews use shady practices to get what they want	Jews have too much power in the business world	Jews still talk too much about what happened to them in the Holocaust	Average of five indicators of antisemitism for 2002 and 2004	Count of anti-Jewish violence for 2001–2003	Rate of violence per 10,000 Jews 2001–2003
Few Jews (0) and Few Muslims (0)	Austria	0.29	0.50	0.26	0.33	0.56	0.39	7	8.8
	Belgium	0.23	0.48	0.16	0.41	0.39	0.33	39	11.1
	Denmark	0.15	0.41	0.14	0.11	0.30	0.22	12	15.0
	Switzerland	0.32	0.48	0.24	0.36	0.52	0.38	14	7.8
	Average	0.25	0.47	0.20	0.30	0.44	0.33	18	10.4
Few Jews (0) and Many Muslims (1)	Italy	0.27	0.58	0.19	0.36	0.43	0.36	9	2.6
	Netherlands	0.15	0.46	0.09	0.19	0.35	0.25	8	2.7
	Spain	0.29	0.60	0.29	0.55	0.57	0.46	11	2.8
	Average	0.24	0.55	0.19	0.37	0.45	0.36	9	2.7
Many Jews (1) and Many Muslims (1)	France	0.18	0.35	0.16	0.38	0.41	0.29	154	2.4
	Germany	0.27	0.53	0.22	0.28	0.57	0.37	69	6.9
	UK	0.14	0.37	0.12	0.21	0.27	0.22	134	4.8
	Average	0.20	0.42	0.16	0.29	0.42	0.30	119	3.5
	Ten-country average	0.23	0.47	0.18	0.32	0.44	0.33	46	3.8

UK = United Kingdom. Counts of Anti-Jewish violence are from the website of the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism. The measures of antisemitic attitudes are averages of the data from the 2002 and 2004 Anti-Defamation League (ADL) reports. The data are proportions responding positively as indicated.

Table 8.5 groups the responses to these indicators and indexes of disaffection from Israel by the typology that classifies the countries according to the sizes of their Jewish and Muslim populations. It also relates these measures to the five-item antisemitism index and to the cumulative count of violence for 2001 through 2003. In general, countries with many Muslims and many Jews have among the lowest average scores for disaffection from Israel, disagreement with Israel's stance on peace, and anti-Semitism, but these countries have the highest counts of violence. However, the scores for PAProPeace show a different pattern: Of these items, PAProPeace has the most consistent effects on counts of violence and is the pivotal indicator of mobilization.[15] The two-item index of disagreement with Israel's stance on peace does not have strong effects.

Assumptions about the Counts of Violence

The ADL surveys coordinated to the earlier phases of the intifada were taken in June and September 2002 (Time 1) and in April 2004 (Time 2), creating a repeated measures longitudinal study in which these observations are separated by about two years. For these multivariate analyses, an analogous two-year time period was created by using the counts for 2001 (Time 1) and 2003 (Time 2) as response variables; the counts for 2004 and 2005 were not yet available. If the effects of anti-Jewish violence on antisemitic attitudes were being studied, then these data would be fine because the model would predict variables later in time (the attitudes) by variables earlier in time (the counts of violence). However, this chapter studies the effects of the attitudes on violence and this creates a problem concerning the time ordering of the variables. It therefore assumes that the attitudes are in equilibrium and hold for the year prior to the actual date of the taking of the surveys; that is, the attitudinal measures from the 2002 surveys hold for 2001 and the attitudinal measures from the 2004 surveys hold for 2003.[16]

As Table 8.6 indicates, this assumption is very reasonable for the unweighted data used in the main analysis because at the country level ($N = 10$ countries × 2 time points) the reductions in the levels of the indicators of antisemitic attitudes are not statistically significant from 2002 to 2004, and there was only one significant change in the indicators of disaffection from Israel during this time period—pro-Palestinian attitudes dropped off by 6 percentage points. The increase in the count of violence was about 8.7 from Time 1 to Time 2, but this statistically insignificant overall change masks a large increase among the countries with many Jews and many Muslims.

Table 8.6 also reports the significance levels when the data are frequency weighted by the $N = 500$ respondents for each survey; these tests of significance approximate those that would be obtained for respondent-level data. Because of the large N due to the frequency weighting, the effects become statistically significant, which is consistent with the ADL's (2004, 11) finding based on the responses of individual respondents that both antisemitic and pro-Israel attitudes tended to

Table 8.5 Jewish and Muslim population classification, indicators of disaffection from Israel, antisemitic attitudes, and anti-Jewish violence

Classification of country	Country	The Palestinian authority truly wants peace	Respondent sympathizes with Palestinianas	Israel does not want peace	Respondent views Israel unfavorably	Average of previous two items about Israel's stance	Average of four items about disaffection from Israel	Average of five antisemitism items	Count of anti-Jewish violence 2001–2003
Few Jews (0) and Few Muslims (0)	Austria	0.32	0.32	0.47	0.31	0.39	0.35	0.39	7
	Belgium	0.46	0.30	0.49	0.34	0.41	0.40	0.33	39
	Denmark	0.40	0.30	0.49	0.31	0.40	0.38	0.22	12
	Switzerland	0.38	0.35	0.52	0.40	0.46	0.41	0.38	14
	Average	0.39	0.32	0.49	0.34	0.42	0.38	0.33	18
Few Jews (0) and Many Muslims (1)	Italy	0.41	0.28	0.44	0.20	0.32	0.33	0.36	9
	Netherlands	0.31	0.27	0.52	0.39	0.46	0.37	0.25	8
	Spain	0.42	0.30	0.47	0.37	0.42	0.39	0.46	11
	Average	0.38	0.28	0.48	0.32	0.40	0.36	0.36	9
Many Jews (1) and Many Muslims (1)	France	0.48	0.23	0.41	0.29	0.35	0.35	0.29	154
	Germany	0.30	0.23	0.42	0.26	0.34	0.30	0.37	69
	UK	0.48	0.27	0.28	0.24	0.26	0.32	0.22	134
	Average	0.42	0.24	0.37	0.26	0.31	0.32	0.30	119
	Ten-country average	0.39	0.28	0.45	0.31	0.38	0.36	0.33	46

UK = United Kingdom. Counts of anti-Jewish violence are from the website of the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism. Measures of disaffection from Israel and antisemitic attitudes are averages of the data from the 2002 and 2004 Anti-Defamation League (ADL) reports. These data are proportions responding positively as indicated.

Table 8.6 The values of the indicators of antisemitic attitudes and of disaffection from Israel do not change much from 2002 to 2004

	Change from Time 1 to Time 2	Unweighted Data		Frequency weighted data	
		Significance		Significance	
Effect on:		*t*	Probability	*t*	Probability
Indicators of antisemitism					
Own kind	−0.02	−0.58	0.57	−13.6	<.0001
Loyal to Israel	−0.07	−1.60	0.13	−37.7	<.0001
Shady	−0.04	−1.24	0.23	−29.2	<.0001
Too much power	−0.08	−1.34	0.20	−31.6	<.0001
Too much Shoah talk	0.004	−0.08	0.94	−1.8	0.07
Five item index	−0.04	−1.10	0.29	−25.9	<.0001
Indicators of disaffection from Israel					
PA Pro peace	−0.03	−0.87	0.40	−20.4	<.0001
Pro Palestine	−0.06	−3.35	0.004	−78.8	<.0001
Israel anti peace	0.004	0.11	0.91	2.7	0.007
Views Israel unfavorably	0.06	1.79	0.09	42.2	<.0001
Disagrees with Israel's stance	0.03	0.99	0.33	23.4	<.0001
Four item index	−0.01	−0.44	0.66	−10.5	<.0001
Anti-Jewish violence					
Count of events	8.7	0.99	0.33	23.4	<.0001

Ordinary least-squares unstandardized regression coefficients for the differences between the values of the variables in the 2004 and 2002 surveys. The frequency weighted data approximate the significance levels for the unaggregated surveys of 500 respondents per country. These results corroborate the ADL's (2004, 11) assertion that antisemitism had significantly declined from 2002 to 2004. The Poisson regression analyses in this paper do not frequency or otherwise weight the data. The N is 20 in the analysis of the early phases of the intifada and 50 in the analysis of the whole time period.

decline from 2002 to 2004. The statistical analyses do not apply frequency weighting because the models of the unweighted data are more parsimonious and easier to interpret.

Table 8.7 examines the change in counts of violence from Time 1 to Time 2 for countries that are classified by the sizes of their Jewish and Muslim populations. All three groups of countries exhibit an increase in the count of violence; only Denmark and Spain exhibit a decline. Once again, countries with few Jews, regardless of the number of Muslims, experience much lower counts of violence than countries with many Jews and many Muslims. The total yearly count for the latter three countries increases from 80 to 156. Among the countries with few Jews and few Muslims, Belgium has the highest count but not the highest rate per 10,000 Jews. Austria and Denmark have low counts but have 6.3 violent incidents per 10,000 Jews, the highest rate of any of the ten countries. The grouping of the violence counts by the population typology may suggest that Muslim's are the perpetrators, but in some countries other people commit the violence.

Table 8.7 Anti-Jewish violence at Time 1 and Time 2

Classification of country	Country Jewish Population 2004	Count of anti-Jewish violence Time 1	Count of anti-Jewish violence Time 2	Increase in anti-Jewish violence Time 2 − Time 1	Count of anti-Jewish violence Time 1 + Time 2	Rate of violence per 10,000 Jews
Few Jews (0) and Few Muslims (0)	Austria 8,000	0	5	5	5	6.3
	Belgium 35,000	8	9	1	17	4.9
	Denmark 8,000	4	1	−3	5	6.3
	Switzerland 18,000	1	5	4	6	3.3
Sub total	69,000	13	20	7	33	4.8
Few Jews (0) and Many Muslims (1)	Italy 35,000	2	4	2	6	1.7
	Netherlands 30,000	2	6	4	8	2.7
	Spain 40,000	4	2	−2	6	1.5
Sub total	105,000	8	12	4	20	1.9
Many Jews (1) and Many Muslims (1)	France 650,000	27	71	44	98	1.5
	Germany 100,000	15	35	20	50	5.0
	UK 280,000	38	50	12	88	3.1
Sub total	1,030,000	80	156	76	236	2.3
Total	1,204,000	101	188	87	289	2.4

Counts of anti-Jewish violence are from the website of the Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism. Estimates of the Jewish populations are from various sources. UK = United Kingdom.

Who Are the Perpetrators?

Are Muslims the main perpetrators of the anti-Jewish violence during the period of the second intifada? Qualitative analyses by the European Union Monitoring Centre (EUMC) both substantiate the role of Muslim youths and enlarge the assignment of blame to include the extreme right and neo-fascist thugs (EUMC 2004a, 2004b, 2004c; Roth Institute 2006, 2007).

Regarding the countries with few Jews and few Muslims: The anti-Jewish violence in Austria is not attributable to Muslim youths or Muslim adults but to elements of the radical political right—young white men who are skinheads

(EUMC 2004a, 14). The violence in Belgium is primarily attributable to Muslim youths of North African origin (not of Turkish origin) who presumably are influenced by speeches in mosques during Friday prayers and then commit anti-Jewish offenses such as fire-bombing of Jewish property and physical assaults (EUMC 2004a, 14–15).[17] Anti-Israel propaganda from the extreme right and from the radical left may provide rationales for this violence.[18] A friend in Belgium has this to say (9/24/2004):

> Until now the violence against Jews in Belgium has come mainly if not solely from youngsters from Maghrebian origin. They are, this is my feeling, encouraged implicitly by leftists who join them in manifestations against "the war", where the distinction between anti-Sharon and anti-Jew becomes most of the time quite vague. Moreover, many Belgian intellectuals think and write that the Jewish Community, as a whole, should condemn Israeli policy

In Denmark, the perpetrators are primarily young males with Arabic-Palestinian-Muslim backgrounds rather than extreme right-wing white youths, as earlier (EUMC 2004a, 14). Because Switzerland is not a member of the European Union the EUMC has no data about its perpetrators.

Regarding the countries with few Jews and many Muslims: In Italy violence against Jews is not attributable to Arab Muslim youths or adults (EUMC 2004a, 15); instead, the extreme radical right is responsible for the very few violent incidents. Spain has the highest proportion (0.46) exhibiting antisemitic stereotypes and, for 2001–2003, low counts of violence (11)—the perpetrators are unknown. In the Netherlands, the violence mostly takes place in Amsterdam where most Jews live; 80% of the perpetrators were classified as "white" compared with 5% who were classified as ethnic minority youths from Islamic Moroccan circles (EUMC 2004a, 14).

Regarding countries where the Jewish and Muslim populations are the largest and the violence counts the highest: Muslim youths and adults have been identified as primary perpetrators of the violence, but violence also emanates from pro-Palestinian sympathizers and from the extreme radical right (Craig Smith, 8/13/2004). In Germany, most of the anti-Jewish offenses concern cemetery desecrations, incitements, and vicious propaganda—aggressive antisemitic letters, emails, and threatening phone calls—and not assaults on individuals (EUMC 2004d, 2; Stephen Roth Institute 2004, 2). The majority of the perpetrators have been males, between 15 and 24 years old, with low educational achievement—right-wing thugs, but violence from Muslim youths has increased (EUMC 2004a, 13 15).

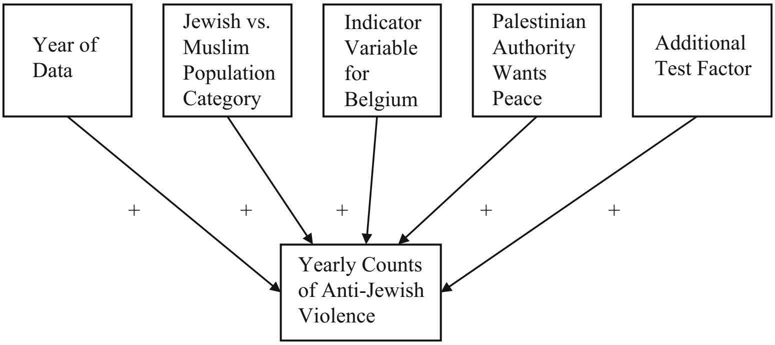
In France, Jews have been assaulted and insulted; synagogues, cemeteries, and Jewish property have been vandalized; and a school destroyed by fire (EUMC 2004d, 3). For 2001–2003, France has the highest count of anti-Jewish violence (154) but a very low proportion (0.29) of the country exhibiting antisemitic stereotypes. The perpetrators are young Muslims: violence attributable to the extreme right was 68% in 1994, 14% in 2001, and only 9% in 2002 (EUMC 2004a, 13).

In the UK, physical and verbal attacks against Jews by Muslims and Palestinian sympathizers have increased, as have attacks against synagogues and Jewish cemeteries (EUMC 2004a, 14; 2004d, 3). The UK has the second highest cumulative count of violence for 2001 through 2003 (134) but the lowest proportion with antisemitic attitudes (0.22).[19]

Developing and Testing Models

Study Design

These explorations suggest the following study design for the quantification of the effects of structural and attitudinal variables on the longitudinal counts of violence against Jews; diagrammatically, for each of the ten European countries:


Year of Data
Jewish vs. Muslim Population Category
Indicator Variable for Belgium
Palestinian Authority Wants Peace
Additional Test Factor
+
+
+
+
+
Yearly Counts of Anti-Jewish Violence

Each country has a yearly count of violence against Jews taken at two points in time in the main analysis and at five points in time in the replication. These yearly counts for a country are the level-1, or microlevel variables, which may vary from year to year of the second intifada. At the level-2 or macrolevel of analysis, the various countries have structural and attitudinal variables. The structural variables include the year of the observations, a typology of a country's Jewish and Muslim population sizes, and an indicator variable for Belgium, which has a much higher than expected count of violence. These characteristics create a societal predisposition toward the counts of violence.

However, a model that comprises only these structural aspects of a country does not fit the data very closely. By adding to this model an aggregated attitudinal variable, a cultural variable, about the Palestinian's desire for peace, the mobilization model is produced. This model combines the cultural stimulus and the societal predisposition to produce a close fit to the data. This model is tested by adding to it one at a time these additional country-level test factors: anti-Israel sentiments, antisemitism, human and economic development, deportations during the Holocaust, and reparations for the Holocaust. None of these additional test factors have statistically significant effects compared with the components of the mobilization model.

Missing from these descriptive macrolevel models are explicitly measured intervening variables that specify the explanatory microlevel agents and processes that link up a country's structural predisposition and the attitudinal stimulus to the country's violent response. This chapter develops plausible intervening microlevel processes by assessing the performance of the mobilization model compared with two other theoretical models that focus on factors that allow the violence to take place: the unresponsive bystander model and the cognitive ambivalence model. Attitudinal measures of these variables are added to the complete structural model one at a time and the goodness of fit of these models are compared to that for the mobilization model. Mobilization of perpetrators by the events in the Middle East and the cognitive ambivalence of ordinary Europeans provide better explanations of the violence than bystander unresponsiveness, given the limitations of these data and the multilevel Poisson regression models that produces the estimates and the goodness-of-fit statistics.

The Poisson Model

To model the counts of violence, the multilevel Poisson regression model is an appropriate first choice (Agresti 1996, 4–6, 80–97; Littell et al. 1996, 453–460; 2006, 557–566; Raudenbush and Bryk 2002, 309–317): for each country, the counts may range theoretically from 0 to positive infinity and the imagery of counts of violent anti-Jewish events in a country for a time period evokes analogous models of counts of murders, accidents, riots, and bombs hitting targets in which statisticians have applied the Poisson sampling distribution (Coleman 1964, 114–115; McCullagh and Nedler 1989, 204–208).

Table 8.8 presents the logic of the pivotal multilevel statistical model.[20] Its levels are defined in ascending alphabetical order: the lowest level—the time of the observation—by *i*; the country, by *j*; and the population typology—the nesting attribute—by *k*. For the main analysis, the fixed explanatory components include a time period variable (YearDummy) coded 1 for 2004 and 0 for 2002; a cross-sectional difference (ManyJewsMuslims) coded 1 for a country that has many Jews and many Muslims and coded 0 for all of the other countries; an indicator variable coded 1 for Belgium and 0 for all other countries (not Belgium); and, in addition to these structural characteristics, aggregated attitudinal variables indicative of sentiments that are, respectively, pro-Palestinian, pro-Jewish, or both pro-Palestinian and pro-Jewish.

In order for SAS to calculate the least-squares means, the binary indicators need to be treated as classification variables. For such "class" variables SAS uses the highest categorical code as the base category rather than the lowest categorical code. It then quantifies the effect of the 0 code relative to the 1 code. Thus, in the subsequent tables, the reported effects for dichotomous variables have a minus sign when the effect for the category coded 0 is less than the effect for the category coded 1; simply change the sign of the coefficient to obtain the effect of the category coded 1.

There are two covariance parameters: the random effect $d_{j(k)}$ between different countries when they are classified by their population predisposition toward violence and, when the countries are thus classified, the residual variance for different countries in different years e_{ijk} (this is the same quantity as the extra-dispersion scale). The analysis aims to uncover the factors that reduce the first variance component to insignificance and the factors that reduce the residual dispersion.

Table 8.8 A statistical model extimated in the main analysis

The subjects are the ten countries surveyed by the Anti-Defamation League (ADL) in 2002 and 2004; these are classified by their Jewish and Muslim population predisposition category. Categories of binary indicators are coded 1 or 0 with the former the base category. The intercept and the two random effects compose the initial model. In an elaborated model, the intercept, four explanatory properties of countries, and two random effects determine the logarithm of the country's count of antisemitic incidents. This statistical model is:

$$\eta_{\mathrm{ijk}} = \log\left[\lambda_{ijk}\right] = \mu + \tau_k + \beta_{1i} + \beta_2 + \beta_3 + \mathrm{d}_{j(i)} + \mathrm{e}_{ijk} \tag{1}$$

where:

λ_{ijk} is the conditional mean count of antisemitic violent incidents for a country *j* in population predisposition category *i* in time period *k* given the structural variables and the random effects:

μ is the intercept term;

τ_k is the time period (YearDummy), *k* = *1 is 2004, k* = *0 is 2002*;

β_{1i} is the fixed parameter for a country having many Jews and Muslims (ManyJewsMuslims, i = Yes = 1, i = No = 0);

β_2 is the fixed parameter for Belgium (Belgium = 1, Not Belgium = 0);

β_3 is the fixed parameter for a country's relevant attitudinal proportion (PAProPeace, Values Jewish Lives, or Values Jewish and Values Palestinian Lives);

$\mathrm{d}_{j(i)}$ is the random effect associated with the j^{th} country grouped by population predisposition category *i*, $\mathrm{d}_{j(i)} \sim iid\,N(0,\sigma^2_{j(i)})$;

e_{ijk} is the residual random effect associated with the j^{th} country in population predisposition category *i* at time period *k*, $\mathrm{e}_{ijk} \sim iid\,N(0,\sigma^2_{ijk})$. This is the extra-dispersion scale factor.

In words, the countries are classified as having many Jews and Muslims or not. Equation (1) states that the natural logarithm of the count of antisemitic incidents for a country thus classified at a time point depends upon the intercept level of antisemitic violent incidents, plus the effect of the time period, plus the effect of having many Jews and many Muslims, plus the effect of the proportion holding the relevant attitude, plus the effect of Belgium, plus the random effect associated with a classified country, plus the random effect associated with a classified country at a particular time.

These instructions for the GLIMMIX procedure specify a multilevel model with a Poisson error and a log link. The method is residual pseudo-likelihood (i.e., rspl). The information criteria ic = pq requests that the penalties include the number of fixed effects parameters. The random _residual_ statement adds the multiplicative overdispersion parameter. The covtest statement tests the significance of the level-2 covariance parameter. The SAS code is:

```
Title1 ‘This run produces estimates for the basic model’;
proc glimmix data = Adlten method = rspl ic = pq;
   class country year ManyJewsMuslims Belgium;
   model violence0103 = year ManyJewsMuslims Belgium PAProPeace/
   link = log dist = poisson s;
   random country(manyjewsmuslims)/s;
   random _residual_/s;
   covtest ‘zerog = no G-side covariance’ zerog / cl Wald estimates;
run;
```

If $d_{j(k)}$ is reduced to insignificance when the fixed covariates are sequentially introduced, then these factors will have clarified why these ten countries have different levels of anti-Jewish violence.

Fit Statistics

SAS's Glimmix macro and its Glimmix procedure for estimating generalized linear mixed models use the pseudo-likelihood techniques of Wolfinger and O'Connell (1993) described in Littell et al. (1996, 433–437; 2006, 538–542) and in the Glimmix users manual (SAS Institute 2005, 116–118); Proc Glimmix estimated the models in this chapter using its default residual pseudo-likelihood (RSPL) algorithm. Although Glimmix computes values for "Fit Statistics"—the AIC, AICC, BIC, CAIC, and HQIC—that analysts often use to test the goodness of fit of models, the SAS Institute does not recommend using these statistics to test models estimated through pseudo-likelihoods.[21] Consequently, in the tables that follow this chapter reports these measures for assessing model fit: an R^2 analog at level-2; an R^2 analog at level-1; the deviance, the scaled deviance, the pseudo -2 × residual log likelihood, and the pseudo-BIC.

R^2 analogs

When multilevel models do not have random slope coefficients for the fixed covariates, Kreft and DeLeeuw (1998, 116–119) suggest using the scaled difference between the variance components of the null and new models as a way to calculate an R^2 analog statistic for each level of the model. In a two-level model, the null random intercept model has two covariance parameters (one variance component at each level), and there are no fixed explanatory variables. If new fixed explanatory variables are added to the null model, then new values of the two variance components will result. Consequently, there will be two R^2 analogs, $\hat{R}_B^2$ based on the between-country variance components and $\hat{R}_W^2$ based on the within-country variance component. To calculate $\hat{R}_B^2$, one takes the level-2 variance component of the new model and subtracts it from the level-2 variance component of the null model and then divides this difference by the level-2 variance component of the null model. To calculate $\hat{R}_W^2$, the same logic is followed but using the level-1 variance components. (These are the extra-dispersion parameters that also are used to scale the deviances and to correct the standard errors.) For example, below in Table 8.9 for Model 5 the level-2 $\hat{R}_B^2$ is 1 ((1.34 − 0)/1.34) compared with 0.96 ((1.34 −0.06)/1.34) for Model 4; for Model 5 the $\hat{R}_W^2$ is 0.65 compared with 0.63 for Model 4—both R^2s give a very slight edge to Model 5 over Model 4.

Although these R^2 analogs quantify the proportion reduction in unexplained variance they have some rather obvious limitations: the calculations may produce a negative sign, in this chapter the variance components ($d_{j(k)}$ and e_{ijk}) are on the natural log scale, the logic does not apply when there are random slopes, tests for significant differences between these R^2s if such tests exist are not well-known, and Snijders and Bosker (1994) suggest using in the calculations of $\hat{R}_B^2$ the total between-variance rather than the level-2 variance component.

Table 8.9 The elaboration and testing of the substantive effects toward the mobilization Poisson regression model

Models	Model 1, Intercept only	Model 2, Intercept and year	Model 3, typology added	Model 4, Belgium added	Model 5, Mobilization added	Model 6, anti-Israel added	Model 7, antisemitism added
Covariance parameters							
$d_{j(k)}$	1.34	1.43	0.096	0.057	0	0	0.003
z	1.84	1.98	0.98	1.04	–	–	0.13
Pr(z)	0.033	0.024	0.163	0.148	–	–	0.45
$e_{\boldsymbol{ijk}}$	3.81	1.71	1.60	1.39	1.32	1.4	1.31
z	2.35	2.19	2.37	2.56	2.74	2.65	2.51
Pr(z)	0.009	0.014	0.009	0.005	0.003	0.004	0.006
Ratio of $d_{j(k)}$ to $e_{\boldsymbol{ijk}}$	0.35 to 1	0.84 to 1	0.06 to 1	0.041 to 1	0 to 1	0 to 1	0.002 to 1
Fit statistics							
Probability that $e_{\boldsymbol{ijk}} = 0$ fits data	0.0003	< .0001	0.0444	0.0366	1	1	1
Generalized chi square for model	72.3	30.8	27.3	22.2	19.9	19.6	18.3
Degrees of freedom of model	20	18	17	16	15	14	14
Level-2 R^2 analog	0.00	−0.06	0.93	0.96	1.00	1.00	1.00
Level-1 R^2 analog	0.00	0.55	0.58	0.63	0.65	0.63	0.66
Deviance	50.0	20.1	24.3	20.9	22.2	22.0	20.4
Scaled deviance	13.1	11.7	15.1	15.1	16.8	15.7	15.6
−2 × Residual log pseudo-likelihood	59.9	54.1	36.1	32.2	24.9	21.8	21.5
Pseudo-BIC, IC = q	64.5	58.7	40.7	36.9	27.2	24.1	26.1
Pseudo-BIC, IC = pq	68.7	65.7	50.3	48.9	41.2	40.3	42.6

(continued)

Table 8.9 (continued)

Models	Model 1, Intercept only	Model 2, Intercept and year	Model 3, typology added	Model 4, Belgium added	Model 5, Mobilization added	Model 6, anti-Israel added	Model 7, antisemitism added
The fixed covariates							
Intercept	2.12	2.3	3.91	4.96	3.42	3.13	4.35
t	5.23	5.72	19.15	10.71	5.7	3.14	4
Pr > *t*	0.0005	0.0003	< .0001	< .0001	0.0007	0.016	0.005
Year is 2002	–	−0.62	−0.62	−0.62	−0.68	−0.67	−0.63
t	–	−3.85	−3.98	−4.28	−4.75	−4.44	−4.22
Pr > *t*	–	0.0039	0.0032	0.0021	0.0014	0.003	0.0039
Country has {Not many Jews and Muslims}	–	–	−2.33	−2.55	−2.40	−2.45	−2.38
t	–	–	−8.09	−9.44	−11.18	−9.23	−10.84
Pr > *t*	–	–	< .0001	< .0001	< .0001	< .0001	< .0001
Country is {Not Belgium}	–	–	–	−1.04	−0.76	−0.74	−0.84
t	–	–	–	−2.41	−2.19	−2.05	−2.31
Pr > *t*	–	–	–	0.04	0.060	0.080	0.054
Believing PA is ProPeace	–	–	–	–	3.04	3.18	2.07
t	–	–	–	–	3.23	3.06	1.54
Pr > *t*	–	–	–	–	0.012	0.018	0.167
Additional test factor	–	–	–	–	–	0.664	−1.60
t	–	–	–	–	–	0.380	−1.04
Pr > *t*	–	–	–	–	–	0.718	0.33
						Not Sig.	Not Sig.

Notes: e_{ijk} equals the generalized chi square divided by degress of freedom. Models 1 through 7 use the counts of violence for 2001 and 2003. A country's disagreement with Israel's stance on peace is assessed by averaging the proportion who say that Israel truly does not want a peace agreement with the proportion viewing Israel unfavorably. The index of conventional antisemitic attitudes averages the values of the five indicators of antisemitic stereotypes.

Deviance and scaled deviance

To test the relative goodness of fit of models based on quasi-likelihoods, the SAS Institute recommends using the difference in the deviances and scaled deviances between models; significantly smaller values indicate a better fit (Littell et al.1996, 432, 445–446; 2006, 536). The deviance is the difference in quasi-likelihood for the full data (i.e., the saturated model) and the model being tested. The deviance is approximately χ^2 distributed with N-p degrees of freedom where N is the number of observations and p is the number of fixed effect parameters. For example, in Table 8.9 the deviance for Model 5 is 22.2 and that for Model 4 is 20.9, for an increase of 1.3 in deviance for an expenditure of 1 degree of freedom. Using the χ^2 test, this difference in deviance between the models is not statistically significant (p ~ 0.25), but it favors Model 4.

Poisson models are often fit to overdispersed data, having a sample variance that is larger than the sample mean. Glimmix can use the extra-dispersion parameter to correct the tests of significance for overdispersion or underdispersion, and to scale the deviance. Models with dispersion parameters very close to 1 are preferred to models whose dispersion parameters differ considerably from 1. For example, Model 5 is less overdispersed than Model 4, 1.32 compared with 1.39, but this difference of 0.07 no doubt is not statistically significant. The scaled deviance is the quotient when the deviance is divided by the extra-dispersion scale (which is identical to the level-1 covariance parameter for these Poisson models.) The scaled deviance for Model 5 is 16.8 and that for Model 4 is 15.1, their difference indicates a nonsignificant increase of 1.7 (χ^2 test p ~ 0.19). Consequently, both the deviance and scaled deviance give a very slight edge to Model 4 over Model 5.

$-2 \times$ Residual log pseudo-likelihood and pseudo-BIC

The SAS Institute is very cautious about comparing the values of the residual log pseudo-likelihood across different statistical models, even if such models are nested. The Glimmix manual states (2005, 163): "It is possible that between two nested models the larger model [i.e., the model with fewer free parameters] has a smaller pseudo-likelihood. For this reason, IC = none is the default for GLMMs fit by pseudo-likelihood methods." However, inspection of the $-2 \times$ residual log pseudo-likelihoods in Table 8.9 below shows that these $-2 \times$ log pseudo-likelihoods decrease in size monotonically with no reversals across the five successive models, as they should. Moreover, Models 6 and 7, each of which uses one more parameter than Model 5, have the smaller values of $-2 \times$ log pseudo-likelihood. This monotonic decline in $-2 \times$ log pseudo-likelihoods when the degrees of freedom are successively reduced provides a rationale for the use of fit statistics like Schwarz's (1978) Bayesian Information Criterion (BIC) as calculated under the IC = pq option. This option requests that the fit statistics such as the BIC be calculated so that models that use more parameters are penalized for their lack of parsimony. The computational formula Glimmix uses (2005, 22) is pseudo-BIC $= -2l + d \times \log n$ where $d =$ the dimension of the model, n is the size of the data, and $-2l$ is the value of the $-2 \times$ residual log pseudo-likelihood. The subsequent analyses tend to prefer models with smaller values of pseudo-BIC to those with larger values.[22]

Creating the Mobilization Model

Table 8.9 presents the estimates for the sequentially elaborated models that lead to the mobilization model. This model includes social structural aspects of the countries and this indicator of mobilization: the belief that the Palestinian Authority truly wants to reach a peace agreement with Israel (PAProPeace)—this sentiment brings the conflict in the Middle East to the streets of Europe. In all of these models the countries are classified as having many Jews and many Muslims (1) or as not (0); the categories of this classification predispose the country toward its characteristic level of violent events. Model 1, which includes the random intercept and the two variance components, provides benchmarks for assessing the subsequent models in which the fixed variables are introduced sequentially, one at a time. Model 2 adds the indicator for the year of the survey, followed by Model 3 that adds the cross-sectional difference for the categories of the Jewish and Muslim population size; it thus quantifies the effects of two predisposing conditions—year of intifada and Jewish and Muslim population size. Logically, the next model would add the interaction between this cross-sectional difference and the time period, but this effect is not statistically significant ($p = 0.55$) so this model is not reported—a significant effect would imply that violence increased disproportionately in countries with many Jews and many Muslims. Model 4 completes the social structural model; it adds an indicator variable for Belgium because its count of violence is much higher than its population predisposition would suggest.[23] Model 5, the mobilization model, introduces an "exciting cause" or stimulus variable, namely the belief that the Palestinian Authority truly wants to reach a peace agreement with Israel. This model is tested by adding to it these additional indicators of mobilization: an index of disagreement with Israel's stance on peace (see Model 6) and the five-item measure of conventional antisemitic attitudes (see Model 7). (Additional country-level measures—human and economic development, deportations of Jews during the Holocaust, and reparations to victims of the Holocaust—have little effect; see endnote 25.)

Model 1, the baseline

This null model provides a baseline for assessing the improvement of the subsequent models that include the various substantive variables that may or may not be associated with the level of violent events. Its two variance components provide information that is needed to calculate the intraclass correlation coefficient, which quantifies the proportion of variance due to differences between the countries, and the reliability of the sample mean for each country (Bryk and Raudenbush 1992, 62–63). The intraclass correlation is 0.26—there is a lot of variability between countries to explain—and the average reliability of the sample means based on the 1,000 survey respondents per country is 0.99, very reliable. This model does not closely fit the data: the extra-dispersion scale is 3.81, the deviance is 50, and the pseudo-BIC is 68.7; these values are higher than those for the other models. Each model through Model 5 incrementally reduces the over-dispersion scale and the pseudo-BIC, but sometimes not the deviance.

Model 2, adding the time period

The addition of the time period to Model 1 does not reduce the variability between countries (it increases slightly from 1.34 to 1.43—the $\hat{R}_B^2 = -0.06!$) but this model does reduce the overdispersion, from 3.81 to 1.71 ($\hat{R}_W^2 = 0.55$). Its parameters indicate that violence increased from Time 1 to Time 2, the coefficient is +0.62 ($p = 0.004$) on the logarithm scale. This model reduces the pseudo-BIC from 68.7 to 65.7 and the deviance from 50 to 20.1; Model 2 fits better than Model 1 but there is much variability left to explain.

Model 3, the predisposing factors

The introduction of the population typology along with the time period as fixed covariates improves the explanation of the violence. Compared with the baseline, the variance between countries is only 0.09 (compared with 1.34) and the residual variance is 1.6 (compared with 3.81). The fit statistics reflect this improvement: $\hat{R}_B^2 = 0.93$ and $\hat{R}_W^2 = 0.58$, the deviance of the model is cut by half from that of Model 1, and the pseudo-BIC is considerably less (50.3 compared with 68.7). The substantive parameters on the logarithm scale clearly indicate that violence has increased from Time 1 to Time 2 (+0.62, $p = 0.003$) and that countries with many Jews and many Muslims have the highest counts of violence (+ 2.33, $p < .0001$).

Model 3 has higher deviance (24.3 compared with 20.1 for Model 2) because Belgium has higher rates of violence than it population category suggests—it is classified as having few Jews and few Muslims but it has relatively high counts of violence. If Belgium is mainly responsible for the increased deviance of Model 3, then inspection of the residuals (the difference between actual and predicted values) should indicate that Belgium has higher values than other countries and it does. Model 3 underestimates the violence in Belgium: at Time 1 the actual count is 8 events whereas the predicted count is 3.7 (2, 7.3) and at Time 2 the actual count is 9 events whereas the predicted count is 7 (3.7, 13.2). The Time 1 raw residual of 4.2 and the chi square adjusted residual of 1.7 are larger than any other residuals. This under-prediction of the violence in Belgium contributes to the larger value of the deviance of Model 3; the next model adds an indicator variable for Belgium that reduces this dispersion.

Model 4, the full social structural model

The indicator for Belgium completes the social structural model; it reduces the variability between countries from 0.09 to 0.06, thereby improving the $\hat{R}_B^2$ from 0.93 to 0.96, and it reduces the extra-dispersion estimate from 1.60 to 1.39, thereby improving the $\hat{R}_W^2$ from 0.58 to 0.63. Model 4's deviance of 20.9 is lower than that of Model 1 (50) and Model 3 (24.3), and its pseudo-BIC of 48.9 is the lowest value thus far. The estimated effects on the log scale indicate that countries with many Jews and many Muslims have higher violence, by 2.55 ($p < .0001$); a country at Time 2 has higher violence than it does at Time 1, by 0.62 ($p < .0021$); and Belgium has higher violence, by 1.04 ($p = 0.04$), when compared with all other countries grouped together and all of the previous factors controlled.[24]

Model 5, adding pro-Palestinian mobilization

A country's attitudes about the conflict in the Middle East can be expected to influence its count of violent events against Jews. More specifically, a country's level of belief that the Palestinian Authority truly wants peace with Israel—a key pro-Palestinian attitude—may foretell its level of violence, especially since its late leader Yasir Arafat seems not to have wanted peace (Roth 2004). When this indicator is added, the resulting model fits the data better than any model considered thus far: the variability between the countries disappears ($\hat{R}_B^2 = 1$); the extra-dispersion estimate is further reduced to 1.32 ($\hat{R}_W^2 = 0.65$), and the pseudo-BIC is 41.2, lower than that of any previous model. But its deviance of 22.2, which is much lower than that of the baseline value of 50, is higher than that of Model 2 (20.1) and Model 4 (20.9).

The effect sizes on the logarithm scale suggest that PAProPeace has the largest effect: a unit increase from 0 to 1 increases the log of the violence count by 3.04 ($p = 0.012$); the indicator for countries with many Jews and Muslims, compared with those with few Jews regardless of the number of Muslims, increases the log of the violence count by 2.40 ($p < .0001$); and the change from Time 1 to Time 2 increases the log of the violence count by 0.68 ($p = 0.0014$). The intercept is 3.42 ($p = 0.0007$). These effects on the log scale translate into the following least-squares means (and their lower and upper values) on the scale of the raw counts of violence:

> Time 1 = 10.6 (7, 16.1); Time 2 = 20.9 (14.5, 30.1); Difference = + 10.3 incidents.
> Not ManyJewsMuslims = 4.5 (3, 6.7); Many = 49.4 (30.8, 79.1); Difference = + 44.9.
> Not Belgium = 10.2 (8, 13); Belgium = 21.8 (10.6, 44.8); Difference = + 11.6.

For Belgium at Time 1 and at Time 2, the proportion saying the Palestinian Authority truly wants peace is 0.46, a fairly high proportion; its predicted count of violence for Time 1 is 5.7 (the actual count is 8) and for Time 2 it is 11.3 (the actual count is 9). In Germany the proportion saying the Palestinian Authority truly wants peace is 0.33 at Time 1 and 0.27 at Time 2. If Belgium had these proportions, then the model predicts a drop in violence as support for the Palestinians drops: the predicted counts would now be 3.9 and 6.3, which are lower than the actual counts by 4 and 3 violent incidents, respectively.

Model 6, adding disagreement with Israel's stance on peace

The index of disagreement with Israel's stance on peace (a component of disaffection from Israel) is the simple average of a country's proportion saying that Israel truly does not want peace and its proportion viewing Israel unfavorably. These two items have a Spearman rank correlation of $r_s = 0.68$ ($p = 0.001$) and compose an index that is reliable ($\alpha = 0.78$) and normally distributed (Shapiro–Wilk test, $p <$ W is 0.98). Model 6 adds this index to the mobilization model but its positive effect on counts of violence is not statistically significant ($p = 0.72$). Consequently, this variable does not threaten the statistical conclusion validity of the mobilization model.

Model 7, adding antisemitic attitudes

The index of antisemitic attitudes is the simple average of the five indicators of antisemitic stereotypes: a country's proportions on "own kind," "loyal to Israel," "shady practices," "too much power," and "too much talk about the Holocaust." As noted earlier, the five items are significantly inter-correlated and compose an index that is reliable ($\alpha = 0.89$) and normally distributed (Shapiro–Wilk test, $p <$ W is 0.29). Model 7 adds this index of antisemitism to the mobilization model; paradoxically, countries with *higher* proportions on this index have *lower* counts of antisemitic violence, but this effect is not statistically significant. Additional analyses show that the negative sign of the effect of antisemitic stereotypes on violence holds for each of the five indicators and their dimensions—Jewish in-group loyalty, Jewish corruption, and too much talk about the Holocaust. Apparently, in the context of the second intifada against Israel, a country's belief that the Palestine Authority truly desires peace (and not its disagreement with Israel's stance on peace and not its level of antisemitic stereotyping) combines with the Jewish and Muslim population predisposition to intensify the anti-Jewish violence.[25] This model can be accepted as an explanation for the mobilization of the perpetrators of the anti-Jewish violence, but it does not specify the factors that allow the violence to take place.

What Factors Foster the Violence?

Why do France, the UK, and Germany have high counts of anti-Jewish violence and low proportions exhibiting antisemitic beliefs? If ordinary people in these three countries are not strongly antisemitic, why do they not more aggressively express outrage at the anti-Jewish violence that takes place in their cities? Perhaps these ordinary Europeans are similar to unresponsive bystanders—they may like Jews personally and have Jewish friends, but they do little to stop the anti-Jewish violence, which they know about through first-hand observations, reports in newspapers, television news, and contacts with Jewish friends. They may believe that stopping the violence is not their responsibility—the police will address this problem. Moreover, their ambivalence about the conflict in the Middle East may dampen ameliorative actions.

Unresponsive bystanders

The bystanders who observed the multiple stabbings of Catherine Genovese did not dislike her personally, but they did nothing to stop the murderer because they thought others would intervene, and it was not in their direct interest to become involved. Ordinary Europeans may be similar to these bystanders because they are aware of the victimizations, they do not dislike the victims, and they do not act to stop the perpetrators. The hypothesized effect of bystander unresponsiveness can be tested by adding to the structural model a reliable index ($\alpha = 0.70$) of valuing Jewish lives, which is composed of honoring the Holocaust and stating that Israel truly wants peace. Although this index says nothing directly about living European Jews—it taps attitudes toward murdered European Jews and living Israeli Jews—for these ten countries it is strongly negatively correlated with the five-item measure of antisemitic stereotypes (-0.786, $p < .0001$) and negatively correlated with dislike of Israel (-0.498, $p = 0.026$). Moreover, the three countries with many Jews and many Muslims have a score of 0.50 on this index compared with the score of 0.45 for the other countries.

Countries that value Jewish lives may not exhibit personal animosity toward Jews, but these countries exhibit higher counts of violence, suggesting that bystanders—the general public—are unresponsive; see Table 8.10, which for ease of comparison also reports the baseline model and Model 5. Both Model 5, the mobilization model, and Model 8, the bystander unresponsiveness model, fit the data about equally as well and have appropriate theoretical rationales. Even though Model 5 explains all of the variability that is between countries, its $\hat{R}_B^2 = 1$ is only very slightly more favorable than the $\hat{R}_B^2 = 0.98$ of Model 8. Model 8 has some favorable fit statistics: its dispersion is smaller than that for Model 5, producing a noticeably larger value of $\hat{R}_W^2$, 0.70 compared with 0.65; its scaled deviance is smaller, 15.1 compared with 16.8; and its deviance is significantly smaller, 17.5 compared with 22.2 (delta = 4.7, chi-square test with 1 degree of freedom, $p \sim 0.03$). But Model 5 has the smaller, more favorable pseudo-likelihood and a noticeably smaller pseudo-BIC, 41.2 compared with 44.3.

Apparently, the mobilization of the perpetrators and bystander unresponsiveness both contribute to the anti-Jewish violence in Europe. Model 5 directly taps pro-Palestinian attitudes toward the conflict in the Middle East; it assumes that valuing Palestinian lives stimulates Palestinian sympathizers some of whom become

Table 8.10 Three models that aim to explain why anti-Jewish violence is prevalent in Europe: mobilization, unresponsiveness, and ambivalence

Models	Model 1, Intercept only	Model 5, Mobilization model	Model 8, Unresponsive bystanders	Model 9, Cognitive ambivalence
Covariance parameters				
$d_{j(k)}$	1.34	0	0.028	0.017
z	1.84	–	0.76	0.57
Pr(z)	0.033	–	0.225	0.283
e_{ijk}	3.81	1.32	1.16	1.17
z	2.35	2.74	2.57	2.60
Pr(z)	0.009	0.003	0.005	0.005
Ratio of $d_{j(k)}$ to e_{ijk}	0.35 to 1	0 to 1	0.024 to 1	0.014 to 1
Fit statistics				
Probability that $e_{ijk} = 0$ fits data	0.0003	1	0.1007	0.1977
Generalized chi square for model	72.3	19.9	17.4	17.6
Degrees of Freedom of model	20	16	15	15
Level-2 R^2 analog	0.00	1.00	0.98	0.99
Level-1 R^2 analog	0.00	0.65	0.70	0.69
Deviance	50.0	22.2	17.5	18.6
Scaled Deviance	13.1	16.8	15.1	15.8
$-2 \times$ Residual log pseudo-likelihood	59.9	24.9	25.4	24.6
Pseudo-BIC, IC = pq	68.7	41.2	44.3	43.6
The fixed covariates				
Intercept	2.12	3.42	3.48	3.37
t	5.23	5.7	4.89	4.88
Pr > t	0.0005	0.0007	0.001	0.002
Year is 2002	–	−0.68	−0.70	−0.69
t	–	−4.75	−5.09	−5.08
Pr > t	–	0.0014	0.001	0.001
Country has {Not many Jews and Muslims}	–	−2.40	−2.38	−2.37
t	–	−11.18	−10.01	−10.45
Pr > t	–	< .0001	< .0001	< .0001
Country is {Not Belgium}	–	−0.76	−0.93	−0.87
t	–	−2.19	−2.54	−2.45
Pr > t	–	0.060	0.035	0.04
Effect of aggregated attitudinal variable	–	3.04	2.76	3.00
t	–	3.23	2.48	2.73
Pr > t	–	0.012	0.038	0.026

Notes: e_{ijk} equals the generalized chi square divided by the degrees of freedom. All models use the counts of violence for 2001 and 2003. "Remembrance" is "Too Much Shoah Talk" = 1. "Values Jewish Lives" is the average of "Israel Desires Peace" and "Remembrance." "Values Both Jewish and Palestinian Lives" is the average of "PAProPeace," "IsraelProPeace," and "Remembrance."

mobilized to attack Jews. Model 8 assumes that many Europeans do not express antisemitic stereotypes, they honor the Jewish dead of the Holocaust, and they believe that Israel desires peace in the Middle East. But they are in effect unresponsive bystanders who do not actively intervene to protest or to stop the violence. They may believe that they are not responsible for the violence, other people will address this problem, and their self-interest is dissonant with their becoming involved.

Dissonant cognitive forces

Ordinary Europeans may value Palestinian lives as well as Jewish lives and this cross-pressure could dampen ameliorative actions creating bystander unresponsiveness. The proportion of a country stating that the Palestinian Authority truly wants peace can be viewed as an indicator of the value people place on Palestinian lives (the higher the proportion, the greater the value). The three countries with high counts of violence, compared with the other countries, have on average a higher proportion who value Palestinian lives, 0.42 compared with 0.38. Countries that place a high value on Palestinian lives also place a high value on Jewish lives, Spearman's $r_s = 0.56$ ($p = 0.01$), and this fact provides a basis for this theorizing:

The conflict in the Middle East between Israel and the Palestinian Authority creates ambivalence in Europeans that can be conceptualized as two opposing cognitive forces (Lewin [1951] 1997; Verba 1961, 226–228), which is a root conception of cognitive dissonance theory (Festinger 1957, 12–24): One force moves people to support the Palestinians; the other force moves people to support Israel. In periods when there is little conflict between Israel and the Palestinians, these opposing cognitive forces are relatively small and the tension level, viewed simply as the sum of these opposing forces, is minimal (this tension level is indicative of the level of cognitive dissonance). Assume that a suicide bomber strikes Israel killing and maiming many people. The reports of this act by the news media creates sympathy for the Israelis and increases the value of the force moving people to support Israel; the tension level (i.e., dissonance) has increased, however. Then, the news media report that Israel retaliates by destroying the homes of the families of the suicide bomber and by tightening its control over the territories.[26] This retaliation increases the value of the force moving people to support the Palestinians and, in combination with the high level of the force supporting Israel, creates a very high level of tension (and of cognitive dissonance). At this point, violence against Jews in Europe may intensify because Muslim youths are mobilized to seek revenge for Israel's actions in the territories, as they perceive them.

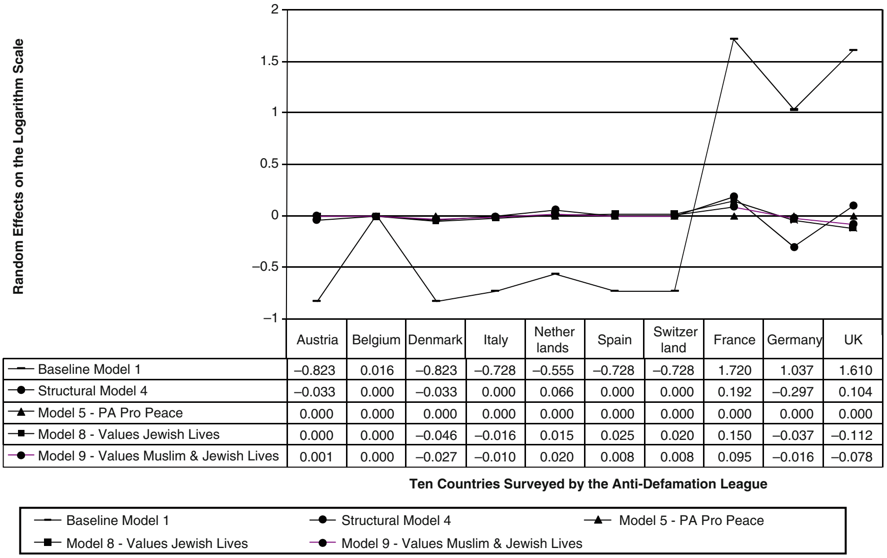
Ordinary Europeans may become ambivalent at a high level of dissonance because they value both Palestinian and Jewish lives and peace in the Middle East and, consequently, they do very little to stop the violence against the Jewish communities in Europe. If they actively supported the Palestinians by protesting Israeli actions, then this would add to their dissonance because they also value Jewish lives. If they actively protested the anti-Jewish violence, then this would also add to their dissonance because they also value Palestinian lives. The dissonance engendered by doing nothing to stop the anti-Jewish violence is less than the dissonance that would result by their supporting either the Palestinians or defending their Jewish neighbors, so ordinary Europeans do little to stop the violence.

Ambivalent people may say a plague on both your houses and withdraw. A public opinion poll found that the French disapproved of both Prime Minister Sharon and Palestinian leader Arafat, 50% and 54%, respectively. Many wanted sanctions against both sides—38% wanted to cut off aid and 33% wanted to block military exports (EUMC 2004b, 90). European Jews are ambivalent: Jewish interviewees in Belgium, France, Germany, Spain, and Greece complained about the conflation of "Jews" and "Israelis" made by their fellow citizens, who blame them for the situation of the Palestinians, which they themselves deplore (EUMC 2004c, 25). Muslim youths are ambivalent: A survey of French young people of North African origin found no massive antisemitism and more tolerance of Jews than the whole group of French youth between 15 and 24 (EUMC 2004b, 105). Moreover, Jewish interviewees perceive the European people as ambivalent because they honor the Jewish victims of the Holocaust but withhold respect for the living—respect is testified to the dead and not yet accorded to the living (EUMC 2004c 31).

The ambivalence of ordinary Europeans is indexed by the value they place on Jewish and Palestinian lives. The simple average of a country's proportions stating that the Palestinian Authority truly desires peace, Israel truly desires peace, and Jews do *not* talk too much about the Holocaust creates a reliable index ($\alpha = 0.75$). If this measure is added to the structural model, its effect on violence is conjectured to be positive. It captures the immobilizing ambivalence of Europeans, leading them to actively withdraw and not to protest the anti-Jewish violence perpetrated by Muslim youth and others. Consequently, and now less paradoxically, those countries that place a high value on Jewish and Palestinian lives can be expected to have higher counts of violence.

Model 9 of Table 8.10 quantifies the effects on violence of the value a country places on both Palestinian and Jewish lives. Compared with Model 5, both explain the variability between countries (0 compared with 0.017, $\hat{R}_B^2 = 1$ to $\hat{R}_B^2 = 0.99$) but Model 9 explains more of the level-1 dispersion (1.17 compared with 1.32, $\hat{R}_W^2 = 0.69$ compared with 0.65), its scaled deviance is smaller (15.8 compared with 16.8), and its deviance is noticably smaller, 18.6 compared with 22.2 (delta = 3.6, chi-square test with 1 degree of freedom, p ~ 0.051). But Model 5 has the smaller pseudo-BIC, 41.2 compared with 43.6. Comparing Model 8 and Model 9, there is little difference—comparable fit statistics round to the same whole number.

Figure 8.2 further clarifies the performance of these models, showing how they reduce each country's residual random effect. For Model 1, the baseline null model, these random effects range very widely, from 1.720 for France to −0.823 for Austria and Denmark. The model that includes only the structural variables flattens the random effects noticeably; these range from 0.192 for France to −0.297 for Germany. The mobilization model (5) that includes PAProPeace explains all of this variability. The bystander unresponsiveness model (8) that adds valuing Jewish lives to the structural model leaves some of the variability unexplained: the random



	Austria	Belgium	Denmark	Italy	Nether lands	Spain	Switzer land	France	Germany	UK
Baseline Model 1	–0.823	0.016	–0.823	–0.728	–0.555	–0.728	–0.728	1.720	1.037	1.610
Structural Model 4	–0.033	0.000	–0.033	0.000	0.066	0.000	0.000	0.192	–0.297	0.104
Model 5 - PA Pro Peace	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Model 8 - Values Jewish Lives	0.000	0.000	–0.046	–0.016	0.015	0.025	0.020	0.150	–0.037	–0.112
Model 9 - Values Muslim & Jewish Lives	0.001	0.000	–0.027	–0.010	0.020	0.008	0.008	0.095	–0.016	–0.078

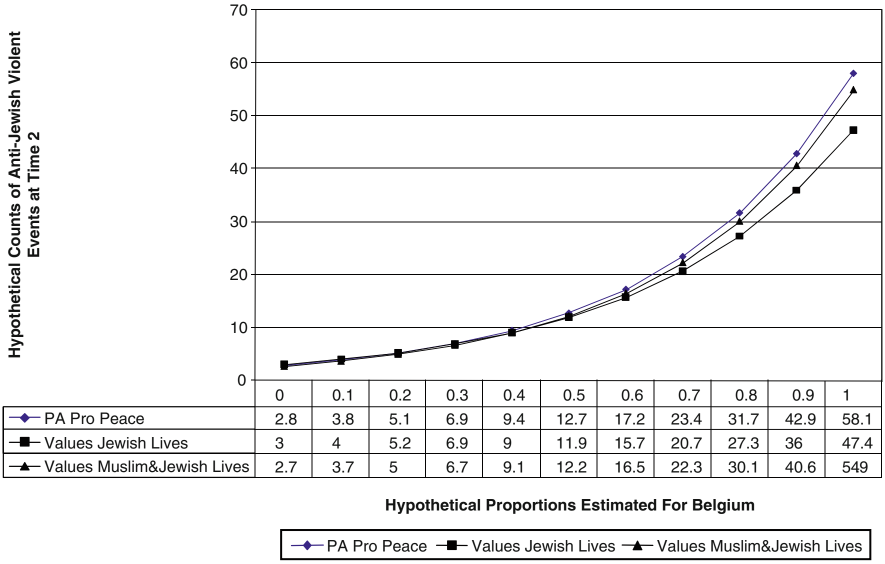
Fig. 8.2 The random effects approach zero in the better models

effects range from +0.15 for France to −0.046 for Denmark. The ambivalence model (9) that includes valuing Jewish and Palestinian lives also exhibits a slight range, from 0.095 for France to −0.078 for the United Kingdom.

Not one of these models totally dominates the other two; taken together the models suggest that, given the globalization of the Arab-Israeli conflict, a country's counts of violence depend upon the sizes of its Jewish and Muslim populations, the mobilization of the perpetrators, the unresponsiveness of bystanders, and the ambivalence of its ordinary citizens.[27] All three models add to our understanding of the causes of contemporary anti-Jewish violence in Europe. However, they leave unanswered two questions: how the factors' effects may unfold, especially in Belgium, the country with an outlying residual, and whether the findings are robust when data for the full period of the intifada are modeled.

Parameter Study of Belgium's Violence Counts

To explore how these factors operate, Fig. 8.3 depicts the results of a parameter study of Belgium's Time 2 violence counts for the three models (Models 5, 8, and 9). Belgium is the example because its outlying counts of violence, given its Muslim and Jewish population category, are usually overshadowed by the higher counts for Germany, France, and the UK—the resulting patterns may hold for these countries and for the others as well. For the range from zero to one, this parameter study varies in increments of 0.10 each model's hypothetical proportion for the



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
◆ PA Pro Peace	2.8	3.8	5.1	6.9	9.4	12.7	17.2	23.4	31.7	42.9	58.1
■ Values Jewish Lives	3	4	5.2	6.9	9	11.9	15.7	20.7	27.3	36	47.4
▲ Values Muslim&Jewish Lives	2.7	3.7	5	6.7	9.1	12.2	16.5	22.3	30.1	40.6	549

Fig. 8.3 Parameter study of three models predicting anti-Jewish violence in Belgium

three attitudinal variables: PAProPeace, valuing Jewish lives (the average of Israel wants peace and Holocaust remembrance), and valuing Muslim and Jewish lives (the average of the three variables). Belgium's actual means for these attitudes are, respectively, 0.46, 0.48, and 0.49.

For each of these three models, the counts of violence increase as the hypothetical proportion of the Belgium population holding the attitudes increases, but there are some subtle differences. Close inspection of the data table of Fig. 8.3 indicates that for low values of the hypothetical parameters the bystander unresponsiveness model produces the highest violence count. At 0.3, the mobilization model catches up and thereafter it produces higher counts than the other models. By 0.4, the cognitive ambivalence model begins to dominate the bystander unresponsiveness model and thereafter this model produces the second highest counts of violence. Given the actual mean scores for Belgium's attitudes, which ranged from 0.46 to 0.48, this parameter study suggests that mobilization and ambivalence were the key drivers of the violence, bystander unresponsive was less important. This pattern seems to hold in general when five years of data are analyzed.

Replication of the Main Analysis

Thus far, this chapter's interpretation of the anti-Jewish violence is based on only two years of data about violent incidents and on attitudinal measures that were assumed to hold contemporaneously with these events. Consequently, the main findings may be suspect. To address this potential issue here this chapter replicates the analysis (see Table 8.11) using an expanded data set that includes the trichotomous population typology that classifies countries, the Roth Institute's violence counts for the period

Table 8.11 Replication of final models of anti-Jewish violence during the second intifada, 2001 through 2005

Models	Replication, Model 1	Replication, Model 5	Replication, Model 8	Replication, Model 9
Covariance parameters				
$d_{j(k)}$	1.82	0	0.08	0.037
z	1.99	–	0.82	0.59
Pr(z)	0.02	–	0.21	0.28
e_{ijk}	4.11	1.85	1.75	1.82
z	4.49	4.06	3.7	3.75
Pr(z)	< .0001	< .0001	0.0001	< .0001
Ratio of $dj(k)$ to e_{ijk}	0.44 to 1	0 to 1	0.046 to 1	0.02 to 1
Fit statistics				
Probability that $e_{ijk} = 0$ fits data	< .0001	1	0.0057	0.046
Generalized chi square for model	201.19	61.15	57.79	60.14
Degrees of freedom of model	49	33	33	33
Level-2 R^2 analog	0	1	0.96	0.98
Level-1 R^2 analog	0	0.59	0.60	0.56
Deviance	182.3	66.6	59.4	63.0
Scaled deviance	44.4	35.9	33.9	34.6
−2 × Residual log pseudo-likelihood	133.7	67.5	71.21	70.1
Pseudo-BIC, IC = q	138.3	69.8	75.8	74.7
Type 3 tests of fixed effects of covariates				
Time 2001 to 2005	–	F Value = 1.72	F Value = 1.78	F Value = 1.78
Num DF = 4, Den DF = 27	–	Pr > F = .175	Pr > F = .162	Pr > F = .163
Three population categories	–	F Value = 117.1	F Value = 48.97	F Value = 68.4
Num DF = 2, Den DF = 6	–	Pr > F < .0001	Pr > F = .0002	Pr > F < .0001
Time × three population categories	–	F Value = 2.24	F Value = 2.35	F Value = 2.24
Num DF = 8, Den DF = 27	–	Pr > F = .056	Pr > F = .046	Pr > F = .056
Belgium	–	F Value = 11.8	F Value = 7.77	F Value = 9.6
Num DF = 1, Den DF = 27	–	Pr > F = .002	Pr > F = .0096	Pr > F = .005
Attitudinal variable	–	F Value = 32.6	F Value = .93	F Value = 4.41
Num DF = 1, Den DF = 27	–	Pr > F < .0001	Pr > F = .34	Pr > F = .045

Notes: e_{ijk} equals the generalized chi square divided by degrees of freedom. All models use the counts of violence for 2001 through 2005. "Remembrance" is "Too Much Shoah Talk" = 1. "Values Jewish Lives" is the average of "Israel Desires Peace" and "Remembrance." "Values Both Jewish and Palestinian Lives" is the average of "PAProPeace," "IsraelProPeace," and "Remembrance." In a Type 3 Test all of the variables are entered and the covariate of interest is then deleted. The significance of the deleted covariate is then calculated by comparing the model absent that covariate to the model that includes it. To create the flat file attitudinal data for missing years were assumed to be in equilibrium, the same as for the previous year.

2001 through 2005, and attitudinal measures from the ADL's surveys taken in 2002, 2004, and 2005; to create a full analytic file, reasonable assumptions were made about the persistence of attitudes across the gaps in these data. This analysis uses the full trichotomous typology of countries and crosses this typology with the indicators of the five yearly time periods. The indicator variable for Belgium very noticeably reduces the values of the extra-dispersion scale, so it was included in the final models. With the structural variables of time, population typology, the interaction of time and typology, and the indicator for Belgium already in the models, each of the three attitudinal measures were added to the model one at a time. The mobilization model and the ambivalence model fit these date more appropriately than the unresponsive bystander model. The fixed effect of the latter's test variable (valuing Jewish life) is not significant ($p = 0.34$), whereas the other two test variables have significant fixed effects (PAProPeace, $p < .0001$ and valuing both Muslim and Jewish lives, $p = 0.045$). Once again, the mobilization of the perpetrators by the events in the Middle East and the ambivalence of ordinary Europeans stemming from this conflict are the key drivers of the anti-Jewish violence; bystander unresponsiveness, antisemitic attitudes, and dislike of Israel are less important.[28]

Discussion

Summary

During the period of the second intifada against Israel, which began in autumn 2000 and, with the renewed conflict in the Middle East, continues today; a European country's population of Jews and Muslims creates its predisposition toward violent anti-Jewish incidents. The larger these populations are, the greater the counts of anti-Jewish violence. Countries with few Jews, regardless of the number of Muslims, have lower counts of violence than countries with many Jews and Muslims.[29] A model composed of the countries' social structural predispositions toward violence is not sufficient to reproduce the counts of violence closely. The stimulus provided by the globalization of the conflict in the Middle East also is needed to mobilize the perpetrators; it is the match that lights the fire. Most likely, many of the acts of violence against Jewish people are perpetrated by young Muslim males of North African origin who think their destructive acts against the European Jewish communities are a way of fighting Israel, and by neo-Nazis thugs. This globalization of the Arab-Israeli conflict reinforces parochial identities and sentiments of Muslims in Europe: globalization does not necessarily lead to homogenization of cultures. However, why do ordinary Europeans allow this violence to take place by not exhibiting forceful moral outrage against it? This analysis suggests that bystander unresponsiveness, indicated by valuing Jewish life, and especially the ambivalence stemming from conflicting pro-Jewish and pro-Palestinian attitudes, which are mutually neutralizing, lead many ordinary Europeans to ignore or discount these anti-Jewish violent events.

Implications

Some ordinary Europeans and Americans participated in protests against the Israeli intervention in Gaza that began toward the end of 2008 and ended 22 days later (12/27/2008–1/18/2009). This intervention spiked the count of anti-Jewish violent incidents in January of 2009 (Roth Institute 2009):

> With the start of Operation Cast Lead in the Gaza Strip on December 27, a wave of antisemitic manifestations swept the world. These included both violent activities (arson attack on synagogues, assaults on Jewish individuals, desecration of cemeteries, and vandalizing of Jewish property and Holocaust monuments) and verbal and visual expressions (insults, threats, gruesome caricatures, and stormy demonstrations). Although most of these activities featured traditional antisemitic motifs, their use was more extreme, intensive, and vociferous than was hitherto known. Muslim activists and organizations worldwide, and especially the radicals among them, showed a high degree of mobilization and were the moving force behind the demonstrations, together with leftist and human rights activists, and to a lesser degree extreme right circles. Jews and former Israelis also took part in some of the rallies, mostly in the US. ... Based on the data we have received to date, we estimate that there were close to 1,000 manifestations of antisemitism of all types in January world wide. The violent cases (including use of arms, assaults on persons, and desecrations) numbered close to 90, three times that of January 2008.

The mass media did not focus on the provocative rocket attacks that threatened the lives of ordinary Israelis and the integrity of the State of Israel; rather, the media portrayed the harm the Israeli intervention caused to civilians in Gaza and the widespread destruction there (Roth Institute 2009, 6). Such images of Israelis and of Israel in the mass media and on the internet may encourage Muslim youths to perpetrate these violent acts and reinforce, if not create, the ambivalence of ordinary Europeans. The actions of the former may be stimulated by the blatant antisemitic propaganda that Fig. 8.2 summarized; the ambivalence of the latter may be solidified by the anti-Israel bias of the mainstream media in Europe, which the content analyses of Media Tenor document (2006, 18–23). For example, on German television for the period July 2001 to July 2006 the ratio of negative to positive assessments was about 20 to 1 against the democratically elected government of Israel, and negative aspects of the Israeli army dominated coverage.[30] Media Tenor (2006, 23) recommends that the media improve the fairness of their news reports thereby enabling TV viewers to form their own opinions.[31]

Policy Recommendations

To reduce this violence, Jewish and Muslim communities should encourage mutual understanding and respect and put a stop to propaganda that is defamatory; the news media should strive for fairness regarding Muslims and Jews and their conflict. Ordinary Europeans should pressure the police, courts, and criminal justice systems to protect Jewish institutions and people; to arrest suspected perpetrators and sponsors of this violence; and to convict guilty assailants and vandals.[32] Most importantly, Israel and the Palestinian Authority, Hamas, and Hezbollah should create the conditions for a sustainable peace in the Middle East, which, no doubt, would enhance the prosperity of the region and reduce anti-Jewish violence throughout our globalized world.[33]

. . .

In countries with advanced economies, stably employed working and middle class employees no doubt regard the nature of their jobs as more salient than the events in the Middle East and their impacts on Jews and Muslims: local concerns are more salient than the global. Employees desire substantively complex work that allows them to make decisions and to experience a sense of autonomy. When the substantive complexity of their work is curtailed, employees are likely to express their discontent. Studying this problem, the next chapter introduces the two-equation approach for developing multilevel models. The multilevel model is composed of variables on employees at level-1 who are grouped into offices at level-2, which are then classified by explanatory typologies of office characteristics.

Acknowledgements The author derived this chapter from his article in *Quality & Quantity* (2008, vol. 42, issue 2, 135–180; © 2006 Springer). He thanks Greg Maney and other reviewers for their helpful comments, and Philip Gibbs of the SAS Institute for clarifying aspects of Proc Glimmix.

Endnotes

[1] This description of the second intifada was gleaned from the Wikipedia encyclopedia.

[2] The Roth Institute (2007, 11) offers this interpretation of why countries that honor the holocaust may have elevated antisemitism: "Paradoxically, the antisemitic discourse and antisemitic manifestations are often triggered by bills, appeals or laws against antisemitism or Holocaust denial, which, it is claimed, violate freedom of expression. These discussions force a confrontation with the past, whose perception in West European states depends on their level of cooperation with the Nazi regime."

[3] Nicholas D. Kristoff quoted King's statement in his Op Ed "The American Witness," *New York Times*, March 2, 2005, page 19.

[4] Useem (1998, 220–223) reviews these and more recent studies on this topic.

[5] For examples of the application of multilevel Poisson regression models to social structural analyses see Sampson et al.(1997), Raudenbush and Bryk (2002, 309–317), and Zheng et al. (2006).

[6] Media Tenor (2007, Charts 66–70) groups 12 countries as Western and 12 countries as Islamic and then characterizes the themes their mass media stress. Concerning the West, Judaism, Christianity, and secular ideologies, in all five of these analyses the average ratings of the Islamic countries were more negative than the ratings of the Western countries. The Western countries did not exhibit noticeably more negative evaluations of Muslims and Islam than did the Islamic countries.

[7] Lindemann (2000, 77–91) organizes European anti-Semitic themes for the period 1914–1933 as tapping their power (Jews as powerful) and evilness (Jews as shirkers and subversives); thus, Jews should be abolished. Slezkine (2004) suggests that Jews will prosper when prejudice does not impede there advancement.

[8] At the country level of analysis a two-item index of disliking Israel's stance of peace, which is composed of the belief that Israel does not truly desire peace and that the respondent does not favor Israel in the conflict in the Middle East, is positively associated with a five-item index of conventional antisemitic attitudes ($r = 0.22$, $p = 0.353$) and several of its subscales: Jewish in-group loyalty ($r = .21$,

$p = .38$) and Jewish corruption ($r = 0.16$, $p = 0.51$). These measures are described later in this chapter. If the data are frequency weighted by the sample of respondents then these correlations are significant at the $p < .0001$ level.

[9] Operario and Fiske (2004) trace the results of studies of stereotypes and stereotyping from their genesis in Walter Lippmann's *Public Opinion* (1922) to the present.

[10] Cooper et al. (2004, 255–263) trace aspects of cognitive dissonance theory from Festinger's early formulations through today's research.

[11] For the year 2005, for example, the counts of major attacks and violent incidents, respectively, for four European countries are as follows: UK, attacks = 1, incidents = 89; France, attacks = 7, incidents = 63; Germany attacks, = 0, incidents = 37; and Belgium, attacks = 1, incidents = 8. In Canada there were 0 attacks and 44 incidents, considerably more than the 19 incidents (zero attacks) in the USA.

[12] Taylor Nelson Sofres (TNS) also conducted the 2007 and 2009 surveys for the Anti-Defamation League.

[13] As mentioned earlier, "Knowledge about the Holocaust, Holocaust Remembrance, and Antisemitism" (Smith, April 17, 2004) found that countries with higher levels of Holocaust remembrance and knowledge about the Holocaust have the higher levels of antisemitic violence. This paradoxical finding is consistent with the finding in this chapter that countries with lower levels of conventional antisemitism (including Holocaust remembrance) tend to have higher counts of antisemitic violence. For ordinary Europeans antisemitic stereotypes seem not to be key drivers of the violence.

[14] Analyzing the respondent-level ADL surveys for 2004, Kaplan and Small (2005) find that anti-Israel sentiment predicts antisemitism in Europe. Their data and codebook were available on the internet circa 2008 at http://jcr.sagepub.com.

[15] PAProPeace has the following Spearman rank correlations (r_s) with the other indicators of disaffection–all are not statistically significant: with ProPalestine $r_s = +0.09$ ($p = 0.70$); with IsraelAntiPeace $r_s = -0.415$ ($p = 0.069$); and with ViewsIsraelUnfavorably $r_s = -0.26$ ($p = 0.26$). Paradoxically, PAProPeace has significant correlations with IsraelProPeace, $r_s = 0.57$ ($p = .009$) and with Holocaust Remembrance, $r_s = 0.44$ ($p = 0.049$). The latter two variables are significantly correlated with each other, $r_s = 0.48$ ($p = 0.03$).

[16] Comparisons between 2007 and 2009 regarding the distributions on indicators of anti-Semitism in seven European countries document the stability of such attitudes as these. Among these countries only in the UK did the percentages decline (Anti-Defamation League, 2009, 17–26). Julius (2010) analyzes the history of anti-Semitism in England focusing on historical, literary, political, and anti-Israel roots.

[17] Belgian nationals of Moroccan origin formed the core of a terrorist group that presumably planned to target leaders of the European Union with a suicide attack during their two-day summit meeting in Brussels in December 2008. Six of the fourteen who were arrested and not let go because of lack of evidence included Ms. Malika El Aroud, who writes inflammatory jihadist propaganda on the internet under the byline "Oum Obeyda." Her late husband received training from Al Qaeda and participated in the killing just prior to 9/11 of Ahmed Shah Massoud, an anti-Taliban resistance leader in Afghanistan. Her current husband, Moez Garsalloui, was arrested and imprisoned for three weeks in 2007 for promoting violence. Upon his release he fled to Pakistan and Afghanistan and is one of the three arrested men who recently just returned from training camps along the Afghanistan–Pakistan border; at least one of these was to be the suicide bomber. Several of these suspects have direct ties to Al Qaeda. For the details see Steven Erlanger's article in the *New York Times*, December 13, 2008.

[18] Why the political left supports the Palestinians and is anti-Israeli is puzzling, since Israel has a democratic-socialist past and is a democracy while few if any Muslim countries are democracies (e.g., at a seminar at Harvard University in 2010, Turkey and Iraq were mentioned as being the only democratic Muslim countries by a reporter from Al Jazeera). The assumption that the political left is not anti-Semitic as those on the Left claim, but merely anti-Israel because of its harsh occupation of Gaza and the Left Bank is questionable. The latter view may mask an effort to coopt via anti-Israel rhetoric and actions the large populations of Muslims in European countries to support particular political agendas, whatever these may be. The political left's hostility may be further exacerbated by Israel's change from a socialist

economy to a free market economy, and its change from a primarily secular society to one with militant religious minorities settling in Palestinian areas.

[19] The BBC's rather objective reporting of the violence in the Middle East may moderate anti-Jewish and antisemitic sentiments in Great Britain. For example, during the period of the conflict in Lebanon between Israel and Hezbollah (July 21, 2006 to August 3, 2006), assessments of the protagonists in their TV news coverage of this war was more restrained than in other TV reports (Media Tenor 2006, 10).
[20] Raudenbush and Bryk derive the Poisson multilevel model using their two-equation approach (2002, 309–317).

[21] SAS Institute (2005, 22–23) explicates the computations for these fit statistics. The computer output, however, states this caveat about their use: "Fit statistics based on pseudo-likelihoods are not useful for comparing models that differ in their pseudo data." However, one can compare scaled deviances between two competing models. The difference in scaled deviances is approximately chi-squared distributed with the degrees of freedom equal to the difference in the number of parameters between the two models.

[22] The pseudo-BIC of a model gauges how well it fits its pseudo-data rather than how well it fits "the" data. Comparison of the pseudo-BICs of two nested models therefore indicates how well each of the models fits its own pseudo-data, with smaller values of BIC indicating the better fit. That a given model fits its pseudo-data closely does not necessarily imply that it would fit "the" data better than the other model.

[23] Goldstein (1987, 21–27) develops an example in which he introduces a fixed indicator variable for a school that has an outlying residual.

[24] The Roth Institute (2006, 8) reports that in Belgium the number of violent events increased from 9 in 2005 to 16 in 2006, including this extreme event: "a religious Jewish couple and their baby were attacked in early January on the train to Antwerp. The father, who tried to protect his wife and baby, confronted the attacker who was armed with a knife, sustaining superficial wounds."

[25] The mobilization model was further tested by adding to it measures of human and economic development, an indicator of whether or not the country experienced deportations of Jews during the Holocaust, and whether or not a country made reparations to victims of the Holocaust. None of these additional effects attained statistical significance. On the log scale, an increase in a country's human development score reduced anti-Jewish violence by -1.94 ($t = -0.10$, $p = 0.92$); a country's increase in gross domestic product per person reduced anti-Jewish violence by -4.5 ($t = -0.57, p = 0.59$), countries that experienced deportations had reduced violence by -0.08 ($t = 1.29, p = 0.78$), and countries that had made reparations (Germany, Austria, and Switzerland) had reduced violence by -0.32 ($t = -0.75, p = 0.47$).

[26] The Roth Institute (2006, 5) describes the shifting views of ordinary Europeans as a consequence of the media's reporting, especially of the war in Lebanon: "The short period of sympathy which Israel enjoyed at the beginning of the war in July 2006 was swiftly reversed after the Qana incident, in which civilians, including children, were killed." The Media Tenor (2006, 5) analysis of the content of TV reports in Germany about the war in Lebanon indicates that segments on the war did not emphasize that Hezbollah hides its forces in kindergartens, hospitals, and elsewhere among the civilian population. The news segments focus on the Israeli forces and their attacks whereas Hezbollah's attacks are not mentioned or shown.

[27] An ambivalent political and social culture regarding Muslim immigrants and citizens exists in many of these European countries, as exemplified by the reaction in the Netherlands to the murder of Theo van Gogh by a Muslim extremist. Van Gogh produced a film that was highly critical of Islam's treatment of women. His murder unleashed retaliatory attacks on Muslims and their mosques by white youths and facilitated the formation of a new political party that aims to curtail immigration for five years. It also led to demonstrations for political tolerance in Amsterdam. The Dutch value an open society and democracy but take exception to the lack of assimilation of many of their Muslim residents and fear their radicalisation. A friend of van Gogh underscored the society's ambivalence saying: "Thirty percent of the Dutch people are racist, 30% are not, and the rest do not know what they think" (quoted by Ian Bicketon, *Financial Times*, November 17, 2004, 15). Perhaps actions by the radicalized minority of Muslims against ordinary Europeans will mobilize the latter to crack down on Muslim violence against Jewish people.

[28] Using the five years of data and estimating the multilevel model when the five-item measure of antisemitism is the test variable, the sign of the coefficient is negative and its fixed-effect is not significant (-0.37, $p = 0.79$). When dislike of Israel is the test variable, its coefficient is positive but not significant (3.76, $p = 0.23$). Note that the average count of violence dropped in 2005 from its 2004 value (from 38.2 to 22.4) across the ten countries. For countries with many Jews and many Muslims the average count dropped from 79 in 2004 to 65.7 in 2005. For countries with few Jews and many Muslims the average count dropped from 3.3 in 2004 to 0.67 in 2005. However, for countries with few Jews and few Muslims the average count increased from 5.25 in 2004 to 6.25 in 2005. Within this group the count for Denmark increased from 3 to 10 incidents whereas for Belgium the count decreased from 14 to 9. Overall, this pattern of change supports the view that changes in the intensity of Israeli-Arab violence are reflected in the counts of anti-Jewish violence in Europe.

[29] Data on violent manifestation in 2008 from the Roth Institute (2009, 57) confirm this conjecture for the European countries in this sample. The number of violent incidents in the UK = 112, Germany = 82, and France = 50. Belgium is again an outlier, with 22 incidents; the other countries in the sample had much lower counts.

[30] The empirical data for these trends are 2,355 new stories in ARD Tagesschau and Tagesthemen and ZDF heute and heute journal (Media Tenor 2006, 19).

[31] Some of Media Tenor's (2006, 23) specific recommendations for fairness in TV newscasts during conflicts are: "Provide context; Same air-time for both sides; Show record of cease fires; Put separate incidents into context; Correct false reports; Daily reference to the problem of correct information during wartime; Equal coverage on victims; Were is democratic control? Put judgments into context;" and so forth.

[32] Mark Huband (*Financial Times*, November 17, 2004, 15) reports that France has demanded that Muslims integrate within French society, whereas the UK and the Netherlands have taken a more multicultural path. To control the radicals, the British police have established strong lines of communication with Muslim leaders whereas the Netherlands relies more on domestic intelligence and security services and may not be sensitive enough to Muslim concerns. The recent violence against ordinary citizens and their property in the UK and France suggest that all of these countries may face threats from radical Islamists who believe that violence advances their cause.

[33] The 12-nation study by the Anti-Defamation League (May 2005, 10–11) suggests that about 29% of ordinary Europeans said that their opinion about Jews was influenced by the actions taken by the State of Israel. Of those who changed their opinion about Jews, 53% said their opinion of Jews was worse because of the actions taken by Israel. Although these results did not factor in the respondents' levels of conventional antisemitic beliefs, these finding do underscore the linkage between the events in the Middle East and opinions about Jewish people. Negative opinions about Jews could lead to bystander unresponsiveness and cognitive ambivalence toward anti-Jewish violence when it erupts in their cities and countries.

Chapter 9
Will Claims Workers Dislike a Fraud Detector?

An intervention DAG model can be justified only to the extent that it fits the behavior of the world in the setting to which it is intended to apply.

—A. P. Dawid (2002, 170)

White-collar workers in insurance companies spend almost all of their working hours sitting in cubicles and interacting with personal computers. The typical suite of office programs may enable them to express their creativity and individuality by writing memos, performing spreadsheet calculations, and designing presentations—such work is substantively complex, requiring thought and independent judgment (Kohn 1969, 1977). Contrariwise, the typical administrative computer program restricts the user's creativity by structuring the information that is needed—this work requires conformity to the demands of the program (Kohn 1969, 1977).[1]

Clarifying aspects of a change from substantively complex to more routine work, this chapter assesses how a computerized expert system—a fraud detector—affects the attitudes of white-collar insurance workers (Carley 1988; Bainbridge et al. 1994, 412–415). At present, most employees whose job it is to identify fraudulent claims use their judgment (Ross 1970, 87–175). They gather information from claimants, assess this information, and classify claims for either routine processing or for further inquiry concerning possible fraud. By imposing on the workers the need to enter structured information, the fraud detector may reduce the substantive complexity of their work.

Because they are concerned about expense ratios and profitability, executives of insurance companies want to reduce the cost of underwriting insurance policies, claims processing, and fraud detection. Because they value economic rationality, they are receptive to computerized algorithms that would replace expensive skilled workers with less expensive data-entry employees. Consequently, managers in an insurance company sponsored this evaluation of the computerized fraud detector (CFRD), an algorithm that is based on empirical findings about automobile insurance fraud. After claims workers enter the required data, CFRD assigns suspicion scores to claims and then categorizes them into one of three tracks: refer to special investigation, gather more information, or process routinely. Thus, CFRD's

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_9, © Springer Science+Business Media B.V. 2011

automation of the process of fraud detection may constrain the judgment and skills of the employees (Burris 1998, 146–149).

The initial design of the evaluation called for a pilot testing of CFRD in three target claims offices; its impacts would be assessed relative to three closely matched comparison offices. However, qualitative field observations and interviews uncovered that in two target offices and in one comparison office a second, more comprehensive administrative computer system—Millennium 2000 (M2K)—also was being installed. The implementation of this second computer system severely threatened the validity of the original study design.[2] This shortcoming eventually led to the creation of a new study design and the application of hierarchical linear modeling for quantification of the effects.

New Contributions

Assessing the determinants of the claims workers' attitudes about the computerization of fraud detection, this chapter probes the effects of the simultaneous introduction of two new computer systems, and the effects of the singular implementations of the two new systems, all relative to two comparison offices that are not experiencing technological change. It thus advances the current substantive literature that probes the impacts of only one new computer system (Liker and Sindi 1997, 150–151; Burris 1998; Liker et al. 1999).

This research also elucidates a methodology for hierarchical modeling. It presents a series of contextual analyses that initially followed the elaboration procedure of Bryk and Raudenbush (1992, 2002), as explicated by Singer (1998, 323–339) for SAS's Proc Mixed. The logic of the present analyses differs from theirs in that the pivotal contextual variable is a typology of offices defined by the number and kinds of new computer systems that are being installed (Littell et al. 1996, 149–155; 2006, 75–81). With the six offices designated as random and the offices classified by the typology, the analyses quantify the effects on worker attitudes of fixed individual-level covariates and the fixed office-level categories of the typology. Since the offices are designated as random, and since the fixed effects are bundled within the offices, the key results may allow inferences concerning the fixed effects that apply to the entire population of the insurer's claims offices (SAS Institute 1997a, 582).[3]

Hypotheses

Regarding computerization, office workers can be expected to act cognitively rationally. Their beliefs, which are derived from good reasons, are not irrational nor do they result from a hyper rational analysis of costs and benefits (Boudon 1996, 2003; Etzioni 1988; Smelser 1998, 1–4). To learn how to use a new computer system, they must change their old habits and absorb new information and instructions; this may be disruptive cognitively and annoying—a hassle—especially when the new system does not work well. Clearly, if employees must learn how to use two new computer systems at about the same time, as well as to conduct their usual business on the old system, then the technological intrusion into their workplace will be even more severe than if they have to adapt to only one new system. Thus:

> *Hypothesis 1*: The simultaneous introduction of two new computer systems will engender more anticomputerization sentiment than the introduction of only one new system, which in turn will engender more anticomputerization sentiment than the introduction of no new systems.[4]

An employee's receptivity to innovation—a general predisposition—is responsible in part for the employee's specific attitude toward computerized fraud detection. This direction of effect between these attributes of an employee is consistent with the ordering principle of Lazarsfeld (1955b, *xi*) and Davis (1985, 17); namely, that a general predisposition (e.g., Democratic Party identification) has specific manifestations (e.g., votes for Democratic candidates). Thus (Merton 1957a, 149–153):

> *Hypothesis 2*: Employees who are receptive to innovations, generally, will be likely to approve of computerized fraud detection, specifically.[5]

Because CFRD may limit the autonomy and the substantive complexity of the work of skilled claims specialists, and also pose a risk to their job security, it would be rational for these skilled employees to dislike computerized fraud detection more strongly than the less-skilled claims representatives—the former have more to lose (Merton 1957b, 564; Mills 1956, 231–232; Zuboff 1988, 129–150; Burris 1998, 146–149; Hage 1999, 609–610). Thus, regarding deskilling:

> *Hypothesis 3*: Employees with jobs of higher rank, claims specialists and supervisors, will be more likely to dislike computerized fraud detection than employees with jobs of lower rank, primarily the claims representatives.[6]

Because the use of CFRD could lead to efficiencies that might engender more layoffs than M2K, we have:

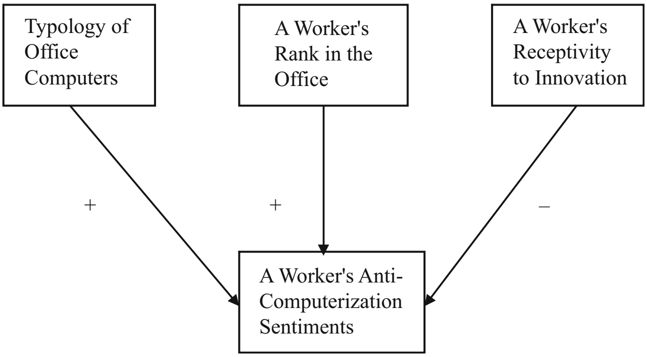
> *Hypothesis 4*: The office with a singular implementation of CFRD will have a higher level of discontent than the office with a singular implementation of M2K.[7]

Method

Study Design

The original research question asked whether implementation of CFRD would engender resistance to the use of this innovation by the claims workers. To answer this question, the researchers selected six geographically dispersed claims offices: three of these would receive the fraud detector treatment, and three closely matched offices would receive the null treatment. To measure worker attitudes and to control for potentially spurious effects, each worker would fill out a questionnaire that covered their background characteristics, work experience, awareness and knowledge of fraud, and attitudes about innovation and computerized fraud detection; the latter would be measured by an "anti-index." The analysis would exemplify causality as an effect of an intervention, thus applying Rubin's causal model, as explicated earlier in Chapter 3. The average causal effect of the fraud detector treatment would be the difference between the expected values on the anti-index Y in the treatment t and control c groups: $\delta = E(Y_t) - E(Y_c)$. To control for spuriousness, various mean-centered variables would be held constant.

This original design was abandoned when it became apparent that another computer system also was being installed in three of the offices at about the same time that CFRD was being installed: Two offices had both CFRD and M2K, one office had CFRD, another M2K, and two offices had neither. These variations suggested a typology of interventions that could serve as fixed covariate and also a classification (i.e., nesting) variable for the six random offices in the multilevel model; other covariates would be indicators of the worker's rank and resistance to innovation; diagrammatically, for each claims office:


Typology of Office Computers
A Worker's Rank in the Office
A Worker's Receptivity to Innovation
+
+
−
A Worker's Anti-Computerization Sentiments

In the multilevel model, the workers and their attitudes are the level-1 units, the six random offices are level-2 units that contain the workers, and the typology of office computer systems classifies the offices and is a fixed covariate in the model.

The Data

To capture the attitudes of the claims workers, the evaluators conducted exploratory focused interviews, drafted a questionnaire, and pre-tested it in two extra offices. To improve item coverage and clarity, they revised the questionnaire taking into account the results of the pretest and their first-hand observations of two target offices. The final questionnaire probes attitudes toward computerization, morale, office cohesion, self-rated efficiency in identifying fraudulent claims, suggestions for improving claims processing, position in the office, and personal characteristics.

The beginning dates of the pilot implementations of CFRD were: Florida, August 14; Boston-area, September 1; and Philadelphia-area, September 27. In the latter two target offices, and in the Philadelphia-area comparison office, the process of installing M2K began circa October 15. After each target office had at least three month's experience with CFRD, a researcher transmitted the questionnaires via electronic mail to the managers of the offices. Using email and their office distribution list, the managers transmitted their copy of the questionnaire to each of their fraud claims workers and urged them to complete it. In each office, one person was responsible for gathering the anonymous questionnaires and mailing them to the evaluators. After two weeks, 198 questionnaires had been returned; the response rate was high, estimated to be about 90%. A professional data-entry service prepared a machine-readable data file. By counting whether or not a respondent carefully answered several open-end questions, an index of the quality of a returned questionnaire was created. A control for this quality index has no effect on the results. Because of missing data, the hierarchical data analyses are based on replies to 193 questionnaires.

Measures

Office types: the stimulus variables

The office typology groups the six offices according to the number and kinds of new computer systems—quasiexperimental interventions—to which their employees were exposed. The Boston- and Philadelphia-area target offices are classified as having had two interventions, CFRD and M2K ($n = 66$, 34%); the Florida target office only had CFRD ($n = 13$, 6.7%), the Philadelphia-area comparison site only had M2K ($N = 61$, 31.4%), and the Florida- and Boston-area comparison sites had neither ($n = 54$, 27.8%). To balance the design, some analyses in this chapter group together the two offices that only had one new system ($n = 74$, 38.1%). As reported later, the reliability of the sample mean for each office is sufficiently high (even for the Florida target site).

To assess whether there are significant compositional differences among the various types of offices, which could possibly bias the results, a series of exploratory one-way analyses of variances were conducted; Table 9.1 presents the F ratios and their significance. Regarding the employees' background characteristics and indicators of their morale, it reports no statistically significant differences among the office means. For four office contexts or for three, the null hypothesis of equal means can not be rejected; the Bonferroni post hoc tests for multiple comparisons indicate no statistically significant pairwise mean differences among the office types for the ten dependent variables. However, regarding the use of CFRD and the subjective evaluations of it by users, it reports statistically significant differences among the office types. The obtained probabilities of $p < .0001$ are less than the Bonferroni-corrected probability for the rejection of the null ($p = 0.0012$), which implies that the office means differ regarding CFRD.[8]

To specify the above analysis, the responses to this open-end question were coded: "If you have used the CFRD system, how would you evaluate it?" Respondents in offices that did not have CFRD installed indicate no use of that system,

Table 9.1 In different types of office contexts, employee background characteristics and morale are similar, exposure to CFRD varies

	Four types of offices		Three types of offices	
	F ratio	Significance Probability	*F* ratio	Significance Probability
Background characteristics				
Q37, Gender	0.92	0.43	1.26	0.29
Q38, Educational level	0.31	0.82	0.21	0.81
Q36, Usually handles no fault claims	0.17	0.92	0.25	0.78
Q33, Job title	0.23	0.88	0.09	0.92
Q34, Job grade	0.51	0.68	0.11	0.90
Indicators of office morale				
Q28, Trusts coworkers	0.08	0.97	0.01	0.99
Q22, Work group is terrific	0.83	0.48	0.16	0.85
Q27, Supervisor is very effective	0.91	0.44	1.37	0.26
Q20, Special investigation unit (SIU) handles claims very well	2.04	0.11	1.60	0.21
Q26, SIU is helpful	1.30	0.27	0.07	0.94
Exposure to CFRD				
Q40, Has used CFRD	229.42	< .0001	135.60	< .0001
Q41, Difference between types of offices in subjective evaluations of CFRD	207.86	< .0001	205.01	< .0001

The Bonferroni-corrected probability-value for the comparison of each of the exposure measures to the family of ten background and morale indicators is 0.0012, which is more probable than the obtained probabilities, but is still very statistically significant. Three background characteristics—educational level, type of claim handled, and job title—compose a factor *Jobs of Higher Rank* that is used as a control in the subsequent analyses. The other indicators are not used as controls because they have very little effect on attitudes toward computerized fraud detection.

whereas respondents in offices designated as having CFRD indicate almost universal use (about 87%). Overall, about 11.4% of the seventy users of CFRD have favorable attitudes toward it, 15.7% are neutral, and 72.9% are unfavorable. Users in target offices with both new systems are more unfavorable than users in the Florida target office were only CFRD was installed. The four responses coded as very favorable (+ ++) are all from that office, whereas the 32 responses coded as very unfavorable (– – –) are all from the two target offices that have joint implementations.

Anti-computerized fraud detection: the response variable

Qualitative explorations

Prior to the formulation of the survey questionnaire, the researchers visited the target sites near Boston and Philadelphia and conducted exploratory interviews with users of CFRD; their attitudes were not favorable—"It stinks." During the first few weeks of the test period, under certain conditions, CFRD caused the user's computer to crash. The vendor soon fixed the technological incompatibility between CFRD and the insurer's FACES interface and taught the users how to avoid system crashes. (FACES is an acronym for Fast Automated Claim Entry System; M2K will replace it.) Even so, the belief persisted in the offices that the use of CFRD caused the computer to crash—"I have tried to use CFRD, but it kept making my computer crash"—and this belief inhibited some people from using it.

Since CFRD was not integrated into the FACES system, but ran in parallel with it, the users of both systems had to enter the same data twice, once for FACES and once for CFRD. This dual data entry, which was an unavoidable shortcoming of the pilot implementation, led users of CFRD to perceive that it wasted their time—"Useless and time consuming," "A waste of time."

A third negative attitude was prevalent among supervisors and claims specialists; namely, that they could detect fraud better than CFRD—"A well-educated claims person beats a machine anytime!" These skilled employees may have perceived CFRD as more threatening to their jobs than M2K. Since CFRD is an expert system, it may limit the judgment exercised by the claims specialists and could empower the less skilled claims representatives to accomplish the specialists' work. From the point of view of all these employees, M2K is merely a new user interface that promises to make their job easier.

The complex of the users' negative attitudes was thus composed of these aspects: CFRD caused systems to crash, wasted time, and was not useful—people could detect fraud better than computerized systems. (Contrariwise, a proprietary part of the evaluation found that CFRD detected suspicious claims more effectively than the claims workers.)

Forming the index

Since the employees in the comparison offices did not have the opportunity to use CFRD, and most probably did not know about it, a direct question about their attitudes toward CFRD would be inappropriate. Consequently, to form a parsimonious, simple index of attitudes toward computerized fraud detection that could be explained to a lay audience, this chapter uses these two general questions that are based on the above qualitative observations:

> Please rate how potentially useful the following tools for fighting fraud would be: Question 16. Computerized Fraud Detector—Not Useful at all (1) to Very Useful (7).
>
> Question 29. Claims employees can detect fraudulent claims better than computerized systems: Strongly Disagree (1) to Strongly Agree (5).

When, for simplicity, the above questions are dichotomized, their response distributions are as follows. About 38.9% are grouped as indicating that a computerized fraud detector would not be useful—[scores 1 through 3 = 26.7% and 4 = 12.2%] versus [5 through 7 = 61.1%]. About 75.7% are grouped as indicating that people can detect fraud better than a computerized system—[scores 4 and 5 = 75.7%] versus [3 = 21.3%, 1 and 2 = 3%]. To form the trichotomous index, these dichotomized responses, which are coded 1 for unfavorable responses and 0 for favorable responses, are summed.[9] The resulting distribution is rather symmetric: about 37% gave anticomputer replies to both questions and have index score 2; about 43% gave one anticomputer response and have index score 1; about 20% gave no anticomputer response and have index score 0—the mean of the scores is 1.17.[10] The full items that compose this index, hereafter referred to as the anti-index, have a Cronbach reliability coefficient of alpha = 0.73, sufficiently high.

Validity

The distinctions of the anti-index and the intervention typology are valid. The correlation of the anti-index with having used CFRD is 0.40 ($p < .001$, $n = 196$) and, for seventy users, the correlation of the anti-index with unfavorable attitudes toward CFRD is 0.32 ($p < .008$). Also, the individual items have similar unitary consequences on a range of criterion variables (Back 1951).

Because of the system crashes, dual data entry, and possible threat to jobs, the target offices, and not the comparison offices, have the higher percentages of employees with attitudes that are anticomputerized fraud detection. Using the dichotomized anti-index (index score 2 versus scores 1 and 0), the effect parameter (Coleman 1981, 19–62) is 0.408 ($t = 6.32$, $p < .001$). However, the statistical significance of this difference is primarily due to the simultaneous installation of both systems—in those offices, the difference between target and control is much larger and statistically significant. In the Philadelphia-area offices, 65.7% of the respondents in the target office are against computerized fraud detection, compared with 20.3% in the matched office, the effect parameter is 0.454 ($t = 4.88$, $p < .001$). In the Boston-area offices, 60% of the employees in the target office are against compared with 15.4% in the matched office, the effect parameter is 0.446 ($t = 3.76$, $p < .001$). When these offices are pooled, the difference between target and control is 0.44 ($t = 6.16$, $p < .001$). The small difference of 0.057 between these two target offices is not statistically significant ($t = 0.47$, $p = 0.640$). In the Florida target office, in which only CFRD was installed, 33.3% are opposed compared with 13% in the matched Florida office; this difference of 0.203 is not statistically significant ($t = 1.43$, $p = 0.164$). The difference of –0.297 between the Florida target office and those with both interventions approaches significance ($t = 1.94$, $p = 0.056$). Moreover, when answers to the open-end question about CFRD are coded from 1 (most unfavorable) to 7 (most favorable), they indicate that the Florida target site differs from the other two target sites; the scores are 4.73 to 1.98 ($t = 5.64$, $p < .001$).

Consistent with Hypothesis 1, when all the sites are organized on the basis of the number and kinds of new systems being installed, the relationship between type of site and the anti-index is linear.[11] In the sites with the simultaneous implementation of the two systems, the scores on the anti-index are very high (63%). These scores decrease when there is only one implementation—33% for CFRD Only and 20% for M2K Only (the difference between these two offices is not significant, $t = 0.98$, $p = 0.33$), reaching a minimum of discontent when neither system is being implemented (14%). The offices with both systems significantly differ from those with only one system, the effect parameter is 0.405 ($t = 5.21$, $p < .001$); and they significantly differ from the offices with no new systems, the effect parameter is 0.488 ($t = 6$, $p < .001$). The difference of 0.083 between the two offices with only one system and those with none is not significant ($t = 1.04$, $p = 0.302$). Thus, without controls for the effects of the covariates, the joint implementation of two new computer systems appears to be the crucial difference among these offices.[12]

Regarding Hypothesis 4 (that CFRD is potentially job threatening whereas M2K is not), the minimal difference in discontent between the two offices with these singular implementations suggests that job insecurity is not a salient source of the discontent.[13] Rather, it is the number of new systems being installed. (Later on, this chapter presents a parameter study that assesses how sample size affects the finding of no statistically significant difference between these two offices.)

The covariates

The covariates assess the employees' receptivity to innovation and the rank of their job. Whereas the office typology gauges the number and kinds of interventions—that is, changes—and thus connotes causal effects, an employee's scores on the covariates are associational attributes of the employee. Because these attributes are not changed through manipulations, this chapter conceptualizes them as associational and not causal in the sense of Rubin (1974) or Holland (1986, 945–948). Moreover, within the context of this study, because these characteristics have no significant correlation with the office typologies, they are conceptualized as *static*, i.e., as fixed attributes of the employees. To obtain scores for these constant characteristics, indicators of the employees' attitudes toward innovations and indicators of the rank of their job were factor analyzed, and mean values were substituted for the very few cases with missing values. For the nine items in the factor analysis the average missingness was 2.7% with a low of 1.5% and a high of 5%. Extraction by principal components, the Varimax rotation, and the regression method for the calculation of factor scores produced fully standardized factor scores (mean = 0, standard deviation = 1) for each of the two orthogonal factors ($r = 0.000$). When the effects of these covariates are evaluated at their mean value of zero, the predicted mean level of discontent for a cell of the design typology reduces to the sum of the intercept and the effects of the binary-coded (0 or 1) design variable(s) in the regression equation.[14]

Receptivity to innovation is assessed by the respondents' ratings (1 = not at all useful to 7 = very useful) about how potentially useful the following new tools for fighting fraud would be (the mean score for each question is enclosed in parentheses followed by its loading, which is the correlation between the item and the factor, and the number of cases with missing values): Q12, Internal Red Flag System (mean = 5.21, loading = 0.567, missing $n = 3$); Q13, Medical Provider Profiles (mean = 5.27, loading = 0.816, missing $n = 7$); Q14, Lists of "Bad" (i.e., corrupt) Lawyers (mean = 5.10, loading = 0.811, missing $n = 5$); Q15, Lists of Fraud Rings (mean = 5.79, loading = 0.788, missing $n = 4$); Q17, More Investigation Resources (mean = 5.88, loading = 0.660, missing $n = 4$); and Q18, Better Coordination Between SIU and Claims Reps (mean = 6.02, loading = 0.661, missing $n = 4$).[15] A scale based on these items has a reliability coefficient of alpha = 0.82.[16] (The opposite of this construct is *resistance to innovation,* which results by simply multiplying the factor scores by –1.)

If the type of office intervention determines the level of a person's receptivity to innovation, then this effect would contradict the conceptualization of the latter variable as an intrinsic characteristic of the employee. Contrariwise, the lack of a significant correlation between the intervention typology and receptivity would indicate the intrinsic, associational (and not causal) aspect of this variable. To test this assumption, the number and kinds of new computer systems being introduced can be conceptualized as an ordinal variable (ordered by the severity of the intrusion in the workplace). When there are four levels of intrusion (no systems, M2K Only, CFRD Only, both systems), its Spearman correlation of $r_s = -0.063$ with receptivity to innovation is not significant ($p = 0.387$); when there are three levels (no systems, one system, both systems) the $r_s = -0.056$, which also is not significant ($p = 0.435$).

Consequently, receptivity to innovation is best viewed as a personal characteristic of the employee that may be correlated with other personal characteristics. Hypothesis 2 suggests that employees who are favorably disposed toward innovations in general also will favor computerized fraud detection, the correlation of 0.301 ($p = 0.0001$) corroborates this. College graduates are slightly more receptive ($r = 0.15$, $p = 0.04$). Office personnel with jobs of higher rank are generally less receptive ($r = 0.19$, $p = 0.01$), as are employees with seniority at the company ($r = 0.18$, $p = 0.02$) or with seniority in the industry ($r = 0.19$, $p = 0.01$).

Indicators of *Jobs of Higher Rank* are: handling bodily injury claims (BI) rather than no fault (personal injury protection) claims (mean = 1.56, loading = 0.864, missing $n = 10$); having a college degree (mean = 0.516, loading = 0.738, missing $n = 4$); and the type of job, dichotomized as supervisors and claims specialists versus special investigators, field representatives, and claims representatives (mean = 0.592, loading = 0.721, missing $n = 7$). A scale based on these items has a reliability coefficient of 0.69.[17] Because employee attributes compose this construct, the associational (and not causal) aspect of this variable is apparent. As expected, the construct is unrelated to the ordinal variable based on type of office. When there are four levels of intrusion, the r_s of –0.010 is not significant ($p = 0.892$); when there are three levels the r_s is zero ($p = 0.996$).

Consistent with Hypothesis 3, the higher the rank of the employee's job, the higher the percentage against computerized fraud detection ($r = 0.131$, $p = 0.072$). Using the dichotomized anti-index, the percentages against are: 24.1% of the claims representatives, 28.6% of the special investigators and field employees, 37.9% of the claims specialists, and 42.9% of the supervisors (but not necessarily top managers of the offices).

In sum, from a causal point of view, the typology of office interventions portrays the different numbers and kinds of varying causal interventions (i.e., causality as an effect of an intervention), whereas the two covariates are fixed, associational characteristics of the employees that do not vary during the study period (i.e., causality as stable association).[18]

Statistical Models

The subsequent analyses apply *hierarchical linear modeling* or, synonymously, *multilevel modeling* to assess the fixed and random effects on the anti-index of the types of office contexts, controlling for the workers' attitudes toward innovation and the rank of their job—the individual-level covariates (Mason et al. 1983, 72–103; Raudenbush and Bryk 2002, 68–95; Bryk et al. 1993, 257–271; Littell et al. 1996, 260–266; 2006, 326–330; Singer 1998, 323–339). These models need to be applied here because the data may be clustered within offices and because there is variance between offices that needs to be explained. SAS's Proc Mixed can provide the appropriate estimates of effect (SAS Institute 1997a, 573–582). Since the response variable and the two covariates are clustered within offices, it is reasonable to assume that at least the values of the response variable in a given office are more similar to each other than they are to the values of that variable in another office; that is, within an office some of the observations are correlated and not independent. Consequently, this correlation violates the assumption of independent observations made by ordinary least-squares regression analysis; Proc Mixed can fix this problem. Since there may be office-to-office (i.e., between-office) variance ($\hat{\sigma}_o^2$) in the response variable that needs to be explained, this quantity should be modeled explicitly; Proc Mixed can quantify the effects of different sets of predictors on this variance.

To explain the between-office variance in anticomputerization sentiment, controlling for the effects of individual-level covariates, this chapter elaborates the effects of the types of offices by progressing from simple to more complex models. By estimating an unconditional means model with the claims office designated as random, Model 1 assesses the extent to which the employees' anticomputerization attitudes vary across the six offices; there is noticeable variance that is statistically significant (Littell et al. 1996, 135–149; 2006, 58–74). Model 2, which corroborates Hypothesis 2 and Hypothesis 3, analyzes how the two individual-level covariates affect the anti-index, whether they reduce the variance among employees within offices ($\hat{\sigma}_e^2$) and the variance between office means ($\hat{\sigma}_o^2$), and whether the two covariates should be conceptualized as fixed and not random (Littell et al. 1996, 171–187; 2006, 244–263).

With the offices designated as random and classified by the office typology, and with the two fixed individual-level covariates in the equation; to test Hypothesis 1 Models 3a through 3d examine in succession the fixed-effects of each binary-coded office type on the anti-index and on $\hat{\sigma}_o^2$. If the type of the office—the classificatory test variable—produces a large reduction in $\hat{\sigma}_o^2$, then that variable would be a cause of the between-office variance in the workers' anticomputerization sentiments (Lazarsfeld 1955b, *xi*). Model 3a assesses the fixed-effect of CFRD; Model 3b, M2K; Model 3c, One System (either CFRD or M2K but not both); and Model 3d, Both Systems. Of these, the best-fitting model includes the fixed office-level indicator for the joint implementation of both new computer systems; the between-office variance $\hat{\sigma}_o^2$ disappears when the effect of this variable is introduced into the equation. Contrariwise, the fixed-effects on the reduction of $\hat{\sigma}_o^2$ of CFRD, M2K, and One System are inconsequential. To further test Hypothesis 1, Model 4 assesses the fixed-effects of the covariates and three office contexts—Both Systems, One System, and No New Systems—these results also support Hypothesis 1 (Littell et al. 1996, 149–155; 2006, 75–81).

To test Hypothesis 4, that CFRD causes more discontent than M2K (because the former may lead to more job loss than the latter), Model 5 disaggregates the indicator for One System by including indicators for the offices with singular implementations of CFRD and M2K. Because the difference in effect between these variables is not statistically significant, Hypothesis 4 is suspect. To determine whether the rejection of this hypothesis is due to the small sample size, the data will be weighted by hypothetical numbers of employees. A parameter study based on these different numbers of employees suggests that the difference may become statistically significant if there would be considerably more cases.

Results

Table 9.2 presents the main results. In Models 1 through 3d, the six offices are designated as random; with appropriate caveats, this assumption allows their effects to be generalized to the universe of the employer's claims offices.[19] In Models 4 and 5, all of the variables are fixed. Because these variables explain all of the between-office variance, the range of inference need not be restricted to the six offices and the employees in this study, as it would be in most fixed-effects analyses.[20]

Model 1—The Baseline Unconditional Means Model

Regarding the anti-index, the unconditional means model (that is, the model with no control variables) provides baseline estimates of the between-office variance $\hat{\sigma}_o^2$, the within-office variance among employees $\hat{\sigma}_e^2$, the intraclass correlation $\hat{\rho}$, the reliability of the sample mean $\hat{\lambda}_j$ in any office j, and the value of the grand mean anticomputerization sentiment. Raudenbush and Bryk ([1992] 2002, 68–72); Singer (1998, 326–330) and Littell et al. (1996, 141–149; 2006, 64–74) define this type of model as a one-way analysis of variance (ANOVA) with random effects. Subsequently, this chapter applies

Table 9.2 Determinants of employee attitudes that are anti-computerized fraud detection

Models	Model 1, baseline model	Model 2, Level 1 only	Model 3a, CFRD only	Model 3b, M2K only	Model 3c, one system	Model 3d, both systems	Model 4, both systems and one system	Model 5, both systems, CFRD, and M2K
Covariance parameters								
σ_o^2	0.106	0.103	0.127	0.126	0.128	0.001	0.000	0.000
Standard error	0.077	0.074	0.099	0.099	0.101	0.011		
z	1.39	1.39	1.28	1.27	1.27	0.11		
Pr z	0.082	0.082	0.100	0.102	0.101	0.457		
σ_e^2	0.456	0.411	0.412	0.411	0.411	0.413	0.407	0.408
Standard error	0.047	0.043	0.043	0.043	0.043	0.043	0.042	0.042
z	9.68	9.62	9.62	9.62	9.62	9.62	9.70	9.67
Pr z	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Model fit criteria								
AIC	413.7	400.6	400.5	400.5	401.0	393.2	389.7	390.1
BIC	413.3	400.2	400.1	400.0	400.6	392.8	392.9	389.9
–2 Residual *ll*	409.7	396.6	396.5	396.5	397.0	389.2	387.7	388.1
Fixed effects								
Intercept	1.173	1.172	1.226	1.029	1.119	1.566	0.828	0.828
t	8.17	8.35	3.05	2.82	4.12	18.83	9.51	9.51
Pr > |*t*|	0.0004	0.0004	0.038	0.048	0.015	< .0001	< .0001	< .0001
For innovation		–0.202	–0.202	–0.200	–0.201	–0.187	–0.200	–0.200
t		–4.28	–4.28	–4.26	–4.26	–4.01	–4.28	–4.26
Pr > |*t*|		< .0001	< .0001	< .0001	< .0001	< .0001	< .0001	< .0001
Higher rank		0.095	0.095	0.095	0.095	0.094	0.092	0.096
t		2.04	2.04	2.05	2.04	2.04	2.01	2.08
Pr > |*t*|		0.043	0.043	0.042	0.043	0.043	0.045	0.039
Test variable			0.063	–0.174	–0.079	0.606	0.736	0.736
t			0.150	–0.430	–0.240	5.88	6.28	6.28
Pr > |*t*|			0.891	0.688	0.822	0.004	< .0001	< .0001
One system							0.233	0.397
t							2.01	1.94
Pr > |*t*|							0.046	0.054
M2K only								0.201
t								1.66
Pr > |*t*|								0.099

In Models 1, 2, 4, and 5 the Office Type is an indicator variable. In Models 3a, 3b, 3c, and 3d the Office Type is a classificatory variable. In Model 5 One System refers to the singular implementation of CFRD. When M2K is the base category, the effect of CFRD relative to M2K of 0.197 is not statistically significant, $t = 0.97$, $Pr > |t| = 0.332$. There also is no significant difference between No Systems and M2K, the effect of –0.201 is not significant, $t = -1.68$, $Pr > |t| = 0.099$.

the Raudenbush et al. (2000, 2–4, 76–79) convention that the lowest-level unit is denoted by *i*, the next-highest unit by *j*, and the next-highest unit by *k;* Littell et al. use the reverse coding of the variables.

At the employee level, the equation for the anti-index score for the i^{th} employee contained in the j^{th} office is

$$Y_{ij} = \beta_{0j} + r_{ij} \tag{1}$$

for $i = 1, \ldots, n_j$ employees who are contained in office j and $j = 1, \ldots, J$ offices; where it is assumed that $r_{ij} \sim iid N(0, \sigma_e^2)$. Equation 1 thus characterizes an employee's level of discontent with computerization in each office as the sum of the intercept β_{0j}, which is the office mean, and the random error, r_{ij}, for that i^{th} employee who is contained in the j^{th} office.

At the office level, each office's intercept (its mean level of discontent) β_{0j} is portrayed as the sum of an overall mean γ_{00} plus random deviations from that mean that u_{0j} quantifies:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{2}$$

where $u_{0j} \sim iid\, N(0, \sigma_o^2)$. Substitution of (2) into (1) creates this multilevel model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \tag{3}$$

where it is assumed that $u_{0j} \sim iid\ N(0, \sigma_o^2)$ and $r_{ij} \sim iid\ N(0, \sigma_e^2)$. That is, the anti-index score for the i^{th} employee who is contained in the j^{th} office equals the overall mean γ_{00} plus random deviations from that mean u_{0j} due to differences between offices plus the residual associated with the i^{th} employee nested in the j^{th} office.

The results for Model 1 in Table 9.2 suggest that variation in the anti-index appears at both levels. At the employee-level, the REML covariance parameter estimate for the residual provides an estimate of var $(r_{ij}) = \hat{\sigma}_e^2 = 0.456$ ($z = 9.68$). At the office-level, the REML covariance parameter estimate for the offices provides an estimate of the variance of the true office means, β_{0j} around the grand mean, which is $\hat{\sigma}_o^2 = 0.106$ ($z = 1.39$).[21] To assess the proportion of variance in Y that is between the offices, Bryk and Raudenbush (1992, 62–63) recommend the use of the intraclass correlation $\hat{\rho} = \hat{\sigma}_o^2 / (\hat{\sigma}_o^2 + \hat{\sigma}_e^2)$. Here it equals 0.189—about 19% of the variance in anticomputer discontent is between offices—but the z score for $\hat{\sigma}_o^2$ may indicate a lack of statistical significance.[22] Proc GLM's random model analysis of variance provides information to test the null hypothesis H_0: $\sigma_o^2 = 0$ vs. H_{a}: $\sigma_o^2 > 0$. The mean square for the model is 3.86 ($df = 5$) and the mean square for the error is 0.457 ($df = 187$); the F value of 8.46 level decisively rejects the null ($p < .0001$). The large sample normal approximation 95% confidence interval around $\hat{\sigma}_o^2$ ($0.106 \pm 1.96 \times 0.077 = +0.257, -0.045$) appears to include zero, but since σ_o^2 is a nonnegative number, the truncated confidence interval is $0 < \hat{\sigma}_o^2 < 0.26$. Since the number of offices is not large, the confidence interval based on the Satterthwaite approximation is more appropriate.[23] Its asymmetric bounds of $0.039 < \hat{\sigma}_o^2 < 0.96$ clearly indicate that there is variance between offices that needs to be explained.

Bryk and Raudenbush (1992, 63) suggest calculating the reliability of the sample mean in any level-2 unit j by substituting the estimated variance components into this equation:

$$\hat{\lambda}_j = \text{Reliability}(\bar{Y}_{\cdot j}) = \hat{\sigma}_o^2/[\hat{\sigma}_o^2 + (\hat{\sigma}_e^2/n_j)]. \tag{4}$$

For the six offices, the simple average of the reliability scores is $\bar{\hat{\lambda}} = 0.87$, which indicates that the sample means are quite reliable as indicators of the true office means. However, there is variation—because of its small sample size of 13 the mean for this office in which only CFRD was being introduced has the lowest reliability, 0.76, which still is sufficient. The office in which only M2K was being introduced has the highest reliability, 0.94—its sample size is 61. Model 4 combines these two offices and assesses the effects of only one implementation, either CFRD or M2K; the reliability of this combined category is $\hat{\lambda}_j = 0.85$.

The grand mean discontent with computerized fraud detection is 1.17; this is the average office-level anti-index score based on data for these six claims offices.

Model 2—Adding Two Fixed Level-1 Covariates

Before assessing the effects of the office contexts, it may be best to let the employee-level variables explain all that they can (Sewell and Armer 1966). At level-1, the employee-level, the equation for the anti-index score for the i^{th} employee nested in the j^{th} office is now

$$Y_{ij} = \beta_{0j} + \beta_{1j}Z_{1ij} + \beta_{2j}Z_{2ij} + r_{ij}, \tag{5}$$

where (Raudenbush et al. 2000, 76–78):

$\beta_{\mathrm{q}j}$ (q = 0, 1, . . . ,Q) are employee-level coefficients;
$Z_{\mathrm{q}ij}$ is a fully standardized employee-level predictor q for employee i nested in office j;
r_{ij} is the employee-level random effect; and
σ_e^2 is the variance of r_{ij}, the employee-level variance.
Again assume that $r_{ij} \sim iid\ N(0, \sigma_e^2)$.

The office-level equations for the intercept and the two regression coefficients, if the employee-level variables were assumed not to be fixed, but each to have a random component, would be:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j}\\ \beta_{1j} &= \gamma_{10} + u_{1j}\\ \beta_{2j} &= \gamma_{20} + u_{2j}\end{aligned} \tag{6}$$

where:

γ_{qs} (q = 0, 1,. . . , S_q) are office-level coefficients; and
$u_{\mathrm{q}j}$ is an office-level random effect.

But, analyses of these data indicate that models that include either (or both) of the random components u_{1j} and u_{2j} do not estimate well (the gamma matrix is not positive definite). It thus is best to fix these variables by eliminating u_{1j} and u_{2j} from these equations and then substituting the reduced expressions into Eq. 5 to obtain this equation for Model 2:

$$Y_{ij} = \gamma_{00} + \gamma_{10}\ Z_{1ij} + \gamma_{20}\ Z_{2ij} + u_{0j} + r_{ij} \tag{7}$$

where $u_{0j} \sim iid\ N(0,\ \sigma_o^2)$ and $r_{ij} \sim iid\ N(0,\ \sigma_e^2)$.[24]

Compared with Model 1, these additional orthogonal variables, whose effects differ significantly from zero, reduce $\hat{\sigma}_e^2$ from 0.46 to 0.41, explaining about 10.9% of the employee-level variance in the outcome; but these variables do not change $\hat{\sigma}_o^2$. Since the means of Z_{1ij}, Z_{2ij}, u_{0j}, and r_{ij} are zero, when (7) is evaluated at these means, the intercept should equal the grand mean, and it does; it is 1.17.

The results for Model 2 support both Hypothesis 2 and Hypothesis 3: Apropos Hypothesis 2, a worker with a factor score of one for receptivity to innovations, compared with a worker with a factor score of zero, will have less anticomputerization sentiment, by about –0.20 units on the anti-index ($t = -4.28$). Apropos Hypothesis 3, a worker with a factor score of one for jobs of higher rank, compared with a worker with a factor score of zero, will have more anticomputerization sentiment, by about +0.10 units on the anti-index ($t = 2.04$). The Akaike (i.e., AIC) and Schwarz (i.e., BIC) indices of model fit clearly show that Model 2 fits better than Model 1—here smaller positive numbers indicate a better fit than do larger positive numbers.[25]

Model 3a Through 3d—Adding One Office-Level Fixed Classificatory Variable

In this set of models, the two individual-level covariates are entered into the equation first, and then one indicator for the type of office intervention—respectively, CFRD Only, M2K Only, One System, and Both Systems. The analyses focus on the extent to which each indicator variable reduces the size of the between-office variance $\hat{\sigma}_o^2$. If an OfficeType indicator causes $\hat{\sigma}_o^2$ to disappear, then that indicator explains why the workers were against computerized fraud detection. The OfficeType indicator can be conceptualized in two ways. The simplest conceptualization views the type of office as just an office-level characteristic analogous to the level-2 mean socioeconomic status in the example of Bryk and Raudenbush (1992, 64–66) and Singer (1998, 330–333). The second conceptualization classifies the levels of the random offices (the random treatments) by using characteristics of the levels, namely, the OfficeType. When the levels of the random offices are classified into the categories of the OfficeType, the categories (i.e., groups of offices) are considered to be levels of a fixed effect. In the resulting mixed model, the fixed effects correspond to the means of the newly formed groups of offices and the random effects are the levels of the random effect nested within the levels of the fixed effects (Littell et al. 1996, 149; 2006, 75). In the data analyzed here, the two conceptualizations produce identical fixed-effect parameter estimates for the predictors, but the degrees of freedom for the fixed OfficeTypes differ, as do their associated *p* values. When there are random office effects to explain, the second conceptualization is preferred to the first because that model explains part of the variability in the levels of the random offices by using characteristics of those levels. When the random effects have been explained the first conceptualization is preferred to the second because that model is more parsimonious.

The officetype indicator model

This model takes as given the employee-level Eq. 5 above. The office-level expressions for the three coefficients in that Eq. 5 are:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}\mathit{OfficeType}_{1j} + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\mathit{OfficeType}_{1j} + u_{1j}\\ \beta_{2j} &= \gamma_{20} + \gamma_{21}\mathit{OfficeType}_{1j} + u_{2j}\end{aligned} \tag{8}$$

where (Raudenbush et al. 2000, 2–3):

γ_{qs} (q = 0, 1,. . . , S_q) are office-level coefficients;
OfficeType$_{1j}$ is a class variable that provides the office-level binary predictors; and
$u_{\mathrm{q}j}$ is an office-level random effect.

However, there are no significant two-way interactions between OfficeType and the two employee-level variables (Z_{1ij} and Z_{2ij}), so the middle term in the last two expressions (the expressions for β_{1j} and β_{2j}) drop out. Also, given that the employee-level variables are assumed to be fixed, their random elements u_{1j} and u_{2j} are assumed to be zero and these two equations simplify to $\beta_{1j} = \gamma_{10}$ and $\beta_{2j} = \gamma_{20}$. Substitution of these two expressions and that for β_{0j} into Eq. 5 results in this equation for Models 4 and 5:

$$Y_{ij} = \gamma_{00} + \gamma_{01}\mathit{OfficeType}_{1j} + \gamma_{10}\ Z_{1ij} + \gamma_{20}\ Z_{2ij} + u_{0j} + r_{ij} \tag{9}$$

where $u_{0j} \sim iid\ N(0, \sigma_o^2)$ and $r_{ij} \sim iid\ N(0, \sigma_e^2)$.

The Proc Mixed SAS code for this model has this form:

```
proc mixed covtest ratio;
        class sixoffices officetype;
        model anti-index = Zone Ztwo officetype/solution;
        random sixoffices/solution;
        lsmeans officetype/pdiff adjust = Bon;
run;
```

Note that the random offices are not classified by the type of office.

The officetype classification model

Because the random offices will be grouped into types of offices, three subscripts are necessary to describe the model. Paralleling Eq. 5 above for the employee-level, the equation for the anti-index score for the i^{th} employee nested in the j^{th} office, which is classified by the k^{th} type of office is now

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}\, Z_{1ijk} + \beta_{2jk}\, Z_{2ijk} + r_{ijk}, \tag{10}$$

where:

$\beta_{\mathrm{b}jk}$ ($\mathrm{b} = 0, 1, \ldots, B$) are employee-level (level-1) coefficients;
$Z_{\mathrm{b}ijk}$ is an employee-level fully standardized predictor for employee i nested in office j which is classified by OfficeType k;
r_{ijk} is the employee-level random effect; and
σ_e^2 is the variance of r_{ijk}, that is the employee-level variance.

This model assumes that the random term $r_{ijk} \sim iid\ N(0, \sigma_e^2)$.

The office-level expressions for the three coefficients in the employee-level Eq. 10 above are

$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \gamma_{01\mathrm{k}} \mathit{OfficeType}_{1jk} + u_{0j(k)} \\ \beta_{1jk} &= \gamma_{10k} + \gamma_{11\mathrm{k}} \mathit{OfficeType}_{1jk} + u_{1j(k)} \\ \beta_{2jk} &= \gamma_{20k} + \gamma_{21\mathrm{k}} \mathit{OfficeType}_{1jk} + u_{2j(k)}, \end{aligned} \tag{11}$$

where:

γ_{bqk} ($\mathrm{q} = 0, 1, \ldots, Q_b$) are office level coefficients;
$\mathit{OfficeType}_{\mathrm{q}jk}$ is an office-level class variable predictor; and
$u_{\mathrm{b}j(k)}$ is an office-level random effect.

However, given that there are no significant two-way interactions between OfficeType and the two employee-level variables (Z_{1ijk} and Z_{2ijk}), the middle term in the last two expressions (the expressions for β_{1jk} and β_{2jk}) drop out. Also, given that the employee-level variables are assumed to be fixed, their random elements $u_{1j(k)}$ and $u_{2j(k)}$ are assumed to be zero, and these two equations simplify to $\beta_{1jk} = \gamma_{10\mathrm{k}}$ and $\beta_{2jk} = \gamma_{20\mathrm{k}}$. Substitution of these two expressions and that for β_{0jk} into Eq. 10 results in this equation for Models 3a through 3d in which the office random effects are classified by the type of office:

$$Y_{ijk} = \gamma_{00k} + \gamma_{01k} \mathit{OfficeType}_{1jk} + \gamma_{10k}\, Z_{1ijk} + \gamma_{20k}\, Z_{2ijk} + u_{0j(k)} + r_{ijk} \tag{12}$$

where $u_{0j(k)} \sim iid\ N(0, \sigma_o{}^2)$ and $r_{ijk} \sim iid\ N(0, \sigma_e{}^2)$.

The PROC MIXED SAS code for this model has this form:

```
  proc mixed covtest ratio;
          class sixoffices officetype;
          model anti-index = Zone Ztwo officetype/solution;
          random sixoffices(officetype)/solution;
          lsmeans officetype/pdiff adjust = Bon;
run;
```

The six random offices are now classified by the type of office.

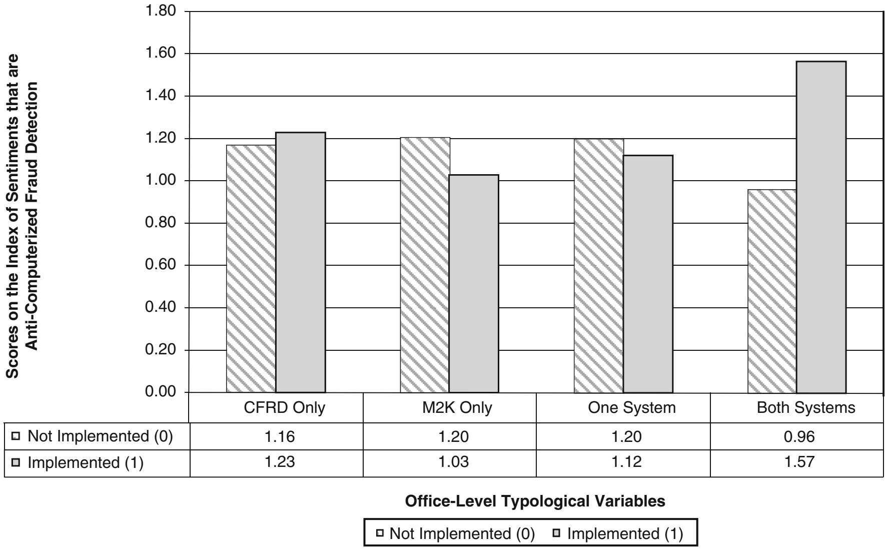
The effects of singular implementations

When the OfficeType classification indicators for singular implementations of CFRD, M2K, or One System define the OfficeType in Eq. 12, that is $k = 1$ or $k = 0$ types of offices, these office contexts do not noticeably reduce $\hat{\sigma}_o^2$ nor do they directly effect anticomputerization sentiment; see the estimates for Models 3a through 3c in Table 9.2. Model 3a provides an answer of "No" to the original evaluative research question; namely, "Did the implementation of CFRD engender sentiment that was anticomputerized fraud detection?"[26] None of these singular implementations explain any of the baseline office-level variance; in fact, the controls for these variables slightly increase $\hat{\sigma}_o^2$.

The effects of joint implementations

When in Model 3d the OfficeType indicator is Both Systems, it obliterates $\hat{\sigma}_o^2$ by reducing it from .106 to .001 ($z = .11$, $p = .46$); the joint implementation of the two computer systems explains about 99% of the office-level variance in the outcome. Both fit indices indicate that Model 3d fits better than any of the other models estimated thus far. Holding constant the covariates in this system, if an office is changed from Both Systems = 0 to Both Systems = 1, then the anti-index score increases by 0.606 (0.808, 0.404). To check the calculations, when Eq. 12 is evaluated at the mean values of its elements, their sum $(1.566 \times 1 + -0.6064 \times 0.66)$ equals the grand mean, 1.17, as it should. When the six offices are designated as random and grouped by Both Systems, the addition of another office-level grouping indicator adds very little to the explanation of $\hat{\sigma}_o^2$. Consequently, the joint implementation of the two new computer systems is the major cause of the workers' discontent with computerized fraud detection.[27]

Figure 9.1 depicts the mean office-level anti-index scores when the two covariates are evaluated at their mean of zero and the various OfficeType grouping indicators are set to zero (for not implemented) or set to one (for implemented). When the OfficeType is CFRD Only, the difference in the anti-index score will be only slightly higher when CFRD is implemented (1), compared to when it is not



	CFRD Only	M2K Only	One System	Both Systems
□ Not Implemented (0)	1.16	1.20	1.20	0.96
□ Implemented (1)	1.23	1.03	1.12	1.57

Fig. 9.1 Fixed effects on the anti-index of different office-level typological variables, fixed employee-level factors controlled

implemented (0); the difference of .063 is not statistically significant (t = 0.150). When the OfficeType is M2K Only, the difference in the anti-index score will be slightly lower when M2K is implemented (1) compared to when it is not implemented (0); the difference of –0.174 is not statistically significant (t = –0.430). When the OfficeType is One System = 1, the difference in the anti-index score will be slightly lower than when One System = 0; the difference of –0.079 is not statistically significant (t = –0.24). However, when the OfficeType is Both Systems = 1, the difference in the anti-index score will be very much higher than when Both Systems = 0, the difference of 0.606 is very statistically significant (t = 5.88, $p < .0001$). Thus, with the six offices designated as random, the best model includes only one fixed office-level variable, the classificatory variable that contrasts Both Systems (scored 1) to all other categories (scored 0).[28] When another indicator of office context is also included in the model the solutions for the random effects are problematical (the gamma matrix is not positive definite).

Model 4—All Variables Fixed, Three Office Contexts

Because there is an additional fixed contextual variable—One System (either M2K or CFRD)—and because the random term for u_{0j} is now zero, applying Eq. 9 the regression equation for this model is:

$$Y_{ij} = \gamma_{00} + \gamma_{01}\text{Both Systems}_{1j} + \gamma_{02}\text{One System}_{2j} + \gamma_{10}\, Z_{1ij} + \gamma_{20}\, Z_{2ij} + r_{ij} \quad (13)$$

where $r_{ij} \sim iid\ N(0,\ \sigma_e^2)$. At the office level, the addition of One System to the equation changes the reference category so that it now includes only the two offices that have no new computer systems; the intercept = 0.828. This change increases the relative effect of Both Systems on the anti-index to 0.736 (0.966, 0.507). The effect of the implementation of One System is a barely statistically significant 0.233 (0.461, 0.006).[29] Thus, for the universe of the employer's claims offices, the predicted mean levels of discontent will be as follows: in offices with Both Systems discontent will be = 0.828 + 0.736 = 1.564, with One System it will be = 0.828 + 0.233 = 1.061, and with No New Systems it will be = 0.828. These effects support Hypothesis 1: offices experiencing the joint implementation of two new computer systems will exhibit more employee discontent with computerized fraud detection than offices experiencing the implementation of only one new computer system, which in turn will exhibit higher scores on the anti-index than offices experiencing no new implementations. The pairwise differences between the anti-index least-squares means follow the expected pattern and are significant: Both Systems engenders more discontent than One System by 0.503 ($t = 4.57$, $p < .0001$); Both Systems engenders more discontent than No New Systems by 0.736 ($t = 6.28$, $p < .0001$); One System engenders more discontent than No New Systems by 0.233 ($t = 2.01$, $p = 0.046$; with Bonferroni adjustment, $p = 0.137$).[30] Also, as expected, when (10) is evaluated at the mean values, the sum of the intercept (0.828) plus the sum of the products of the effects of the indicator variables and their means (0.736 × 0.3402 + 0.233 × 0.3814) equals the grand mean, 1.17.

Model 5—All Variables Fixed, Four Office Contexts

Equation 14 disaggregates the One System indicator variable of Eq. 13 into its two components. With controls for receptivity to innovation and for rank in the office, the resulting model analyzes the differences between implementation of Both Systems, CFRD Only, and M2K Only, all relative to offices with no new systems:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{01}\text{Both Systems}_{1j} + \gamma_{02}\text{CFRDOnly}_{2j} + \gamma_{03}\text{M2KOnly}_{3j} \\ &\qquad + \gamma_{10}\, Z_{1ij} + \gamma_{20}\, Z_{2ij} + r_{ij} \end{aligned} \tag{14}$$

Again assume that $r_{ij} \sim iid\ N(0,\ \sigma_e{}^2)$.[31]

The results for Model 5 are nearly identical to those of Model 4 except for the effects of these two new contextual variables: The effect of implementation of CFRD Only, relative to the implementation of No New Systems is 0.397 (0.798, –0.004) and just fails to obtain significance. The effect of implementation of M2K Only is 0.2005 (0.437, –0.036), also not significant. When M2K Only is used as the base category, there is no significant difference between it and CFRD Only—the estimate is 0.197 ($t = 0.97, p = 0.332$).[32] This lack of a significant effect contradicts Hypothesis 4 that assumes that CFRD is perceived as a greater threat to the job than

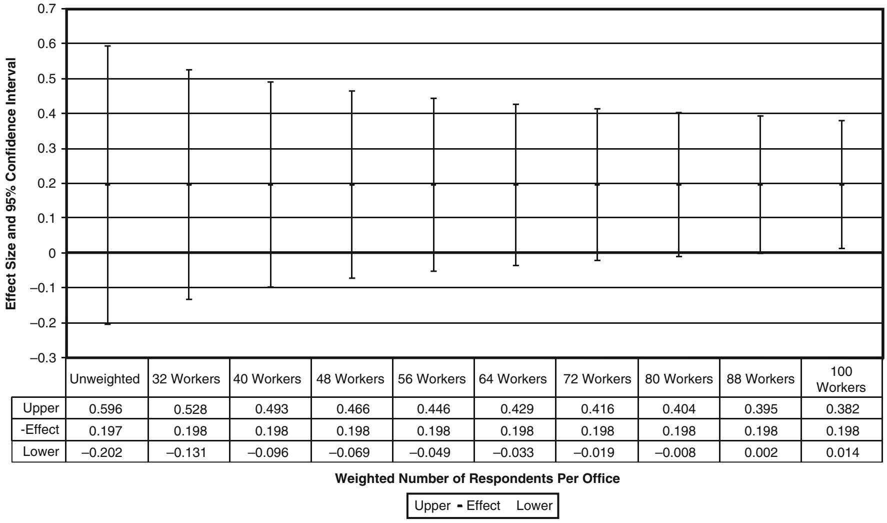
M2K. It is appropriate; therefore, to combine these two categories to form One System, as was done in Model 4. To check these calculations, when the three office contexts are evaluated at their means, the sum of their contributions and the intercept should equal the grand mean, and it does ($0.828 \times 1 + 0.736 \times 0.3402 + 0.397 \times 0.06701 + 0.2005 \times 0.3144 = 1.17$).

Which model is best? Models 2 through 5 indicate that claims workers with higher-ranked jobs and workers who generally disapprove of innovations will have higher levels of discontent about computerized fraud detection. Among the better models that assess the effects of the types of interventions, Model 4 is preferred to Model 5 because the latter includes two effects that are not statistically significant.[33] Model 3d has one more free parameter than Model 4, but the fit statistics for 4 are slightly better. Model 3d designated the six offices as random and classified them by Both Systems; SAS estimated the effects without a problem. Thus, its key result can be generalized with little equivocation to all of the offices in the population—the simultaneous implementation of the two new computer systems will create much discontent with computerized fraud detection. When Model 4 designates the six offices as random, because the fixed variables explain all of the between-office variance, the gamma matrix is not positive and definite. However, the fixed-effects estimates are the same whether or not the offices are designated as random. Consequently, it can be assumed that the key result of Model 4 can also be generalized to the universe of the employer's claims offices—the simultaneous introduction of the two new computer systems will engender more discontent than the introduction of only one new system. Because its key finding is richer than that of Model 3d, Model 4 may be best.

The Effect of Sample Size on Hypothesis 4

Because CFRD poses a possible threat to job security, whereas M2K does not, Hypothesis 4 predicts that CFRD will have a stronger effect on the anti-index than M2K. As reported earlier, when M2K Only is used as the base category, the effect on the anti-index of CFRD Only was larger by about 0.197 ($t = 0.97, p = 0.332$) but was not statistically significant. Because this lack of significance could be due to the small sample size of 13 respondents in the Florida target office, compared with the relatively large sample size of 61 in the M2K Only comparison office, a parameter study will be conducted in which the number of respondents in all offices are first equalized and then increased systematically until the difference between CFRD and M2K achieves statistical significance. Because SAS used 193 observations to estimate each model, this number was divided by 6 offices to obtain 32.16666667 observations per office as the desired number. This number was then divided by the actual number of respondents in each of the six offices to obtain the six weights that equalize the number in the six offices. Weighting the data for an office by its quotient, SPSS was used to estimate the fixed effects—Proc Mixed (SAS version 6.12) did not readily allow a frequency statement for simulated replication; current versions allow this. Now the effective sample size in the regression was about 32 in each office. To boost the sample size to 40, eight was added to 32.16666667 and the procedure was repeated. To boost the sample size to 48, 16 was added to 32.16666667 and the procedure repeated, etc. The parameter study ranges from about 32 respondents in each office to about 100 respondents in each office, in increments of eight respondents.

Figure 9.2 summarizes the results by reporting the effects and 95% confidence intervals first for the actual data and then for the weighted data. The regression model is equivalent to Model 5—all variables fixed—but now the base category is the office with M2K Only. No New Systems, the former base category, now appears in the regression equation explicitly. Because each office has an equal number of respondents, the effect of CFRD Only relative to M2K Only remains constant at 0.198, but the standard errors decline as the number of hypothetical respondents increases. From 32 through 80 workers per office the confidence interval of the effect includes 0. At 88 workers per office the effect just attains statistical significance ($t = 1.984$, $p = 0.048$) and then remains significant thereafter. Because this number of simulated respondents is larger than the number of fraud claims workers in an actual office, most likely if the actual sample size was equalized in each office and increased, then the difference in effect between CFRD and M2K would not attain statistical significance. However, the counternull effect size is 0.40 units on the anti-index and this value is as likely to be true as a null of zero effect.



	Unweighted	32 Workers	40 Workers	48 Workers	56 Workers	64 Workers	72 Workers	80 Workers	88 Workers	100 Workers
Upper	0.596	0.528	0.493	0.466	0.446	0.429	0.416	0.404	0.395	0.382
-Effect	0.197	0.198	0.198	0.198	0.198	0.198	0.198	0.198	0.198	0.198
Lower	−0.202	−0.131	−0.096	−0.069	−0.049	−0.033	−0.019	−0.008	0.002	0.014

Fig. 9.2 A parameter study of how the number of respondents per office affects the statistical significance of the difference in effect on the anti-index of CFRD Only compared to M2K Only, fixed effects estimates

Discussion

Summary

For assessing the effects of a computerized fraud detector (CFRD) on employee morale, the initial study design matched three target offices with three comparison offices. Using the corporate intranet, questionnaires were administered to the personnel in all six offices. After the data were gathered, the evaluators discovered the impact of a second, administrative computer system (M2K) that was being installed in two target offices and in a comparison office, thereby confounding the study design. To assess the effect of CFRD, it was necessary to redesign the study as follows: two offices had both CFRD and M2K, one office had CFRD Only, another office had M2K Only, and two offices had No New Systems. The simultaneous implementation of two new systems engendered strong dislike of computerized fraud detection, especially by employees who generally oppose innovations and those who have jobs of higher rank. The two new computer systems caused more employee discontent than one, which in turn caused more discontent than no new systems.

Interpretations

Technological intrusions in the workplace may be a hassle because employees must learn new procedures, which make obsolete old habits, skills, and ways of performing work. Thus, the hassle of two intrusions causes more discontent than the hassle of only one, which causes more discontent than the status quo. Two characteristics of the employees affect their level of discontent. Some are predisposed cognitively to be more receptive to innovations than others: college graduates are more receptive; supervisors and employees with seniority in their jobs and in the insurance industry are less receptive. Because they may have more to lose in terms of the substantive complexity of their work and their job security, workers with jobs of higher rank—the supervisors and claims specialists—exhibit more discontent with computerized fraud detection than other workers. Some top managers want to reduce costs by enabling the less expensive claims representatives to perform the work of the more highly skilled claims specialists. Since CFRD poses a greater threat to job security than M2K, it follows that the office with the former would have a higher level of discontent than the office with the latter. However, a parameter study indicates that this difference attains statistical significance only when the hypothetical number of employees in an office exceeds the number in a typical office. Consequently, employment security is not a key interpretative factor. The threat of deskilling, which interprets the effect of jobs of higher rank, no doubt engenders some discontent with computerized fraud detection. But its effect is not as salient as the effect of the hassle due to the implementation of two new computer systems in the work place, and the effect of an employee's receptivity to innovation, a prior predisposition.

These findings bear on several disagreements among social theorists. George Homans (1975, 56) questioned Peter Blau's (1960, 179–180) conception of a structural effect because it was based in part on an aggregation of individual predispositions. Blau illustrated a structural effect by noting that people who live in neighborhoods composed of many bigots are more likely to discriminate regardless of their own level of prejudice. The interaction between the composition of the group and their predispositions defined the structural effect on their willingness to discriminate. Homans explained this effect in terms of the rewards (social acceptance) less prejudiced people would receive from their bigoted neighbors when they would agree to discriminate. In this way, Homans's use of methodological individualism enabled him to question the existence of structural effects. This study works against Homans's interpretation because the number of new computer systems is a global characteristic of the claims offices, one not based on the aggregation of the workers' predispositions. The effect of these new tools on reducing the variance in anticomputerization sentiment that is between the offices is a structural effect that cannot be readily explained by methodological individualism. Moreover, this effect is cultural, by the introduction of the new computer systems management was attempting to change the culture of the work place. This cultural effect on worker discontent dominated the effects of the social variable (the rank of the worker's job) and the personality variable (the worker's attitude toward innovation). In this analysis, the change in the tools was the determining variable. It thus supports Parsons (1966, 113), who stated that: "In the sense, and *only* that sense, of emphasizing the importance of the cybernetically highest elements in patterning action systems, I am a cultural determinist, rather than a social determinist."

Generalizations

The analyses applied hierarchical linear modeling to assess the determinants of the variance in discontent that was between the office means, $\hat{\sigma}_o^2$ and the variance in discontent among employees within offices, $\hat{\sigma}_e^2$. With the offices designated as random, and with fixed controls for receptivity to innovation, jobs of higher rank, and the classificatory indicator variable for Both Systems, the latter variable almost completely explains the $\hat{\sigma}_o^2$. This may imply that for the universe of this employer's claims offices, the simultaneous introduction of CFRD and M2K in an office will engender sentiment that is anticomputerized fraud detection. Moreover, employees of higher rank will exhibit more anticomputerization sentiment, and employees who are generally receptive to innovation will exhibit less. The singular implementation of CFRD, M2K, or One System (CRFD or M2K) will generate much less discontent with computerized fraud detection than will their joint implementation.

Recommendations

Employers, who desire to minimize employee discontent with new expert systems like CFRD, should carefully install one new computer system at a time, encourage employees to be receptive to innovations by encouraging lifetime-learning programs, enhance the substantive complexity of work by allowing employees to make judgments and to think, and reduce job insecurity by using attrition and not layoffs to reduce the number of employees.

. . .

Following a general strategy of quantification of the relative effects of several assumed causal variables on a response (Coleman 1964, 189–240; 1981), the chapters in this Part developed contextual analyses by introducing typologies in multilevel models. To formulate policy options, Chapter 6 explicated how each of its three subsequent chapters addresses a social problem by combining theory, study designs, measures, and results from multilevel contextual analyses. Chapter 7 grouped the nations of the world into their regions and found that typologies of democracy and emancipatory employment accounted for the regional disparities in human development. Chapter 8 grouped longitudinal measures on European countries and found that a county's Muslim and Jewish population category influences its predisposition to anti-Jewish violence. Chapter 9 grouped claims workers into their offices and found that the number of new computer systems being installed in an office engendered their discontent. All of these chapters focused on quantifying the relative sizes of the fixed covariates in the multilevel model and determining which typological variables accounted for the variance between the level-2 units.

The next three chapters composing Part 3 illustrate causality as an effect of an intervention. They present longitudinal evaluative research studies in which aggregated performance measures on students at a point in time are key level-1 outcomes, schools are the level-2 variables, and the quasiexperimental condition of the schools—target or comparison—is the classification typology. Although the designs of these chapters control for a number of fixed covariates, their analytic focus is on quantification of the causal effect of the manipulated treatment (Rubin 1974) using difference-in-differences (DIDs). Three key differences are: (1) the difference between the pre to post period means in the target group; (2) the difference between the pre to post period means in the comparison group; and (3) the difference between the first two differences, the latter difference estimates the effect of the treatment.

Acknowledgements In addition to the author, the evaluation team included Herbert Weisberg and Glenn Wolf. Weisberg was responsible for the implementation of CFRD; Wolf represented the insurance company and provided liaison with the claims offices; Weisberg and Wolf primarily focused on the evaluation of the cost savings due to CFRD; the author primarily focused on the development of the survey questionnaire, creation of the database, and the analysis of employee attitudes. Major portions of this chapter appeared in the *Evaluation Review* (vol. 26, no. 1, 3–39; © 2002 Sage Publications). The comments of Richard Berk, the journal editor, were very helpful. The author thanks the publisher for allowing him to reuse this material in this book. The views expressed herein are the author's own and do not necessarily reflect those of his colleagues or those of any organization.

Endnotes

[1] Numerous aspects of Kohn's research program on social structure and personality are reviewed by Smith (2008b, 48–49).

[2] Pearl's ([2000], 2009, second edition, 157–159) $do(x)$ operator clarifies this shortcoming. It signifies doing $X = x$ in an ideal, closely matched quasi-experiment in which the manipulation X is set to $x =$ CFRD and no other variable is changed. The causal effect would be $\delta = y_{\text{CFRD}} - y_{\text{NoCFRD}}$. But in some offices there is another variable being manipulated, Z is set to $z =$ M2K. Thus, the actual quasi-experiment is denoted as $do(x, z)$ and not $do(x)$.

[3] The insurer's claims offices are thought to be of four types. The sample of employee data obtained from an office type is assumed to be a random sample of all employee samples that could be obtained from the office types. Consequently, under this assumption analysis of variance and its generalization in Proc Mixed can be used to make the desired comparisons. For explication of this kind of assumption see Moore and McCabe (1989, 714 –715, 719).

[4] The skeptical reader may accept the hypothesis but reject the postulated underlying mechanism. Field observations and the insights of a professional writer on the topic of anti-computer attitudes provide some anecdotal evidence that supports the mechanism this chapter assumes. At a meeting the author attended in one target office the verbal statements of the coordinator of the implementations of the two computer systems indicated that she was distraught and overwhelmed by her responsibilities—two new systems were too many. The professional writer asks (Mello 2000, November 19):

> Why does technology throw otherwise adept, skillful, and talented people into a state of helpless confusion? Social scientists have spent lots of research money trying to answer that question. Feelings of inadequacy can arise when technology threatens to render useless a skill set someone spent a lifetime cultivating, they say. Those feelings of inadequacy are reinforced by the head-spinning pace of technological change. Even before a person can master one technological intrusion into their workspace, it's altered or replaced with something new.

[5] Although a person's long-term voting pattern does affect the person's party identification through a learning process, there is little if any evidence that an employee's negative attitude toward computerized fraud detection weakens receptivity to innovation, the more general predisposition. In fact, it may have the opposite effect: one supervisor who thought that CFRD was a waste of money and time viewed that such innovations as lists of doctors and lawyers involved with past suspicious claims would be more helpful.

[6] The author derived this hypothesis directly from his field observations and interviews at a target office. The dislike of the fraud detector was evident among the supervisors of the claims representatives. One supervisor clearly stated that she could detect fraudulent claims much better than any computer. She also stated that she devised some hypothetical data that confused CFRD. Many of the lower-rank workers, however, were willing to use the system, once they were confident that it would not cause their computer to crash.

[7] Some top managers viewed CFRD as an innovation that might reduce the expense of detecting fraud by enabling the less skilled claims representatives to accomplish the work presently done by the more skilled claims specialists. The workers with higher-ranked jobs may have been aware of CFRD's possible threat to their job security. M2K did not pose such a threat.

[8] Regarding CFRD usage, the Bonferroni post hoc tests indicate that offices with both systems and the office with CFRD Only have about the same usage and both have higher usage than the office with M2K Only and the offices with neither system. Regarding unfavorable attitudes toward CFRD, the offices with Both Systems are more unfavorable than the office with CFRD Only; both of these types of offices have more unfavorable attitudes than the other two types of offices. The index of unfavorable attitudes assessed here is as follows: Used CFRD and dislike it (2), used CFRD and do not dislike it (1), have not used CFRD and have no opinion about it (0). When there are three office types the results are strictly ordinal for both usage and dislike of CFRD: both

systems > one system > neither new system. With the Bonferroni corrections these results hold at the $p = 0.01$ level of significance.

[9] The groups in an analysis of variance or in its Proc Mixed generalization should come from populations with equal variances. When four or three types of offices are the groups, the trichotomous anti-index passes the Levene test for homogeneity of variance. Across the four types of offices the variances in the anti-index are about equal (Levene statistic = 0.841, $p = 0.47$) and the ratio of the highest and lowest sample standard deviations is 1.18, much less than the rule-of-thumb cut-off value of 2 for conducting an analysis of variance (Moore and McCabe 1989, 722). The pattern of results is similar when there are three office types: the Levene statistic is 1.23 ($p = 0.29$) and the ratio of the highest and lowest sample standard deviations is 1.25. For either set of offices, the dichotomous index does not pass the Levene tests ($p = 0.000$) but the ratios of the highest and lowest sample standard deviations are both 1.38, below the cut-off value of 2.

[10] The skewness of the distribution of the anti-index scores is within bounds for the normality assumption of the analysis of variance; the ratio of the skewness of –.284 to its standard error of .173 (which equals –1.64) falls within the 95% confidence bounds for normality. The negative sign indicates a longer left tail than expected for a normal distribution. The kurtosis ratio of –1.108 to its standard error of .345 (which equals –3.21) falls outside the 95% confidence bounds for a normal distribution. The negative kurtosis indicates shorter tails than expected for a normal distribution. Consequently, both the Kolmogorov-Smirnov and Shapiro-Wilk tests reject the null hypothesis of a normal distribution. Neither the square root nor the log-to-base-e transforms improve the normality tests. Given that the analysis of variance (and its generalization in Proc Mixed) are reasonably robust regarding violations of normality that are not severe (Moore and McCabe 1989, 721–722, 726), for simplicity the untransformed scale will be used in the analyses. Using the untransformed scale avoids the problem of recovering the equivalent of the untransformed results from the analyses based on the transformed data (Manning 1998; Manning and Mullahy 1999).

[11] When there are four office types, for the dichotomized anti-index the null hypothesis that the linear constant is zero (i.e., there is no linear effect) is rejected (weighted linear $F = 43.4$, $p = 0.000$). The alternative hypothesis of a linear effect can be accepted (deviation $F = 1.36$, $p = 0.261$; the quadratic deviation $F = 0.056$, $p = 0.814$); the office means are located on a straight line. For the trichotomous anti-index the linearity test results are similar (weighted linear $F = 41$, $p = 0.000$; deviation $F = .812$, $p = 0.45$; and quadratic deviation $F = 0.086$, $p = .77$). When there are three office types, the null hypothesis that the linear constant is zero is rejected for the dichotomous measure (weighted linear term $F = 39.1$, $p = 0.000$) and for the trichotomous measure (weighted linear term $F = 37.6$, $p = 0.000$)—the increase is monotonic but not linear. The observations are not located on a straight line: the mean discontent for offices with both new systems is significantly above a straight line drawn through the first two points (no new systems and only one new system). For the dichotomous anti-index the deviation F is 6.14 ($p = 0.014$) and for the trichotomous anti-index the deviation F is 4.3 ($p = 0.039$).

[12] For the dichotomized anti-index, the Bonferroni post hoc tests assess the following pairwise mean differences: The mean discontent in offices with both new systems is higher than that in offices with M2K Only (mean difference = 0.43, $p = 0.000$), No New Systems (mean difference = 0.48, $p = 0.000$), and CFRD Only (mean difference = 0.30, $p = 0.17$—not significant). The level of discontent in offices with no new systems is lower than that in the offices with only one new system: for M2K Only the mean difference is –0.06 ($p = 1.000$) and for CFRD Only –0.19 ($p = 1.000$). For the trichotomized anti-index, the pattern of the results is the same: The mean discontent in offices with both new systems is higher than that in offices with M2K Only (mean difference = 0.61, $p = 0.000$), No New Systems (mean difference = 0.74, $p = 0.000$), and CFRD Only (mean difference = 0.42, $p = 0.275$—not significant). The level of discontent in offices with no new systems is lower than that in the offices with only one new system: for M2K Only the mean difference is –0.13 ($p = 1.000$) and for CFRD Only –0.32 ($p = 0.86$)—these differences are not significant. These findings indicate that the offices with both new systems have levels of discontent that are higher than the levels in all of the other offices; these other offices could be pooled.

[13] The Bonferroni pairwise differences also indicate no significant difference in mean discontent between offices with CFRD Only and M2K Only. For the dichotomized anti-index the mean difference is 0.13 ($p = 1.000$) and for the trichotomous measure, 0.18 ($p = 1.000$).

[14] When the design variables are also evaluated at their means, then the sum of the intercept plus the sum of the product of the coefficients for the design variables and their respective means will equal the grand mean.

[15] For simulation studies bearing on the feasibility of assigning equal-interval scales to ordinal variables see H. T. Reynolds (1974), especially pages 405–417. In general, Reynolds (following Blalock) recommends not collapsing variables to form dichotomies and trichotomies but to keep the full range of variability, in this case the seven categories of each of the six variables included in the factor analysis.

[16] The measure of receptivity to innovation has a skewness ratio of –4.8; there is a long tail to the left. The kurtosis ratio of 1.8 is in the bounds for a normal distribution. The tests for normality reject the hypothesis that the distribution is normal ($p = 0.000$). However, the Q-Q plot indicates that the vast majority of the points are close to the expected normal line, the deviations are at the extremes.

[17] The measure of jobs of higher rank has a skewness ratio of –1.40, which is in bounds for the assumption of normality. The kurtosis ratio of –4.023 indicates that the distribution is decidedly not normal, as do the tests for normality ($p = 0.000$) and the Q-Q plot. The histogram and stem and leaf plots suggest that there are four clusters of peak values.

[18] Pearl's ([2000] 2009, second edition, 157–159) $do(x)$ operator provides a language for portraying these causal assumptions. Most simply, $y = \mathrm{a} + \mathrm{b}\ do(OfficeType) + \mathrm{c}\ do(z_1 = z_1) + \mathrm{d}\ do(z_2 = z_2) + e$. The OfficeType intervention may be M2K Only, CFRD Only, One System, or Both Systems. By holding the two covariates constant the equality $z_i = \mathrm{z}_i$ denotes their static, associational nature (z_1 is receptivity to innovation and z_2 is rank of job). The fixed causal effect of Office Type = Both Systems is $b = \mathrm{E}(\mathrm{y} \mid do\ Both\ Systems = 1, z_1, z_2) - \mathrm{E}(\mathrm{y} \mid do\ Both\ Systems = 0, z_1, z_2)$.

[19] Strictly speaking, these six offices are not a random sample from the list of all of the insurer's claims offices. The evaluators selected these offices because they are located in different parts of the country and they are typical offices. Within an office the returned sample of employee surveys may be considered a random sample because any missing surveys probably are missing at random. Using Sudman's (1976, 27) credibility scale for small samples the author gave this sample 30 out of 35 points (0.86) as follows: geographic spread, 8; discussion of limitations, 4; no special populations, 5; sample size, 4; sample execution, 4; and use of resources, 5.

[20] This conclusion follows from the fact that when the offices are designated as random and Models 4 or 5 are estimated, the gamma matrix of random effects is not positive definite and Proc Mixed calculates only the fixed-effects. These fixed-effects estimates are identical to those that result when the offices are not designated as random and the model includes only fixed effects.

[21] Throughout this analysis this chapter reports results based on the restricted maximum likelihood (REML) method of estimation, which generally is preferred to the various other methods of estimation. The other methods produce about the same results. The maximum likelihood (ML) estimates for $\hat{\sigma}_e^2$ and $\hat{\sigma}_o^2$ are respectively, 0.087 and 0.452. The MIVQUE(0) estimates are .109 and 0.458. The method of moments estimates from Proc Varcomp are 0.111 and 0.457. The REML estimates indicate that the two variance components are not correlated, their covariance is -5.87×10^{-5}. By creating an output data set that contains the residuals and other information, and then inputing the residuals into Proc Univariate, the following diagnostics were obtained for Model 1. The residuals have estimates of mean zero and variance 0.45 ($\hat{\sigma}_o^2 = 0.456$). Their distribution has a skewness of –0.19 and kurtosis of 0.45, neither indicate severe departures from normality. The tests for normality, however, reject the null hypothesis of a normal distribution. However, the normal probability plot indicates that it is reasonable to assume that the distribution of the residuals is normal. The box plot confirms that the distribution is slightly left skewed. The interquartile range (IQR) is 1. There are no severe outliers beyond plus or minus 1.5 × IQR, the most extreme residuals are –1.57, –1.57, and –1.51. Within the four types of offices

the sample variances are similar; the ratio of the highest and lowest sample standard deviations is 1.26, well below the rule-of-thumb criterion for conducting an analysis of variance. Within each office type the distribution of residuals has no severe outliers, the skewness and kurtosis are not severe, the tests for normality reject the null, but the normal probability plots are reasonable.

[22] Even if $\hat{\sigma}_o^2$ lacked statistical significance, given that it is important substantively to explain this variation, it should be explained. Moreover, the counternull effect size of 0.21 indicates that this variation should be explained (Rosenthal et al. 2000, 13,14).

[23] The following SAS program will calculate the confidence interval based on the Satterthwaite approximation; for the details see (Littell et al. 1996, 147–149; 2006, 73,74),

```
data satt1;
C = 30.8; *coefficient of var(office) in e(ms office)from calculations of C;
msee = 85.3807713/187; *ms error based on ees from GLM;
msof = 19.3135292/5; *ms model based on offices from GLM;
sa2 = .11059; *estimate of var(office) from varcomp;
v = (sa2**2)/((((msof/c)**2)/5) + (((msee/c)**2)/187)); *Approx df;
c025 = cinv(.025,v); *lower 2.5 chi square percentage point;
c975 = cinv(.975,v); *upper 97.5 chi square percentage point;
low = v*sa2/c975; * lower limit;
high = v*sa2/c025; *upper limit;
run;
data satt2; set work.satt1;
proc print;
run;
```

The elements of C are $n.$ = 193, sum of n_i^2 = 7,527, and $(s - 1) = 5$.

[24] Using Pearl's operator the causal equation is $Y_{ij} = \gamma_{00} + \gamma_{10}\ do\ (z_{1ij} = z_{1ij}) + \gamma_{20}\ do(z_{2ij} = z_{2ij}) + u_{0j} + r_{ij}$. That is, hold the covariates constant at their initial values.

[25] By again creating an output data set that contains the residuals and other information, and then entering the residuals into Proc Univariate, the following diagnostics were obtained for Model 2. The residuals have estimates of mean zero and variance of 0.40 ($\sigma_c^2 = .411$). Their distribution has a skewness of –0.38 and kurtosis of –0.55, neither indicate severe departures from normality. The tests for normality, however, reject the null hypothesis of a normal distribution. However, the stem and leaf plot and the normal probability plot indicate that it is reasonable to assume that the distribution of the residuals is normal. The box plot confirms that the distribution is slightly left skewed. The interquartile range (IQR) is 0.903. There are no severe outliers beyond plus or minus 1.5 × IQR = ±1.35. The residuals beyond that bound are –1.49, –1.46, –1.40, –1.32, and +1.36. Within the four types of offices the sample variances are similar; the ratio of the highest and lowest sample standard deviations is 1.44, well below the rule-of-thumb value of 2 for conducting an analysis of variance. Within each office type the distribution of residuals has no severe outliers, the skewness and kurtosis are not severe, not all tests for normality reject the null, and the normal probability plots are reasonable.

[26] Some diagnostics for Model 3a follows: The residuals for the overall model have estimates of mean zero and sample variance 0.397 ($\hat{\sigma}_e^2 = 0.41$). The distribution of the residuals has a skewness of –0.38 and kurtosis of –0.53, neither indicate severe departures from normality. The tests for normality, however, reject the null hypothesis of a normal distribution. However, the stem and leaf plot and the normal probability plot indicate that it is reasonable to assume that the distribution of the residuals is normal. The box plot confirms that the distribution is slightly left skewed. The interquartile range (IQR) is 0.898. There are no severe outliers beyond plus or minus 1.5 × IQR = ± 1.35. The residuals beyond that bound are –1.51, –1.48, –1.40, –1.39, and + 1.37. Within the office with CFRD Only the estimate of the mean is zero and the sample variance is 0.62. Within the other offices the estimate of the mean is zero and the sample variance is 0.38; the ratio of their sample standard deviations is 1.28, well below the rule-of-thumb value of 2 for conducting an analysis of variance. Within each of these two types of office type the distributions of residuals have no very severe outliers, the skewness and kurtosis are not severe, not all tests for normality reject the null, and the normal probability plots are reasonable. The residuals of the other singular implementations

show similar patterns and the sample variances tend to be identical. For M2K only, when the indicator is one the sample variance of 0.407 about equals that when the indicator is zero, .396; both means are zero. For One System, when the indicator is one the sample variance of .430 about equals that when the indicator is zero, 0.379; both means are zero. For these types of offices, when the office type and the covariates (mean = 0) are evaluated at their mean values, the sum of the intercept and the product of the office type and its mean equal the grand mean, as it should.

[27] Some diagnostics for Model 3d follow. The residuals for the overall model have estimates of mean zero and sample variance of 0.406 ($\hat{\sigma}_e^2 = .41$). The distribution of the residuals has a skewness of –0.33 and kurtosis of –0.66, neither indicate severe departures from normality. The tests for normality reject the null hypothesis of a normal distribution. However, the stem and leaf plot and the normal probability plot indicate that it is reasonable to assume that the distribution of the residuals is normal. The box plot confirms that the distribution is slightly left skewed. The interquartile range (IQR) is 0.91. There are no severe outliers beyond plus or minus 1.5 × IQR = ± 1.37. The residuals beyond that bound are –1.463 and –1.459. When Both Systems = 1 the sample variance is 0.30; the mean is zero; the distribution of residuals is skewed to the left (–0.911); the normality tests reject the null; the two outliers are not severe; and the distribution is rather flat and somewhat bimodal (the kurtosis is –0.14). When Both Systems = 0 the sample variance is 0.47; the mean is zero; the normality tests reject the null but the normal probability plot is rather linear; there are no outliers; the distribution of residuals is somewhat bell shaped with a short second bulge on the left (skewness = –0.18); the kurtosis of –0.86 indicates that the tails are too short for a normal distribution. The ratio of the two sample standard deviations of 1.27 is less than the rule-of-thumb value of 2 for conducting an analysis of variance. The collinearity of the fixed variables is negligible.

[28] Because the notation is simpler, here are the causal equations for the office type model:

$$Y_{ij} = \gamma_{00} + \gamma_{01}\ do(CFRD = 1, Not\ CFRD = 0) + \gamma_{10}\ do\ \left(z_{1ij} = z_{1ij}\right) + \gamma_{20}\ do\left(z_{2ij} = z_{2ij}\right) \\ + u_{0j} + r_{ij},$$

$$Y_{\mathrm{ij}} = \gamma_{00} + \gamma_{01}\ do(M2K = 1, Not\ M2K = 0) + \gamma_{10}\ do\left(z_{1ij} = z_{1ij}\right) + \gamma_{20}\ do\left(z_{2ij} = z_{2ij}\right) \\ + u_{0j} + r_{ij},$$

$$Y_{ij} = \gamma_{00} + \gamma_{01} do(OneSystem = 1, OneSystem = 0) + \gamma_{10} do\left(z_{1ij} = z_{1ij}\right) + \gamma_{20} do\left(z_{2ij} = z_{2ij}\right) \\ + u_{0j} + r_{ij},$$

and

$$Y_{ij} = \gamma_{00} + \gamma_{01}\ do(Both\ Systems = 1, Both\ Systems = 0) + \gamma_{10}\ do\left(z_{1ij} = z_{1ij}\right) \\ + \gamma_{20}\ do\left(z_{2ij} = z_{2ij}\right) + u_{0j} + r_{ij}.$$

Hold the covariates constant at their initial values and, depending upon what intervention is made in an office, change the indicator from 0 for not implemented to 1 for implemented.

[29] The causal equation is: $Y_{ij} = \gamma_{00} + \gamma_{01}$ *do*(*Both Systems* = 1, *Both Systems* = 0) + γ_{02} *do*(*One System* = 1, *One System* = 0) + γ_{10} *do*($Z1_{ij} = Z1_{ij}$) + γ_{20} *do*($Z2_{ij} = Z2_{ij}$) + r_{ij}. Hold the covariates constant at their initial values and in offices with Both Systems being installed change the indicator from 0 to 1 and in offices with One System being installed change the indicator from 0 to 1.

[30] The Bonferroni adjustment may not be necessary here because the ordering was predicted in advance. Some diagnostics for Model 4 follow. The residuals have estimates of mean zero and variance 0.398 ($\sigma_c^2 = 0.407$). Their distribution has a skewness of –.33 and kurtosis of –.55, neither indicate severe departures from normality. The tests for normality reject the null hypothesis of a normal distribution. However, the stem and leaf plot and the normal probability plot indicate that it is reasonable to assume that the distribution of the residuals is normal. The box plot confirms that

the distribution is slightly left skewed. The interquartile range (IQR) is 0.834. There are some outliers beyond plus or minus 1.5 × IQR = ±1.251. The residuals beyond that bound are –1.47, –1.37, –1.344, –1.341, –1.32 and + 1.43. Within the three types of offices the sample variances are very similar; the ratio of the highest and lowest sample standard deviations is 1.11, much less than the rule-of-thumb value of 2 for conducting an analysis of variance. Within each office type the distribution of residuals has no severe outliers, the skewness and kurtosis are not severe, not all tests for normality reject the null (for the baseline offices the null hypothesis of a normal distribution is not rejected), and the normal probability plots vary. All have a linear component; the first two also have a serpentine pattern that wraps around the linear plot.

[31] The causal equation is: $Y_{ij} = \gamma_{00} + \gamma_{01}\, do(Both\ Systems = 1, Both\ Systems = 0) + \gamma_{02}\, do(CFRD = 1, CFRD = 0) + \gamma_{03}\, do(M2K = 1, M2K = 0) + \gamma_{10}\, do\ (Z_{1ij} = Z_{1ij}) + \gamma_{20}\, do(Z_{2ij} = Z_{2ij}) + r_{ij}$. Hold the covariates constant at their initial values and in offices with Both Systems being installed change the indicator from 0 to 1, in offices with CFRD Only being installed change the indicator from 0 to 1, and in offices with M2K Only being installed change the indicator from 0 to 1.

[32] Some diagnostics for Model 5 are as follows. The residuals have estimates of mean zero and variance 0.397 ($\hat{\sigma}_e^2 = 0.408$). Their distribution has a skewness of –0.35 and kurtosis of –0.48, neither indicate severe departures from normality. The tests for normality reject the null hypothesis of a normal distribution. However, the stem and leaf plot and the normal probability plot again indicate that it is reasonable to assume that the distribution of the residuals is normal. The box plot confirms that the distribution is slightly left skewed. The interquartile range (IQR) is 0.835. There are some outliers beyond plus or minus 1.5 × IQR = ±1.25. The residuals beyond that bound are –1.51, –1.47, –1.46, –1.37, –1.35 and + 1.42; on the box plot none are identified as outliers. Within the four types of offices the sample variances are very similar; the ratio of the highest and lowest sample standard deviations is 1.43, less than the rule-of-thumb value of 2 for conducting an analysis of variance. Although the office with CFRD Only passes three of the four tests for normality, the box plot identifies two outliers. Although the office with M2K Only has six observations beyond 1.5 × IQR = 1.002, on the box plot none are identified as outliers.

[33] The parameter study indicates that both of these effects would attain statistical significance if there were 88 employees sampled in each office, a number that is larger than the number of fraud workers in a typical office. Consequently, Model 5 is not the best model.

Part III
Evaluative Research

Chapter 10
Cause and Consequences

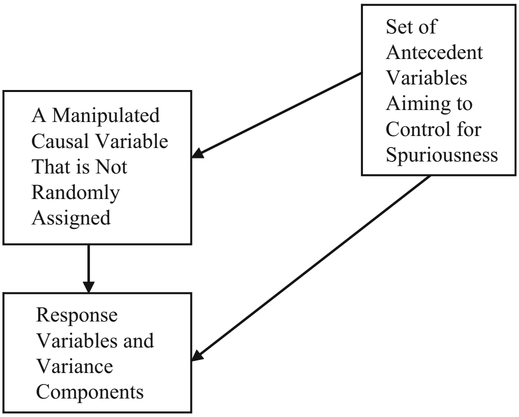
> *Unlike a true experiment, in which treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. In order to control for systematic differences between the control and treatment groups, we need two years of data, one before the policy change and one after the change. Thus, our sample is usually broken down into four groups: the control group before the change, the control group after the change, the treatment group before the change, and the treatment group after the change.*
>
> —Jeffrey M. Wooldridge (2006, 458)

Probing the *causes of effects*, the contextual analyses of the previous four chapters exemplified Coleman's (1964, 116, 189–240) emphasis on quantifying the implied causal effects of several predetermining variables on a response. Probing the *effects of a cause* (Holland 1986, 945; Morgan and Winship 2007, 280–282), these chapters on evaluative research exemplify Rubin's (1974) emphasis on studying the effects of a manipulated cause—here, some consequences of comprehensive school reform (CSR) on measures of achievement. Because these chapters study repeated measures on the same schools, they are similar to studies that analyze panel data. But, because of the students' mobility out of and into these schools, and other changes in the compositions of the schools, these chapters are best viewed as analyzing repeated quasi-panel data.

Evaluating how the activities of educational consultants can improve elementary schools, these chapters apply difference-in-differences (DID) study designs. Applying this design, the longitudinal change in the target group that receives the innovative treatment is compared with the longitudinal change in one or more comparison groups that do not receive this treatment. The difference between these differences quantifies the effect of the intervention. This average causal effect is equivalent to the coefficient on the interaction effect defined by the product of the indicator for the treatment group and the indicator for the post-intervention time period, assessed after potentially spurious effects have been appropriately controlled.

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_10, © Springer Science+Business Media B.V. 2011

This diagram depicts aspects of the logic of these studies:


Set of Antecedent Variables Aiming to Control for Spuriousness
A Manipulated Causal Variable That is Not Randomly Assigned
Response Variables and Variance Components

The notation of Chapter 3 formalizes this diagram as either $yx.T$ (i.e., $x \rightarrow y$, holding constant the set of antecedent test factors T), or as $yx.T^p$ (i.e., $x \rightarrow y$, holding constant the set of antecedent test factors T and the propensity scores, superscript p). Although the covariates may have independent effects on the response variable, they are centered by their means. This centering puts their effects into the modified intercept term and sharpens the focus on discerning the effects of the interventions.

A Pressing Social Problem: Student Achievement and Schools

The academic achievement of the United States of America (USA) works against the view that this country is the leading country of the world. Comparing the achievement of 15-year-old students in 30 affluent countries, which are members of the Organization for Economic Co-Operation and Development (OECD), students in the USA rank twenty-first in science literacy and twenty-fifth in mathematics literacy (Burd-Sharps et al. 2008, 196). Within the USA, children with disadvantaged demographic characteristics—especially lower socioeconomic status (i.e., children from poor families) and minority ethnicity (i.e., children from African-American, Native American, or Hispanic families)—often are found to perform less well on standardized tests than children from more affluent, white families (Burd-Sharps et al. 2008, 178–181). This gap in achievement, which is especially wide for students from families that are both poor and minority, is indicative of numerous social problems that combine to create and exacerbate this disparity, which is itself a major social problem. This achievement gap works against the equalitarian ethos of the USA that emphasizes equality of opportunity and, for some people, equality of results; and it weakens the capacity of the USA to innovate in science and technology and to develop economically.

Four Types of Policy Orientations

Numerous educational policies aim to improve academic achievement and reduce this achievement gap; a fourfold typology clarifies some of their root assumptions. The four types are defined by different emphases on beliefs about the malleability of schools and the malleability of their students; these are: *fatalism* (unchangeable schools and unchangeable students); *pragmatic activism* (unchangeable schools and more malleable students); *pluralism* (malleable schools and less changeable students); and *comprehensive school reform* (malleable schools and malleable students).[1] These definitions are "ideal types" (Weber 1947, 115–118); future empirical inquiries can substantiate these patterns and test their linkages to specific policy options.

Fatalism: unchangeable schools and unchangeable students

The fatalistic pattern of beliefs holds that schools and children cannot be changed much. Why? Schools are difficult to change because the principals and faculties of most schools are burned out, mediocre at best, and securely employed. More importantly, the different social categories of the students are thought to have on average different levels of innate intelligence as measured by IQ tests; most probably, these different levels are thought to be due to genetic differences: Given that IQ is an intrinsic and unitary characteristic of a person, a few people with diverse social backgrounds are gifted intellectually; many others are broadly average (e.g., the middle class of whites and ethnic minorities); and still others are below average (e.g., disproportionately the poor and ethnic minorities).

These fatalistic beliefs lead their adherents to question policies that aim to raise the IQ of those with lower scores by improving their circumstances and the schools that service their children. They would rather channel underachievers into jobs in the vocational branches of the service economy where the work is not too intellectually complex (e.g., carpentry, plumbing, electrical services, driving trucks or taxi cabs, providing security in stores and airports, and so forth), or into repetitive blue collar jobs that are not challenging cognitively. Such adherents perceive little need to provide better schools and educations to those marked by a high probability of low achievement because such efforts are bound to disappoint; it is futile to try to improve schools with the aim that every student can perform above average.[2]

Pragmatic activism: unchangeable schools and more malleable students

Neighborhood schools are difficult to change because they reflect the social and economic characteristics of the residents of the neighborhood, and because administrators and most teachers are tenured; even if their students do not perform well, the teachers' union protects their jobs. Thus, low-performing urban neighborhood schools are in a triple bind: the socio-cultural aspects of their students (i.e., low income families, single parents, too much watching of television, and insufficient reading material in the home) work against academic achievement; the administrators and teachers resist change; and, contrary to federal policy, such schools tend to be segregated on the basis of skin color and ethnicity: predominantly black, Hispanic, or white.

Since in the short run the schools cannot be changed much, and the students' performance would improve if they attended better schools, then an obvious policy recommendation would be to enroll the students in better schools. In their new schools, the minority students would interact with people of different social backgrounds, which may compensate for their disadvantaged home background; the schools would exhibit a more equitable ethnic balance; the students' original school could receive better-performing students from other neighborhoods, becoming more balanced; and these better students would inspire the tenured teachers to strive for excellence, thereby improving the performance of the school.

From such assumptions stem mandated desegregation of schools, bussing to achieve ethnic balance, vouchers that allow students to attend the school of their parent's choice: secular or religious schools, charter schools, and so forth. Evaluative research studies aim to determine the extent to which such changes have the desired consequences on student achievement. All of the involved parties—school administrators, teachers, parents, politicians, unions, and activists—debate the implications of the limited improvements in student achievement these studies report, and they press for policies that they value. Educational research thus engenders debates about policy, and these debates shape the directions of reform (Coleman 1972).[3]

Pluralism: malleable schools and less changeable students

Even though cognitive and social constraints may limit some of the students' performance on standardized tests of achievement, the pluralistic policy orientation holds that children have different blends of talents, skills, and needs, and that their school should be able to educate them taking into account their unique intelligences. For example, Gardner (1983) conceptualizes seven core intelligences: linguistic, logical-mathematical, musical, bodily-kinesthetic, spatial, interpersonal, and intrapersonal; subsequently, he added naturalist intelligence. Different configurations of these intelligences characterize individual students. Consequently, teachers and schools should further develop each student's capabilities along these dimensions, enhancing strengths and reducing weaknesses.

To accommodate these diverse intelligences of the students, schools can be improved by increased monetary investments in new or renovated school buildings; state-of-the art libraries and laboratories; the use of computers in instruction; professional development for teachers and administrators; new leadership and new teachers; strengthened school security; nutritional food at breakfast and lunch; enhanced curriculums for reading skills, mathematics, and science; enriched artistic, musical, and physical education; special courses for students with special needs; and so forth.

However, because of the students' diverse intelligences, not all of them can benefit equally from some of these innovations. Rather than encouraging low achievers to enroll in vocational schools and the very bright to enroll in magnet schools, let all of the students attend the pluralistic, enhanced school, which has different groupings of children who exhibit different abilities, needs, and interests, which the school tries to address. Thus, the school will be balanced ethnically and intellectually, but some classrooms will be composed of the brighter students who take the more rigorous curriculums; there even may be a school-within-the-school for these students. But the pluralistic school encourages all students to develop their own unique configuration of capabilities.

Comprehensive school reform (CSR): malleable schools and malleable students

Comprehensive school reformers assume that all students can learn and that external change agents can improve the school's leadership and teachers, who then will improve the achievement of their students. Educational consultants (i.e., the change agents), some from profit-seeking companies (i.e., the change agencies), accept the given infrastructure, leadership, teachers, and parents of a school and, within these parameters, they attempt to improve the school's effectiveness. Based on its philosophy of educational change—and on prior theory, scientific research, and experience—the change agency will have developed a coherent model for comprehensive school reform that shapes the activities of its consultants, who adapt the model to local circumstances. During the time period (circa 1997–2002) of this book's CSR studies, the low-performing schools that chose to undergo comprehensive reform applied for funds from the Comprehensive School Reform Program (CSRP) of the federal Department of Education, and also funds from other sources. Receiving the funding, the school district signed a multiyear contract with the change agency for the use of its consultants and resources.

At least 29 different reform models of varying effectiveness and cost have been designed, implemented, and evaluated (Borman et al. 2003). Depending on the reform model, the consultants may try to improve the leadership capabilities of the principal; skills of the teachers through professional development and training; curriculum and course material; use of technology; level of parental involvement; and pedagogy, perhaps by shifting to project-based learning or to direct instruction; and so forth.[4]

The implementation of a design for comprehensive school reform is not inexpensive. For example, the costs per school during the 2005–2006 school year for implementing the Success for All (SFA) curriculum, a highly regarded program of reform, were as follows: first year = $88,580 (training = $44,750, materials = $40,080, conferences = $3,750); second year = $58,200 (training = $28,275, materials = $28,275, conferences = $1,650); third year = $34,566 (training = $24,900, materials = $8,016, conferences = $1,650). Thus, the full three-year program costs about $181,346 per school. From 1987 to 2006, SFA was implemented in 1,400 schools (Comprehensive School Reform Quality Center, November 2006, 223–238).

Because of the No Child Left Behind (NCLB) act of 2001, most school reform efforts are now rather narrowly focused on improving the students' performance on standardized tests of verbal and quantitative achievement, so that the school will meet the following NCLB criteria. If an underperforming school fails to meet adequate yearly progress gauged by its percentage of students failing to meet grade-level standards for five consecutive years, then it must engage in planning for restructuring in the sixth year. Under NCLB, the restructuring options include: closing the school and then reopening it as a public charter school; firing the school staff including teachers and the principal and replacing them with new people; contracting with an outside organization that will operate the school; letting the state educational agency takeover the operation of the school; or engaging in some other form of comprehensive restructuring. Apparently, in 2008 all federal funds for comprehensive school reform were channeled to a clearinghouse that provides literature to school districts; for 2009, no funds were available for other change agencies and their consultants. However, the recent Obama stimulus plan provides new funds for educational reform, which may revivify comprehensive school reforms similar to those studied here.

The next two chapters on the comprehensive reform of elementary schools evaluate the effectiveness of consultants from Co-nect, a profit-seeking firm that recently became a component of Pearson Achievement Solutions. The target schools in these evaluations have many minority students: African-American, in Harford County, Maryland; and Mexican-American, in Houston, Texas. Initially, the academic achievement of these students was problematical.

Co-nect's Theoretical Model of Comprehensive School Reform

Co-nect designed its theoretical model of reform under the auspices of the New American Schools (NAS) Development Corporation, a nonprofit organization funded by the private-sector. NAS used these funds to create and support educational design teams that would facilitate the transformation of stagnant elementary and secondary schools into effective organizations through comprehensive reforms. In competition with 600 other design teams, in 1992 NAS selected Co-nect's reform model along with ten others for further development, pilot-testing, and scaling-up to other schools nationwide. Many studies have evaluated the effectiveness of these and other models of comprehensive reform; their results have been meta-analyzed: see Berends et al. (2002); Borman et al. (2003); and Aladjem et al. (2006). These sources have informed this discussion of CSR and Co-nect. Compared with Success for All, and with some of the other reform models (e.g., Direct Instruction, School Development Program), the Co-nect model engendered smaller but promising improvements that called for further empirical documentation. Co-nect hoped that the research underlying these chapters would address this need.

The Co-nect team premised that all students can learn; teachers can enhance their skills by utilizing Co-nect's proprietary teaching resources through the internet; and a nationwide network of Co-nect's certified consultants can directly guide administrators and teachers as they work together to effectively deliver, manage, and assess all aspects of reform. Forming a system of dynamic cumulative causation (Myrdal [1944] 1964, 1065–1070), the basic Co-nect model had these key components: multidisciplinary project-based learning; clustering of students and teachers; decentralized, cluster-based governance of the teachers; comprehensive and multifaceted assessment including exhibitions, performances, and standardized tests; access to the best available computer technology; and the development of strong professional and parental communities.

Guided by the unique needs of the school, the consultants developed a specific portfolio of services that might include extensive courses on professional development for principals and teachers; visits to exemplary schools; facilitation of the shift to project-based learning and to the clustering of teachers and students; training of teachers for their use of the Exchange (i.e., Co-nect's database of resources teachers can access through the internet); identification and dissemination of best practices; alignment of the curriculum with local and national educational standards; development of home and community support for learning; and so forth (Goldberg and Richards 1996).

The total operating costs for the first of several years of services was about $75,000, for one independent school with about 500 students or 40 teachers. For a school district with a cluster of three to five Co-nect schools, the costs for one year ranged from $36,000 to $40,000 per school. The fees varied according to the number and sizes of the schools, and the services provided (Comprehensive School Reform Quality Center 2006, 194–203).

Because the subsequent evaluations of the Co-nect reforms are based on the repeated achievement measures of clusters of students in clusters of schools, multilevel modeling is needed to appropriately estimate the parameters of the difference-in-differences (DID) designs of these studies. This design is especially useful in policy research studies because it can be applied to panel data, repeated cross-sections, and trends; it enables the statistical significance of average treatment effects to be readily calculated using either interactions of indicator variables or differences-in-differences; and it is consistent with the potential outcomes causal perspective.

The Basic DID Design

The activities of change agents during the post-implementation time periods are the quasi-experimental treatments of these studies. Because measures of the specific components of their activities are not available, these evaluations can only assess the overall effects of the interventions. In order to increase the number of data base cases in the statistical modeling, each study pools its data, thereby borrowing strength. If the data were not pooled, and separate regression models estimated for the treatment and comparison units, then the number of covariates most probably would exceed the number of units receiving the target treatment, thus making the estimation difficult or inappropriate.

Binary (0,1) Indicators

Coding the design variables of the regression model as binary (0,1) indicators (i.e., dummy variables), Table 10.1 depicts the indicator variables that specify the basic difference-in-differences model: The time period of an observation on a unit is coded 1 for the period after the implementation of the target intervention (i.e., post), and 0 for the baseline time period prior to the intervention (i.e., pre). The units assigned to the target treatment are coded 1 and those assigned to the comparison treatment are coded 0. Given this setup, the program effect can be estimated directly by the regression coefficient on the product of the target group indicator (1) and the indicator for the post time period (1); that is, it is the estimated coefficient on target × post in a statistical model that also includes the intercept, the effect of the target group, and the effect of time. Except for the indicators of the post time period, the target versus comparison group, and the target × post interaction; all other covariates including any propensity scores are centered by their respective sample means; this puts their effects into the modified intercept term. This modified intercept term equals the sum of the original intercept term (i.e., the intercept for the model when no variables are mean-centered) and the coefficients on the covariates each multiplied by their own sample mean. This modified intercept term appears in all four cells of the fourfold design table. Estimates of the least-squares means for a cell of the design table result when the estimated coefficients are multiplied by the binary codes for the target group (1 or 0), time period (1 or 0), and target × post (1 or 0), and the resulting products plus the intercept are summed for each of the cells; more simply, a coefficient is added to a cell when the relevant components of the cell are coded (1).

The regression coefficient that estimates the program effect equals the difference between two other differences: one difference is for the units receiving the target treatment and the other difference is for the units receiving the comparison treatment. Reading down the two columns of Table 10.1, the pre-to-post difference between the means in the comparison group (i.e., the target group coded 0), from the pre period (i.e., the post time period coded 0) to the post period (i.e., post time period coded 1), is $d(0) =$ the coefficient on the indicator for the post time period. The difference between the means in the target group coded 1, from the post time period coded 0 to the post time period coded 1, is $d(1) =$ the coefficient for the post time period plus the coefficient on the interaction effect, target × post. The difference between these two differences quantifies the program effect; see Eq. (1) on the next page.[5]

Table 10.1 A four fold, difference-in-differences design; two groups at two points in time, pooled data

Time period	Comparison group $G_i = 0$	Target group $G_i = 1$
Earliest, Post period = 0	Intercept (1) Target group (0) Post period (0) [Target group (0) × post period (0)] = (0)	Intercept (1) Target group (1) Post period (0) [Target group (1) × post period (0)] = (0)
Latest, Post period = 1	Intercept (1) Target group (0) Post period (1) [Target group (0) × post period (1)] = (0)	Intercept (1) Target group (1) Post period (1) [Target group (1) × post period (1)] = (1)

Note: Rather than conducting modeling separately for each group and thereby reducing the number of cases in each model, the data are pooled so that the modeling is based on all of the cases. All covariates except the time period and cross-sectional binary variables and their interaction are centered by their sample means. The effects of these mean-centered covariates appear in the modified intercept term; thereby intensifying the focus on the effects of a cause. The estimate of the program effect is the coefficient on the product of the target group indicator (1) and the post period indicator (1). To produce the predicted means, multiply the estimates of the coefficients by either (1) or (0) depending upon the coding for the cell of the four-fold table, and then sum the resulting products. Equivalently, simply add into a cell the coefficients for the variables coded (1). The separate columns for the units in the target and comparison groups imply that there is no leakage of the treatments between these groups; the columns also separate the realized and counterfactual outcomes of a unit.

$$\begin{aligned}\hat{\delta} = d(1) - d(0) &= \text{PostPeriod}(1) + [\text{Target} \times \text{PostPeriod}](1) - \text{PostPeriod}(1) \\ &= [\text{Target} \times \text{PostPeriod}](1) = \text{Target} \times \text{Post}.\end{aligned} \tag{1}$$

This difference-in-differences (DID) model is consistent with the potential outcomes perspective: Prior to the assignment of a unit either to the reform treatment or to the comparison treatment, the unit has potential outcomes under either treatment. After assignment, the units assigned to the target treatment will have realized outcomes under this treatment and counterfactual (i.e., potential) outcomes under the alternative treatment. Conversely, the units assigned to the comparison treatment will have realized outcomes under this treatment and counterfactual outcomes under the target treatment. Because only one realized outcome per unit can be observed, the unit-level causal effect cannot be observed; the fundamental problem of causal inference holds. Consequently, the program effect will be an average causal effect estimated by the difference between the means of an outcome measure in the pre to post period in the target group minus the difference between the means on that outcome measure in the pre to post period in the comparison group (the SAS code in this endnote presents some illustrative SAS code).[6]

An Econometric Formalization

Following Imbens and Wooldridge (2009, 67–68), this section more formally presents this difference-in-differences (DID) model for regression models based on ordinary least squares; this approach underlies the above exposition. Here, with schools being the individual units, let unit i belong to either group 1 or to group 0, respectively, for the target or comparison groups, $G_i \in \{0,1\}$. Units are observed at time period 0 (i.e., pre intervention) and at time period 1 (i.e., post intervention); $T_i \in \{0,1\}$. Given that the units $i = 1, \ldots, N$ are assumed to be random samples from a population, a unit's group identification and the time period of observation are thought to be random variables. All covariates other than the design variables are centered by their own sample mean thus putting their effects into the intercept term α. Then, the equation for the outcome for unit i in the comparison group is $\mathrm{Y}_i\ (0) = \alpha + \beta T_i + \gamma G_i + \varepsilon_i$. The equation for the outcome for unit i in the target group is $\mathrm{Y}_i\ (1) = \mathrm{Y}_i\ (0) + \delta_{\mathrm{DID}} = \alpha + \beta T_i + \gamma G_i + \delta_{\mathrm{DID}} + \varepsilon_\mathrm{i}$. In these equations, let α be the modified intercept term (the covariates are mean-centered); β be the effect of time (i.e., post = 1, 0), γ be a time-invariant group-specific component (i.e., target = 1, 0), δ_{DID} represent the program effect, and ε_i represent unobserved characteristics of the unit; ε_i is independent ($\perp\!\!\!\perp$) of G_i and $\mathrm{T_i}$, $\varepsilon_i \perp\!\!\!\perp (\mathrm{G}_i, \mathrm{T}_i)$ and $\varepsilon_i \sim iid\ N(0, \sigma^2_\varepsilon)$. Then, the standard δ_{DID} estimand in the population is:

$$\begin{aligned}\delta_{\mathrm{DID}} = (\mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)]) &= (\mathrm{E}[Y_i|G_i = 1,\ \mathrm{T_i} = 1) - \mathrm{E}[Y_i|G_i = 1,\ T_i = 0]) \\ &- (\mathrm{E}[Y_i|G_i = 0,\ T_i = 1]) - (\mathrm{E}[Y_i|G_i = 0,\ T_i = 0]) = d(1) - d(0). \\ &\qquad\qquad (\mathrm{I\&W\ eq.\ 35, 68})\end{aligned} \tag{2}$$

The regression model is (Imbens and Wooldridge 2009, 68):

$$Y_i = \alpha + \beta_1 T_i + \gamma_1 G_i + \delta_{\mathrm{DID}} W_i + \varepsilon_i \tag{3}$$

where W_i is the interaction of the group and the time indicators, $I_i = T_i G_i$. The estimate of $\hat{\delta}_{\mathrm{DID}}$ equals the difference between the pre to post means in the cells of the target group minus the difference between the pre to post means in the cells of the comparison group; that is, $\hat{\delta}_{\mathrm{DID}} = (\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00})$.

This $\hat{\delta}_{\mathrm{DID}}$ can be conceptualized as an estimate of the population average treatment effect ($\hat{\delta}_{\mathrm{PATE}}$) if the sample of size N is viewed as a random sample from a large super-population, and interest is in the average effect in the super-population (Imbens and Wooldridge 2009, 15); random-effects models share this goal of inference from a sample value to a super-population parameter. Alternatively, inference for the sample at hand may be the analytic goal. Then, a relevant estimator would be the average treatment effect conditional on the covariates in the sample ($\hat{\delta}_{\mathrm{CATE}}$) (Imbens and Wooldridge 2009, 16); fixed-effects models have this goal of inference for the sample at hand.[7] The PATE or CATE are most relevant when the policy being evaluated would expose all units to the treatment or to none at all (Imbens and Wooldridge 2009, 16).[8]

The next chapter, for example, assumes that the schools being studied are a random sample from a population of schools. This assumption is implemented in the multilevel modeling by designating the schools as random in the SAS program for Proc Mixed. This produces a random-effect estimate for each school and two estimated variance components; one for the unexplained variance between the schools, the other for the unexplained variance within the schools. The tests of significance take these estimated variances into account, adjusting the confidence intervals of the effect parameters; this adjustment corrects for clustering and allows inference from the sample at hand to the population of schools. However, this population must be very carefully defined.

Other Assumptions

To buttress any causal interpretation of the estimators of the program effect δ (either $\hat{\delta}_{\text{CATE}}$ or $\hat{\delta}_{\text{PATE}}$), the following assumptions should be tested empirically: The data are appropriate. The relevant variables are adequately measured and missing data are minimal.

The stable-unit-treatment-value assumption (SUTVA) holds. The target innovation should be crisply implemented with no interference from the units in the comparison group; all units assigned to the target group should receive the identical treatment; all units assigned to the comparison group should not receive the target innovation; and these units should receive the comparison treatment equally, which may be null or another treatment crisply implemented with no interference from the target group.

Unconfoundedness holds. The assignment of a unit to either the target or comparison treatment should not be confounded with the potential outcomes. Ideally, there is random assignment of a unit i to treatment group ($W_i = 1$) and to the comparison group ($W_i = 0$). If that is not possible, then unconfoundedness requires that all biases be removed by adjusting for covariates: that is, there are no uncontrolled predetermining variables that jointly influence the assignment of units either to the target treatment or to the comparison treatment and that also influence the outcomes (i.e., spuriousness is controlled). If the goal is to assess the causal effects of the treatments, then variables that intervene between the treatments and the outcomes should not be controlled (the implied causal effects are not interpreted by an intervening variable); also, variables consequent to the outcomes should not be controlled. Imbens and Wooldridge (2009, 26) formalize this unconfoundedness assumption as:

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i. \tag{4}$$

That is, the assignment to treatment W_i (0,1) is conditionally independent $\perp\!\!\!\perp$ of the potential outcomes Y_i (0), Y_i (1), given the observed covariates X_i.

The overlap condition holds. Each of the covariates X_i has empirical observations on the units in the comparison and target groups. Imbens and Wooldridge (2009, 26) state the overlap assumption more formally: "the [empirical] support of the conditional distribution of X_i given $W_i = 0$ [i.e., the comparison group] overlaps completely with that of the conditional distribution of X_i given $W_i = 1$ [i.e., the target group]".

Strong ignorability of selection bias holds. If both the unconfoundedness and overlap conditions hold, then selection bias is strongly ignorable (Imbens and Wooldridge 2009, 26). By applying multilevel modeling along with, respectively, matching and propensity scores, the next two chapters evaluating comprehensive educational reform aim to reduce selection bias, produce valid estimates of the implied causal effects, and allow appropriate inferences from the sample of schools being studied to a larger population of schools.

Evaluations of Comprehensive School Reform

These evaluations of Co-nect's comprehensive school reforms systematically elaborate the basic DID design by applying, respectively, designs with three treatment groups at two points in time and two treatment groups at three points in time.

Target, Matched, and Not-Matched Schools

Chapter 11 applies a DID design that encompasses three types of elementary schools in Harford County, Maryland: target, matched, and not-matched. Prior to this study, each of the five target schools were matched with two different comparison schools on indicators of parental social class (percentage of students receiving free or reduced-price lunches), ethnicity (percentage of nonwhite students), and school size. The 1996–1997 school year is the baseline preperiod; the target schools received the Co-nect school reforms for three school years, 1997–1998 through 1999–2000. This study treats the last school year as the post intervention period, and absorbs the effects of the two intervening school years into the intercept term. The primary outcomes are aggregated test scores on standardized tests of student achievement. Table 10.2 presents the logic the design and the coding of the design variables.

Study design

This six-fold design, which elaborates the fourfold design explicated earlier, also is consistent with the potential outcomes perspective. Prior to assignment to the treatments, a school has three potential outcomes: one for the target treatment, one for the null treatment of the matched schools, and one for the null treatment of the not-matched schools. After assignment to one of these treatments, the school

Table 10.2 A six fold difference-in-differences design; three types of schools at two points in time, pooled data

Time period	Five target schools Type = 1	Ten matched schools Type = 3	Sixteen not-matched schools Type = 2
Earliest, Post period = 0 "Pre" = SY 1996–97	Intercept (1) Target schools (1) Post period (0) [Target (1) × post period (0)] = (0)	Intercept (1) Target group (0) Post period (0) [Target (0) × post period (0)] = (0)	Intercept (1) Not-matched (1) Post period (0) [Not-matched (1) × post period (0)] = (0)
SY 1997–1998 and SY 1998–1999	Effects are absorbed into the intercept	Effects are absorbed into the intercept	Effects are absorbed into the intercept
Latest, Post period = 1 "Post" = SY 1999–2000	Intercept (1) Target schools (1) Post period (1) [Target (1) × post period (1)] = (1)	Intercept (1) Target group (0) Post period (1) [Target group (0) × post period (1)] = (0)	Intercept (1) Not-matched (1) Post period (1) [Not matched (1) × post period (1)] = (1)

Note: SY is an acronym for school year. Rather than conducting modeling separately for each group and thereby reducing the number of cases, the data are pooled so that the modeling is based on all of the cases. The SAS program for Proc Mixed designates the three types of schools as a CLASSification variable. It uses the grouping of schools with the highest assigned number (here 3) as the base category and creates indicator variables coded 0 or 1. Consequently the matched schools (Type = 3) are the comparison schools. The effects of the target schools and the not-matched schools are compared relative to the comparison schools. The design variables contained in the cells of the table are not centered by their sample means; all the other variables are centered by their own sample mean. Consequently, all six cells of the design share a common intercept that expresses the effects of the mean-centered covariates and the unmodified intercept. Separating the realized and counterfactual outcomes, the separate columns for each type of school also imply that there is no leakage of the target reform innovations into the other groups of schools, and vice versa.

has a realized outcome under that treatment and two counterfactual outcomes; one for each of the treatments it did not receive. Because a unit can receive only one of these treatments, the fundamental problem of causal inference holds: the estimated program effects will be implied average causal effects—"implied" underscores that all causal assumptions may not be met.

This DID design compares the change in test scores of the five target schools that received the comprehensive school reform treatment implemented by the Co-nect consultants with the change in test scores of the ten matched comparison schools. It also compares the change in test scores of the 16 not-matched schools that did not receive this or any other reform treatment with the change in test scores of the 10 comparison schools—these 26 schools received null treatments. Similar to the basic design, the program effect coefficient, which quantifies the effect of the reforms in the target schools relative to the null treatment of the comparison schools, equals the population average difference between the means of an outcome measure in time periods pre to post intervention in the target group, $d(1)_t$, compared with the pre to post difference between the means on that measure, $d(0)_m$, in the matched comparison group:

$$\begin{aligned}\hat{\delta}_1 &= d(1)_{\mathrm{t}} - d(0)_{\mathrm{m}} = \mathrm{Post\,Period}\,(1) + [\mathrm{Target} \times \mathrm{PostPeriod}](1) - \mathrm{Post\,Period}(1)\\ &= [\mathrm{Target} \times \mathrm{PostPeriod}](1) = \mathrm{Target} \times \mathrm{Post}.\end{aligned} \tag{5}$$

The computer program quantifies this coefficient and reports its statistical significance, estimating Bonferroni corrections for multiple comparisons when appropriate.

The comparison of change in the not-matched schools with the change in the matched schools also is informative: if there is no difference in effect between the two null treatments and a difference in effect between the target and null treatments, then this would be persuasive evidence for an effect of the reforms. Again, using the matched schools as the reference category, the estimate of the population average null treatment effect for the not-matched schools is the difference between these two differences:

$$\begin{aligned}\hat{\delta}_2 = d(1)_{\mathrm{nm}} - d(0)_{\mathrm{m}} &= \mathrm{Post\ Period}(1) + [\mathrm{Not\text{-}Matched} \times \mathrm{Post\ Period}](1)\\ &\quad - \mathrm{Post\ Period}(1)\\ &= [\mathrm{Not\text{-}Matched} \times \mathrm{Post\ Period}](1) = \mathrm{Not\text{-}Matched} \times \mathrm{Post}.\end{aligned} \tag{6}$$

Measures

Chapter 11 assesses aggregated measures of student performance for a school at the end of the school year. It first compares the results between the treatment conditions on the school performance index (SPI) and on the overall composite index (CI), which averages a school's third-grade and fifth-grade scores. Next, it compares the results for third-grade students on the composite index and separately for its reading and mathematics components. Then, using these same measures, it focuses on the performance of the fifth-grade students. The composite index measures the students' progress in these six core capabilities: reading, writing, language usage, mathematics, science, and social studies; the SPI gauges a school's distance from satisfactory standards. Because probability plots indicate that these correlated measures are approximately normally distributed, Proc Mixed is used to estimate the multilevel models.

Multilevel models

The multilevel modeling specifies the test scores of a school at a time point as level-1 variables, the schools as the level-2 variables, and the types of schools—target, matched, or not-matched—as the classification typology. To provide estimates of

$\hat{\delta}_1$ and $\hat{\delta}_2$, Proc Mixed estimated this multilevel model in which the effects of the mean-centered covariates appear in the intercept:

$$y_{ijk} = \mu + \alpha_i + \tau_k + (\alpha\tau)_{ik} + \mathrm{d}_{j(i)} + \mathrm{e}_{ijk}. \tag{7}$$

Here, y_{ijk} is the test score at time k for school j in the TYPE condition i; μ is the intercept term; α_i is the TYPE cross-sectional difference, $i = 1 =$ target schools, $i = 2 =$ not-matched schools; and $i = 3 =$ matched schools; τ_k is the indicator for the post time period, $k = 1$ is the post time period and $k = 0$ is the pre time period (the other time periods are mean centered and their effects appear in the intercept); $(\alpha\tau)_{ik}$ is the interaction TYPE × post time period, which provides the estimates of $\hat{\delta}_1$ and $\hat{\delta}_2$; $\mathrm{d}_{j(i)}$ is the random effect associated with the j^{th} school in TYPE $= i,$ where $\mathrm{d}_{j(i)} \sim iid$ $N(0, \sigma^2_{j(i)})$; and e_{ijk} is the random effect associated with the j^{th} school in TYPE $= i$ at time period k, where $\mathrm{e}_{ijk} \sim iid\, N(0, \sigma^2_{ijk})$. The mean-centered covariates for a school include the indicator variables for the two intervening school years, the proportion of students eligible for free or reduced-price lunches, the proportion of African-American students, the proportion of female students, the ratio of students to teachers, the urban or rural location of the school, the lowest grade of the school, and whether or not the school is a targeted poverty school.

Results

On every measure of student achievement, the target group exhibited more favorable effect sizes than the comparison group, whereas there was little difference in performance between the two null treatment groups. Because the estimates of $\hat{\delta}_2$ are not statistically significant, indicating no difference between the two null treatments on the various outcomes, whereas the estimates of $\hat{\delta}_1$ for the target group relative to the matched comparison group are statistically significant, these differences between the two coefficients support the implied causal effect of the school reforms in these target schools, as do the following considerations: The available aggregated measures in the analytic data set exhibited no missing data (no doubt, at the micro-level there were missing data); the stable-unit-treatment-value assumption was addressed by weighting the target observations by a measure of the quality of the implementation of the reforms; there was no evidence that the reform treatments influenced the comparison units; the units in the three treatment groups exhibited dense overlap of the covariates; and matching and multilevel analysis aimed to reduce selection bias. The covariance structures of the models fitted the patterns of the response variables, corrected p-values accounted for the multiplicity of the response variables, and effect sizes facilitated comparisons of the results. The concluding chapter of this book examines the validities of this study and other evidence that the findings achieved the zone of potential outcomes causality.

Using Propensity Scores

Chapter 12 evaluates the effects of comprehensive school reforms in seven elementary schools in the Reagan High School Feeder System of Houston, Texas. The Co-nect consultants facilitated project-based learning, standards-based curriculum alignment, and technology integration. The previous chapter closely matched the target schools with comparison schools prior to the implementation of the study; contrarily, this chapter compares the outcomes of the ongoing reforms with those of a null treatment received by a convenient, roughly matched, comparison group composed of the six remaining elementary schools in this feeder system. Because of the proximity of treatment and comparison schools, some of the target reforms could have affected the comparison schools, thereby reducing the program effect; and selection bias due to the rough matching of the schools could threaten the validity of the results, perhaps increasing the program effect.

Propensity scores

With the aim of minimizing the effects of selection bias, this chapter calculates propensity scores and includes them in the multilevel models as a mean-centered, fixed covariate. Propensity scores are a unit's predicted probability of receiving the quasi-experimental treatment. To obtain these probabilities, the units in the treatment group are coded 1 and the units in the comparison group are coded 0. Then using logistic or probit regression models, this indicator is regressed on a rich set of confounders that are antecedent to and have effects on the treatment and outcomes.[9] Conditioning on the resulting propensity scores can produce independence of potential outcomes and treatment indicators, and exchangeable treatment groups that are perfectly balanced, thereby minimizing selection bias (Rosenbaum and Rubin [1983] 2006; Imbens and Wooldridge 2009, 32).

Because of the statistical problem of separation, the preferred logistic or probit regression models initially did not converge to meaningful values. However, as the chapter explains, Firth's (1993) procedure, when applied to the data previously linearized by Goldberger's method (Achen 1986, 40–41), produced valid propensity scores. For about 200 Houston elementary schools, the binary treatment indicator was regressed on 16 prior characteristics of target group membership using the Firth logistic regression option. The resulting predicted probabilities for the target and comparison schools ranged in value form about 0.22 through 0.38. Because none were extremely close to zero or close to unity, these propensity scores could be used as a mean-centered fixed covariate (Imbens and Wooldridge 2009, 32–33).[10] On average, the target schools had a slightly higher probability of target group membership (0.33) than the comparison schools (0.29).

Study design

The study spans three school years: SY 1999–2000 (the baseline year), SY 2000–2001, and SY 2001–2002. Exploratory research indicated that the target schools experienced rather steady high performance across these three years, but the performance of the comparison schools dipped and then recovered. Consequently, using the earliest time period as the baseline for change, the chapter analyzes the effects of the program through the middle year, from the middle to the final year, and then from the baseline year through the final year.

Table 10.3 presents the indicator variable codes for a DID design for two types of schools at three points in time; the data are pooled. This design is consistent with the potential outcomes perspective, as follows. Schools are the unit. Prior to the assignment of a unit to the treatment group or to the comparison group, which receives the alternative (i.e., null) treatment; each school has a potential value of student achievement under each of these treatments. After assignment to one or to

Table 10.3 A six fold, difference-in differences design; two types of schools at three points in time, pooled data

The time period, A CLASSification variable	Six comparison schools $G_{\mathrm{i}} = 0$	Seven target schools $G_{\mathrm{i}} = 1$
Earliest time period = 0 Third grade tests, SY 1999–2000	Intercept (1) Target (0) Time period (0) [Target × time periods] = (0)	Intercept (1) Target (1) Time period (0) [Target × time periods] = (0)
Middle time period = 1 Fourth grade tests, SY 2000–2001	Intercept (1) Target (0) Middle period (1) [Target × middle period] = (0) Final period (0) [Target × final period] = (0)	Intercept (1) Target (1) Middle period (1) [Target × middle period] = (1) Final period (0) [Target × final period] = (0)
Final time period = 2 Fifth grade tests, SY 2001–2002	Intercept (1) Target (0) Middle period (0) [Target × middle period] = (0) Final period (1) [Target × final period] = (0)	Intercept (1) Target (1) Middle period (0) [Target × middle period] = (0) Final period (1) [Target × final period] = (1)

Note: SY is an acronym for school year. Rather than conducting modeling separately for each group and thereby reducing the number of cases, the data are pooled so that the modeling is based on all of the cases. The SAS program designates the three time periods as the categories of a CLASSification variable. It uses the time period with the highest assigned number as the base category and creates indicator variables coded 0 or 1. Consequently, the baseline time period becomes SY 2001–2002. Relative to this changed baseline, the effects of the target schools are compared with the comparison schools at two points in time, SY 1999–2000 and at SY 2000–2001. The design variables are not centered by their sample means, all of the other covariates are centered by their own sample mean. Consequently, all six cells of the design share a common intercept that expresses the effects of the mean-centered covariates and the unmodified intercept. Separating the realized and counterfactual outcomes, the separate columns imply that there is no leakage of the target reforms into the comparison schools.

the other of these treatments, the unit has a realized value of student achievement after exposure to the assigned treatment and would have a counterfactual value of student achievement after exposure to the other treatment. Ideally, the causal effect for each unit would be the difference between its realized and its counterfactual values of student achievement. Since the latter cannot be observed, the fundamental problem of causal inference holds; the effect of the reform program will be an implied average causal effect.

Two program effect estimators are defined first: $\hat{\delta}_{\mathrm{m}}$ assesses the average change in the mean outcomes from the first time period through the middle time period (m) and $\hat{\delta}_{\mathrm{f}}$ assesses the average change on the mean outcomes from the first time period through the final time period (f). Referring to Table 10.3, each estimate of these average causal effects equals a difference between two differences indexed by treatment 1 or 0 and time 0, 1, or 2; the coefficients on the resulting interactions quantify the effects:

$$\begin{aligned}\hat{\delta}_m &= d_1 - d_0 = (\bar{Y}_{11} - \bar{Y}_{10}) - (\bar{Y}_{01} - \bar{Y}_{00})\\ &= (\text{Middle Period} + \text{Target} \times \text{Middle Period}) - (\text{Middle Period})\\ &= \text{Target} \times \text{Middle Period}\end{aligned} \tag{8}$$

and

$$\begin{aligned}\hat{\delta}_f &= d_2 - d_0 = (\bar{Y}_{12} - \bar{Y}_{10}) - (\bar{Y}_{02} - \bar{Y}_{00})\\ &= (\text{Final Period} + \text{Target} \times \text{Final Period}) - (\text{Final Period})\\ &= \text{Target} \times \text{Final Period}.\end{aligned} \tag{9}$$

Between the second and third time periods, the Houston Independent School District assigned extra teachers to underachieving schools to help prepare their students for the standardized tests to be administered toward the end of fifth grade. Consequently, this study also focuses on discerning the effects of these extra teachers as reflected in the reduced ratios of students-to-teachers. It assesses change from the middle period to the final period using the following DID estimator which equals the difference between the coefficients on the interactions of (8) and (9):

$$\begin{aligned}\hat{\delta}_{fm} &= d_2 - d_1 = (\bar{Y}_{12} - \bar{Y}_{11}) - (\bar{Y}_{02} - \bar{Y}_{01})\\ &= [(\text{Final Period} + \text{Target} \times \text{Final Period})\\ &\quad - (\text{Middle Period} + \text{Target} \times \text{Middle Period})]\\ &\quad - \ [\text{Final Period} - \text{Middle Period}]\\ &= \text{Target} \times \text{Final Period} - \text{Target} \times \text{Middle Period}\end{aligned} \tag{10}$$

This chapter quantifies these average treatment effects using Proc Glimmix to estimate logistic multilevel models that have covariance structures that closely fit the patterns of the response variables.

Measures

The measures were taken on cohorts of students and then aggregated by educational administrators to characterize the schools at different points in time: third grade in SY 1999–2000; fourth grade in SY 2000–2001; and fifth grade in SY 2001–2002. Although at the microlevel, students may leave a school or shift from one school to another, at the macrolevel the quantities on the covariates for the target and comparison schools are remarkably stable across the study period. The following response variables are modeled: the schools' proportion of its students passing tests of reading; the proportion passing tests of mathematics; the average of those proportions; the proportion passing fourth-grade writing; and less saliently, the school quality ratings. Because the achievement measures are based on the total number of students passing a test to the total number of students taking a test, the test scores are treated as binomial variables in the events-trials format, and the multilevel model estimates a logistic regression model using the pooled data. The mean-centered covariates include propensity scores, proportion Hispanic in the school, proportion eligible for free or reduced-price lunches, highest grade of the school (sixth or fifth), students per teacher, and Success for All curriculums.

Multilevel model

The logistic multilevel regression model is:

$$\log[\pi_{ijk}/(1-\pi_{ijk})] = \mu + \alpha_{\mathrm{i}} + \tau_k + (\alpha\tau)_{ik} + \mathrm{d}_{j(i)} + \mathrm{e}_{ijk} \tag{11}$$

in which: π_{ijk} is the proportion passing a test at grade-time period k for a cohort in school j in the treatment condition i; μ is the intercept; α_i is the treatment group, $i = 1 =$ Target; $i = 0 =$ Comparison; τ_k is the grade-time period CLASSification variable, $\mathrm{k} = 2$ is fifth grade in SY 2001–2002, $k = 1$ is fourth grade in SY 2000–2001, and $k = 0$ is third grade in SY 1999–2000; $(\alpha\tau)_{ik}$ is the interaction of the treatment and the grade-time periods, which provides estimates of $\hat{\delta}_{\mathrm{m}}$ and $\hat{\delta}_{\mathrm{f}}$; $\mathrm{d}_{j(i)}$ is the covariance parameter of the j^{th} school grouped in treatment $= i$, where $\mathrm{d}_{j(i)} \sim iid\,N$ $(0, \sigma^2_{j(i)})$; and e_{ijk} is the residual covariance parameter of the j^{th} school grouped in treatment $= i$ at grade-time period k, where $\mathrm{e}_{ijk} \sim iid\,N(0, \sigma^2_{ijk})$. The Glimmix instructions specify the event/trials format, an appropriate covariance structure, the binomial distribution, and the logit link. The covariance structures are selected on the basis of the results of a series of tests using the new Covtest statements of SAS 9.2. All of the covariates are centered by their overall mean values and this transformation puts their effects in the intercept term. For each cell of the design, Proc Glimmix calculates the least-squares means on the logit scale, the difference-in-differences, and the odds ratios. It then uses the inverse link option to produce the predicted proportions and their upper and lower bounds (this endnote presents some illustrative SAS code that calculates the difference-in-differences).[11]

Results

In the target schools, the reforms produced significant improvements from third to fourth grade in their proportions of students passing tests of reading and mathematics, and their average; these schools maintained the improvements through in fifth grade. These schools also exhibited significant improvement in fourth-grade writing from Time 0 to Time 2. Conversely, the performance of the students in the comparison schools generally declined from third to fourth grade and then improved from fourth grade to fifth grade, apparently due to extra teachers who prepared the students for the fifth grade tests. This improvement in the comparison schools resulted in both types of schools performing at about the same level on the reading and mathematics tests administered in fifth grade. However, these extra teachers did not prepare the students for the fourth grade writing tests; consequently the Co-nect schools exhibited significant improvement in writing test scores across the evaluation period. The chapter demonstrates that both the Co-nect reforms and the more favorable ratios of students to teachers had beneficial consequences for the achievement of these ethnic-minority students. The concluding chapter of this book ascertains the validities of this study and whether the main findings achieved the zone of potential outcomes causality.

Policy Implications

The results of these two chapters substantiate the view that external change agents can improve the governance, teaching, and student achievement of initially underachieving schools, especially those that serve minority populations. The positive results develop over the course of the consultations and may drop off after the change agents cease their activities. Such reversions are not only due to the general intractability of reforming schools; changes in the administration of the school district often change the criteria of success and the means to achieve these new goals. However, comprehensive school reforms can work if properly implemented and supported by administrators, principals, teachers, and parents, as meta-analyses across many studies have found.

Endnotes

[1] Etzioni (1968, 1983, 87–88) defines malleability as the extent to which a policy variable is movable or changeable. Policy researchers rank factors according to their malleability and focus their research on the more malleable factors and on the constraints that impede the malleability of the less changeable factors.

[2] The controversial writings of Herrnstein and Murray (1994) and Murray (2008) are consistent with some of these beliefs about stratification, education, and intelligence; and have engendered a very spirited debate, see the edited collections by Fraser (1995) and Jacoby and Glauberman (1995).

[3] Coleman’s many educational policy studies express his pragmatic activism; Smith (2006, *i–xi*) lists and briefly discusses his educational studies.

[4] Ravitch (2010), believing that both schools and students are changeable, advocates the creation of a widely shared core curriculum that would engender multiple benefits. Hirsch (2010, 18) summarizes some of her thoughts about core curriculums as follows:

> It would assure the cumulative organization of knowledge by all students, would help overcome the notorious achievement gaps between racial and ethnic groups. It would make the creation of an effective teaching force much more feasible, because it would become possible to educate American teachers in the well-defined, wide-ranging subjects they would be expected to teach—thus educating students and teachers simultaneously.
>
> It would also foster the creation of much better teaching materials, with more substance; and it would solve the neglected problem of students (mostly low-income ones) who move from one school to another, often in the middle of the school year. It would, in short, offer American education the advantages enjoyed by high-performing school systems in the rest of the world, which far outshine us in the quality and fairness of their results.

[5] Similarly, reading across the two columns, for time = 0, the difference d’(0) between the target group (1) and the comparison group (0) = the coefficient for the target group. For time = 1, the difference between the target group (1) and the comparison group (0) = d’(1) = the coefficient for the target group and the interaction effect, target × post. The difference between these two differences is again the program effect coefficient:

$$\begin{aligned}\delta{=}d'(1) - d'(0) &= \text{Target (1) + Target} \times \text{Post Period(1)} - \text{Target (1)}\\ &= \text{Target} \times \text{Post Period(1) = Target} \times \text{Post}\end{aligned}$$

[6] Jill Tao of the SAS Institute has graciously provided the following SAS program that will create an illustrative SAS data set and will then call Proc Mixed to calculate the treatment effect as an interaction and as a difference-in-differences. Chapter 12 explains in some detail the logic of similar SAS programs for calculating program effects as differences-in-differences, also see endnote 11 of this chapter for an illustrative application of Proc Glimmix and logistic regression from a DID perspective.

```
*Example of SAS code for a Difference-in-Difference Design for a Continuous
Response Variable;
*This code creates the SAS data set;
data test;
do i = 1 to 20;
trt = round(ranuni(2345), 1);
period = round(ranuni(1234), 1);
y = trt + period + rannor(2687);
p = exp(-y)/(1 + exp(-y));
if ranuni(2356) > p then y2 = 1; else y2 = 0;
output;
end;
run;
proc print data = test;
run;
*This code calls Proc Mixed to estimate the treatment effect
on the continuous response variable;
proc mixed data = test;
class trt period;
model y = trt period trt*period/s;
*This statement will calculate the difference between
the means in the target group;
estimate ‘1,1 - 1,0’ period -1 1 trt*period 0 0 -1 1;
```

```
*This statement will calculate the difference between
the means in the comparison group;
estimate ‘0,1 - 0,0’ period -1 1 trt*period -1 1;
*This statement will calculate the difference-in-differences;
estimate ‘(1,1-1,0) versus (0,1-0,0)’ trt*period 1 -1 -1 1;
run;
```

[7] The acronym PATE refers to the population average treatment effect. The PATE estimand is the population expectation of the unit-level causal effect, $Y_i(1) - Y_i\,(0)$, which is $\delta_{\mathrm{PATE}} = \mathrm{E}[Y_i(1) - Y_i\,(0)]$. CATE refers to “conditional on the covariates in the sample” average treatment effect. The CATE estimand is $\delta_{\mathrm{CATE}} = 1/N \sum_{i=1}^{N} E[Y_i(1) - Y_i\,(0) \mid X_i]$ (Imbens and Wooldridge 2009, 15–17; Rubin 1974, 689–693).

[8] Willard Manning (circa 1990, personal communication) has pointed out that when the data are pooled, the intercept is based on all of the cases, even though the various cells in the design may actually have different numbers of cases. Thus, there is this counterfactual aspect to this procedure: At time 0, all of the cases are placed in the (0, 0) cell and experience the null treatment. Then, all of the cases are placed in the treatment group cell for that time period (0, 1) and they experience the baseline difference between the two groups. Then, at time 1 all of the cases are placed in the comparison group and are exposed to the effect of time (1, 0). Then, all of the cases are placed in the target group at that time (1, 1) and experience the effect of time, the target program, and their interaction; the latter is the program effect coefficient.

[9] At a meeting of the Boston Chapter of the American Statistical Association (May 4, 2010), Til Stürmer (an epidemiologist teaching at the University of North Carolina, Chapel Hill) reported the results of a computer simulation study that clarifies criteria for the selection of potential covariates in the calculation of propensity scores: Antecedent variables that influence both treatment and outcome should be included because they reduce spurious associations. Instrumental variables that influence only the treatment should be excluded because they lead to separation. Variables that influence only the outcome can be included because they improve the efficiency of the estimation.

[10] Imbens and Wooldridge (2009, 33) note several limitations to the use of propensity scores as covariates in regression models: the propensity score does not have substantive meaning; units with propensity scores of 0.45 and 0.50 are more similar than units with scores of 0.01 and 0.06; logit and probit models will produce similar scores in the middle range of the data but different scores for extreme observations; and propensity scores close to 0 or 1 are outliers that may have too much impact in weighting schemes.

[11] The SAS code below by Jill Tao of SAS implements the DID design for a binomial response variable. Chapter 12 explicates similar models in depth.

```
*Example of SAS code for a Difference-in-Difference Design for a
Dichotomous Response Variable;
*This code creates the SAS data set;
data test;
do i = 1 to 20;
trt = round(ranuni(2345), 1);
period = round(ranuni(1234), 1);
y = trt + period + rannor(2687);
p = exp(-y)/(1 + exp(-y));
if ranuni(2356) > p then y2 = 1; else y2 = 0;
output;
end;
run;
proc print data = test;
run;
```

```
*This code calls Proc Glimmix to estimate the treatment effect
on the dichotomous response variable. It calculates a logistic
regression model and converts the logit scale means to odds ratios and
proportions;
proc glimmix data = test;
class trt period;
model y2 (event = “1”) = trt period trt*period/s dist = binary link = logit;
*This statement will calculate the difference between
the means in the target group on the logit scale etc.;
estimate ‘1,1 - 1,0’ period -1 1 trt*period 0 0 -1 1/or ilink cl;
*This statement will calculate the difference between
the means in the comparison group on the logit scale, etc.;
estimate ‘0,1 - 0,0’ period -1 1 trt*period -1 1/or ilink cl;
*This statement will calculate the difference-in-differences on the
logit scale, etc.;
estimate ‘(1,1-1,0) versus (0,1-0,0)’ trt*period 1 -1 -1 1/or ilink cl;
run;
```

Chapter 11
Target, Matched, and Not-Matched Schools

> *Statisticians use the term "borrowing strength" to describe procedures . . . in which individual estimates are enhanced by incorporating information from others with whom he or she shares attributes. In this case, the model-based trajectories are shrunk toward the average trajectory of that person's peer group (those with the same predictor values). This combination yields a superior, more precise, estimate.*
>
> —Judith D. Singer and John B. Willett (2003, 136)

For five elementary schools in Harford County, Maryland, this chapter probes the effects of comprehensive school reforms on change in student test scores from school year (SY) 1996–1997 through SY 1999–2000. External consultants employed by Co-nect, an educational design and professional development firm then based in Cambridge, Massachusetts, were the change agents. The consultants began their core services in December of 1997 and ended them three years later. Controlling for the intervening years, this evaluation uses the average test scores of the schools for May of 1997 as the prereform measures and the average test scores for May of 2000 as the postreform measures, thereby covering the major portion of the consultants' basic contractual activities. These results pertain to this pre to post period and should not be generalized to time periods after the consultants ceased their activities.

New Contributions

Illustrating the strength of "borrowing strength," this study pools the data, applies a difference-in-differences (DID) design, and carefully chooses the model's covariance structure by using SAS's new Covtest statements. It then estimates the DID parameters innovatively via multilevel modeling that weights the computations by the quality of the implementations and adjusts the standard errors (SE), F-statistics, and degrees of freedom *(df)* by applying Kenward-Roger (KR) corrections. The three treatment groups include five target schools, ten matched schools, and 16 not-matched schools. Making numerous comparisons between each target school and

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_11, © Springer Science+Business Media B.V. 2011

its two matched schools, an earlier evaluation found inconsistent effects of the reforms. Contrarily, this chapter borrows strength by pooling the data on all 31 schools and estimates one equation that quantifies the average effects for each group of schools, rather than estimating the singular effects of each school based on comparisons with its two matched schools. The earlier analysis found a very weak "signal" and much "noise"—random variability. This chapter reports stronger signals and less noise—noticeable effects of the reforms even when the *p*-values are adjusted for multiplicity using the step-down Bonferroni algorithm.

The Setting of the Study

Harford County is northeast of Baltimore County and immediately west of Chesapeake Bay, which forms its eastern-most boundary. Except for a high tech corridor near Baltimore, during the period of this study the county was largely undeveloped residentially and industrially, comprising three state parks and many golf courses. Its 224,000 citizens are primarily white and middle class. They value its open spaces, strong school system, recreation facilities, safe neighborhoods, and opportunities for coastal living—pleasure boating, swimming, and water fowl hunting. Of its population (at the period of this study), about 97% are native-born, 9% are African-American, 2% Asian, 2% Hispanic, and 87% white. The ancestral backgrounds of the latter reflect the various immigrations from Europe: 26.5% are of German ancestry; 18% Irish; 11.6% English; 9.5% Italian; 6.9% Polish; and 6.5% American.

The Aberdeen Proving Ground is the county's largest employer, with about 7,000 civilian and 4,000 military personnel. The Upper Chesapeake Health System (1,762 employees), Rite Aid Mid-Atlantic Distribution Center (1,503), and Frito-Lay (475) are the next largest employers. The median household income in 1999 was about $57,200, the per capita income was $24,200, and the median sales price of a home was $150,000—about 4.9% were living below the poverty line. The available labor force includes 118,500 people; in March of 2001 3.4% were unemployed. The median commuting time is about 30 minutes.

Although there are no major state or private universities in the county, the Harford Community College, which offers two-year associate degrees, works with the Higher Education and Applied Technology Center (HEAT) coordinating onsite educational workshops and courses with the following degree-granting institutions: the College of Notre Dame of Maryland, the Whiting School of Engineering of Johns Hopkins University, Towson University, the University of Phoenix, the University of Maryland, and Villa Julie College. These institutions provide the faculty, establish their curriculum requirements, charge their own tuition, and confer their own degrees. The degree programs emphasize practical skills in engineering, computer science, education, nursing, paralegal studies, and business administration, thereby extending the more limited programs of the two public community colleges.

Table 11.1 Profiles of target schools in Harford County, Maryland

Name of elementary school	Darlington	Deerfield	Hickory	N. Harford	W.S.James
Locale	Rural, Inside MSA	Urban Fringe	Rural, Inside MSA	Rural, Inside MSA	Urban Fringe
Description of Locale	Hills and Farms of Darlington	Suburban Edgewood	North of Bel Air	Pylesville Scenic Area	Community of Abington
Physical Condition	Renovated, 1999	Partially Renovated	Renovated, 1997	Built 1984	Built 1976
Computers to Students	1 to 4.6	1 to 12.6	1 to 6.5	1 to 5.5	1 to 5.9
Capacity	197	576	686	511	564
Enrollment	165	594	614	499	600
Excess Capacity	32	–18	72	12	–36
Students to Teacher	15.6	19.3	21	21	20
% Subsidized Lunch	20.1	23.9	9.0	13.1	4.5
Targeted Poverty School	Yes, SAFE	Yes, SAFE	No	No	No
% African American	6.4	21.8	2.1	0.9	3.8
Matched Schools	Norrisville, Dublin	Joppatowne, Meadowvale	Fountain Creek, Jarretsville	Emmorton, Forest Lakes	Abington, Bel Air
Implementation Score	0.52	0.56	0.80	0.79	0.86

Russell and Robinson (April 2000) chose the matched schools and developed the implementation scores. SAFE is an acronym for this program: Maryland Schools Accountability Funding for Excellence.

During the period of this study, the Harford County Public Schools counted nine high schools, eight middle schools, and 31 elementary schools. The consultants provided services to the five elementary schools that Table 11.1 profiles; because they received the school reform treatment this chapter calls them target schools. Three of these five schools are located in rural areas, all are in appropriate physical condition, all have air conditioning, and all use computers as part of instruction. Darlington and Deerfield have relatively high percentages of students who receive either free or reduced-price lunches. Because of their high percentages of students with families in poverty, these schools received aid from this state program: School Accountability Funding for Excellence (SAFE). Their lower implementation scores indicate that the planned reforms were implemented less successfully in these schools than in the others. About a quarter of the students in Deerfield are African Americans, which is the highest percentage among these schools. Deerfield has the least favorable ratio of students to computers and along with W. S. James serves more students than its planned capacity. To provide a comparison group, each target school was matched with two other elementary schools in Harford County (referred to as matched schools) that received a null treatment; that is, no reform activities. The remaining elementary schools that were not matched to the target schools form a second comparison group (referred to as not-matched schools) that also received a null treatment.

What Did the Consultants Do?

Implementing the school reform design, which emphasizes the use of project-based learning, standards-based assessment, teaming of teachers with small groups of students, and the effective use of technology; the consultants provided the following professional development services for all faculty and instructional assistants: One half-day in-service workshops on the design of the reforms and project-based learning; two half-day sessions on developing criteria for performance-based assessments; three half-day sessions on the use of technology and the Internet for project-based learning and standards-based projects; and mini-sabbatical sessions for small groups of teachers responsible for sustaining the design elements in each school. These sessions covered portfolio assessment, technology action plans, and technology roundtable discussions. The school principals received a separate day-long session on leadership and also a training session of three days at sites in Cambridge, Massachusetts.

The teachers had access to the web-based Exchange that provides online professional development, interactive development of project-based curriculum, and nationwide networking opportunities. Teachers had the opportunity to submit school-based projects to an annual contest; some received cash awards. To continuously assess progress toward the highest level of implementation of the design, teachers had access to benchmarks and indicators. Teachers in the five elementary schools participated in reviews of progress in Harford schools and had opportunities to participate in progress reviews of other schools nationwide. Teachers from other schools—Critical Friends—visited the Harford schools as part of the national Critical Friends initiative.

To oversee the reform effort, the school district assigned a supervisor who assisted in the delivery of professional development and facilitated the managerial logistics of the effort. This supervisor worked closely with the Co-nect regional site coordinator, reinforcing the implementation of the reform design; different individuals staffed this position at different times. These consultants provided job-embedded continuous professional development and assisted in the coordination and facilitation of the site-based design teams.

Methods

This section discusses studies of Harford schools, DID designs, repeated measures, covariates, statistical models, selection of covariance structures, corrections for multiplicity, and calculations of effect sizes based on the KR-corrected estimates.

Previous Studies

This chapter builds upon and advances two earlier reports of the school reforms in Harford County: Harkins's "Informational Report to the Superintendent" (2001, July) and Russell and Robinson's "Co-nect Retrospective Outcomes Study" (2000, April). From the former, it takes the time period of analysis and the aggregated measures; from the latter, it takes the matched comparison schools and implementation scores.

On the basis of the demographics of the schools, and without regard to the test scores, Russell and Robinson (2000, April, 3) closely matched each target school with two other schools and eliminated from their study the remaining schools that they did not use as matched schools; those schools appear in the present study as the not-matched schools. The target and matched schools have similar scores on six demographic variables: grades served by the school; school size (total enrollment); ethnicity (the percentage of non-white students); parental socioeconomic status (the percentage of students eligible for free or reduced-price lunches); ethnicity (the percentage of non-white students); English proficiency (Limited English Proficiency, i.e., LEP students); and the mobility rate.[1]

Russell and Robinson assessed the effects of Co-nect reforms in each target school relative to each of its two matched schools. They compared the change in the percentages passing the six component tests of the Maryland School Performance Assessment Program (MSPAP) for a cohort of students beginning in third grade in 1997 and ending in fifth grade in 1999. For a target school and its two matched schools, a typical chart of theirs depicts the change from third to fifth grade in the percentage passing tests of reading, writing, language usage, mathematics, science, and social studies, the six components of the MSPAP. For the five target schools, these depictions create a total of 90 relevant within-chart comparisons (3 target and matched schools times 6 tests times 5 comparisons of target and matched schools).[2]

The researchers summarized the effects of the reforms as positive, negative, or equivocal: A target school whose growth in its percentages passing surpassed both of its matched schools received a positive rating (+); schools that underperformed both of their matched schools received a negative rating (–); and schools with equivocal results got an equivocal rating (=). Of 30 ratings of Co-nect schools (5 target schools time 6 tests), only 2 were positive, 22 were equivocal, and 6 were negative. Because their procedure does not borrow strength, the reported treatment effects are problematical.

Russell and Robinson combined two measures of the quality of the implementation of the reforms—the overall benchmark score and the site director's score. They took a simple average of these scores to create an implementation score that can range from 0.2 to 1.0; then they trichotomized the scores. Given their cut-points, two schools (Darlington and Deerfield) have Medium implementation and three schools (Hickory, North Harvard, and William James) have High implementation. They did not directly incorporate such scores into their comprehensive study as a weighting variable because most of their schools fell into the high end of the middle level.

This chapter uses these scores as weights by first assigning an implementation score of 1 to all 31 schools. Then, for the target schools, it adds their implementation score to 1 creating a new implementation score that differentiates one target school from the others, and from the nontarget schools. The scores for the target schools are: Darlington = 1.52; Deerfield = 1.56; Hickory = 1.8; North Harford = 1.79; and William S. James = 1.86; all other schools have a score of 1. The Proc Mixed runs use these implementation scores (referred to as Impscore) as a Weight variable that slightly changes the importance of a school's contribution to the calculation of the standard errors, depending upon the quality of its implementation of the reforms.

The DID Design

Unlike Russell and Robinson's procedure, which was appropriate for their data set, this study borrows strength by pooling the data of a more comprehensive data set and applying a difference-in-differences (DID) design that groups the schools according to this treatment typology: five target schools (type = 1) receive the reform treatment (i.e., the services sketched earlier); ten matched schools (type = 3) do not receive the reform treatment (these schools receive a null treatment), and the remaining 16 not-matched schools (type 2) do not receive the reform treatment (these schools receive a null treatment). During the period of this study, Harford County schools did not implement any other comprehensive school reforms.

The 31 schools are conceptualized as subjects that are nested within the three treatment types, this nesting is symbolized by: school (Trt). This three-group DID design is stronger than a two-group design for the following reasons: For the compound symmetry covariance structure that appropriately models most of the response variables, the additional grouping of schools facilitates the quantification of its two covariance parameters σ_s^2 and σ_b^2, which decompose the variance of the response variable, $\mathrm{Var}[Y_{ijk}] = \sigma^2$ into two components: The between-schools variance component is σ_s^2 and the residual variance component is σ_b^2. Their inclusion in the model corrects for clustering and may allow the results to be generalized to the universe of schools in Harford County and with extreme caution to schools in similar counties. The third group of schools also provides a benchmark for comparing the effect of the reforms. If both the target and the not-matched schools have favorable effects when compared to the matched schools, then this works against the attribution of a positive treatment effect due to the reforms.

Table 11.2 Illustrative derivations of the difference-in-differences estimators

$\mu_{\mathrm{TrtTime}\,=}\ \beta_0 + \beta_1\mathrm{Trt}_1 + \tau_1\mathrm{Time}_{01} + \delta_{11(}\mathrm{Trt}_1 \times \mathrm{Time}_{01}) + \beta_2\mathrm{Trt}_2 + \delta_{21}(\mathrm{Trt}_2 \times \mathrm{Time}_{01})$

Treatment (Trt):	Five Target Schools Type = 1	Ten Matched Schools Type = 3	Sixteen Not-Matched Schools Type = 2
DID estimators, Time 0 to Time 1			
Time 0 (Baseline), "Pre" = SY 1996–1997	$\mu_{10} = \beta_0 + \beta_1$	$\mu_{30} = \beta_0$	$\mu_{20} = \beta_0 + \beta_2$
Time 1 (Last Period), "Post" = SY 1990–2000	$\mu_{11} = \beta_0 + \beta_1 + \tau_1 + \delta_{11}$	$\mu_{31} = \beta_0 + \tau_1$	$\mu_{21} = \beta_0 + \beta_2 + \tau_1 + \delta_{21}$
Difference Time 1 – Time 0	$d(1)_\mathrm{t} = \mu_{11} - \mu_{10}$ $= \tau_1 + \delta_{11}$	$d(3)_\mathrm{m} = \mu_{31} - \mu_{30}$ $= \tau_1$	$d(2)_\mathrm{nm} = \mu_{21} - \mu_{20}$ $= \tau_1 + \delta_{21}$
Effect $\hat{\delta}_\mathrm{t}$ of Reform Treatment Relative to Matched	$\hat{\delta}_\mathrm{t} = d(1)_\mathrm{t} - d(3)_\mathrm{m}$ $= \delta_{11}$		
Effect $\hat{\delta}_\mathrm{nm}$ of Not-Matched "Treatment" Relative to Matched			$\hat{\delta}_{\ \mathrm{nm}} = d(2)_\mathrm{nm} - d(3)_\mathrm{m}$ $= \delta_{21}$
Effect $\hat{\delta}_\mathrm{tnm}$ of Reform Treatment Relative to Not-Matched	$\hat{\delta}_\mathrm{tnm} = d(1)_\mathrm{t} - d(2)_\mathrm{nm}$ $= \delta_{11} - \delta_{21}$		

This table assumes that the lowest coded value of the treatment typology will be the reference category for the (0,1) indicator variables (i.e., dummy variables) in the SAS runs. The effects for SY 1997–1998 and SY 1998–1999 are centered by their means and thus appear in the intercept.

Equation (1) below develops the DID design using the parameters of Table 11.2

$$\mu_{TrtTime} = \beta_0 + \beta_1 Trt_1 + \tau_1 Time_{01} + \delta_{11}(Trt_1 \times Time_{01}) + \beta_2 Trt_2 + \delta_{21}(Trt_2 \times Time_{01}) \qquad (1)$$

The treatment typology and the indicator for the time period specify the cell mean, which equals the following: β_0, an intercept; plus $\beta_{1,}$ the coefficient on Trt_1, the cross-sectional difference between the reform and matched group; plus $\tau_{1,}$ the coefficient on Time_{01} that gauges the effect of the post time period$_{;}$ plus $\delta_{11,}$ the interaction $\mathrm{Trt}_1 \times \mathrm{Time}_{01}$ of the reform treatment and the post time period; plus β_2 the coefficient on $\mathrm{Trt}_{2,}$ the cross-sectional difference between the not-matched and matched group; and plus $\delta_{21,}$ the coefficient on the interaction $\mathrm{Trt}_2 \times \mathrm{Time}_{01}$ of the "treatment" for the not-matched group and the post time period.

Table 11.2 derives the DID estimators as follows. For the five target schools, let the difference between the means for Time 0 and those for Time 1 be:

$$d(1)_{\mathrm{t}} = \mu_{11} - \mu_{10} = \tau_1 + \delta_{11}. \tag{2}$$

Let the difference between the means in the comparison group composed of the matched schools be:

$$d(3)_{\mathrm{m}} = \mu_{31} - \mu_{30} = \tau_1. \tag{3}$$

Then the DID estimator$\hat{\delta}_t$ for the average treatment effect is simply the difference between these differences, which equals the treatment × time interaction effect, δ_{11}:

$$\hat{\delta}_t = d(1)_t - d(3)_{\mathrm{m}} = \delta_{11}. \tag{4}$$

A similar logic defines the effect $\hat{\delta}_{\mathrm{nm}}$ of the notmatched comparison group relative to the matched comparison group, resulting in:

$$\hat{\delta}_{\mathrm{nm}} = d(2)_{\mathrm{nm}} - d(3)_{\mathrm{m}} = \delta_{21}, \tag{5}$$

which equals the interaction of the nonmatched and the posttime period indicators.

The effect $\hat{\delta}_{\mathrm{tnm}}$ of the reform treatment relative to the not-matched comparison group equals the difference between (4) and (5);

$$\begin{aligned} \hat{\delta}_{\mathrm{tnm}} &= [d(1)_{\mathrm{t}} - d(3)_{\mathrm{m}}] - [d(2)_{\mathrm{nm}} - d(3)_{\mathrm{m}}] \\ &= [d(1)_{\mathrm{t}} - d(2)_{\mathrm{nm}}] = \delta_{11} - \delta_{21}. \end{aligned} \tag{6}$$

These DID estimators of the average treatment effects are consistent with the potential outcomes causal perspective: Prior to assignment to the reform or matched groups, a school has potential outcomes under each treatment, reform or null. After assignment to one of these treatments, the school has realized outcomes under that treatment and counterfactual outcomes under the other treatment. The unit-level causal effect cannot be calculated because information about a school's response to the treatment that was not received is missing. Because the two groups of schools are closely matched, an average causal effect between the mean responses of these groups can be calculated. A similar logic holds for the average causal effect of the notmatched schools relative to the matched, and for the average causal effect of the reform treatment relative to the notmatched schools.

Repeated Measures

From Harkins's research, this study takes its equally spaced time periods (May of 1997 through May of 2000); its focus on aggregated school-level trends based on the students' performance; and two of its summary measures—the school performance index (SPI) and the composite index (CI).[3] Both variables are approximately normally distributed and are very strongly correlated (r = .999). Later on in this chapter, Figs. 11.1 and 11.3 will depict the trends in overall performance of the schools using these measures. Given their near identity, these figures represent a replication and not independent findings. By forming one composite estimate from these confounded effects, this chapter will address the bias due to their multiplicity.[4]

The school performance index (SPI) is conceptualized as the weighted average of a school's relative distance from the standard for satisfactory performance. For an elementary school, the SPI combines the MSPAP results of its students on the six content areas for grades three and five, and attendance. The matched schools have the highest scores on the SPI and the notmatched schools have the highest variability. For the full sample of schools across the years of this study the mean SPI score is 79.7 (45.4 to 107.2; SD = 15.3). The mean SPI score for the target schools is 77.9 (56.1 to 92.1; SD = 11.2), for the matched schools it is 87.7 (61.9 to 107.2; SD = 11.3), and for the not-matched schools it is 75.4 (45.4 to 105.2; SD =16.8).

The CI indicates the average performance of students across all six content areas of the MSPAP. The numerator for the CI is the number of students achieving satisfactory or better in each content area summed across all six content areas. The denominator is the number of students eligible for the MSPAP tests for each content area summed across all content areas. The resulting number is the average of the percentages of students achieving satisfactory or better performance across all content areas weighted by the number eligible. Number eligible is the total number of students taking each test plus the number of students absent and excused at each grade level.[5] Once again, the matched schools have the highest achievement scores and the not-matched schools have the highest variability. For the full sample of schools across the years of this study, the mean CI score is 54.5 with a range from 28.6 to 75.3 and a standard deviation (SD) of 11.6. The mean CI score for target schools is 53.2 (36.6 to 63.8; SD = 8.5), for the matched schools it is 60.5 (41 to 75.3; SD = 8.6), and for the not-matched schools it is 51.2 (28.5 to 73.5; SD = 12.7).

For third grade and for fifth grade, the CI and its reading and mathematics components are analyzed separately. Fig. 11.4 through 11.6 will portray the results for third grade, and Figs. 11.7 through 11.9 will portray the results for fifth grade. Depicting the consequences of the KR corrections, Figs. 11.10 and 11.11 will compare the effect sizes of the target and not-matched groups for all of the outcomes.

The Covariates

To minimize potentially spurious effects, this study controls for the following key covariates that could influence the schools' average test scores: the proportion of students eligible for free or reduced-price lunches, the proportion of African-American students, the proportion of female students, and the ratio of students to teachers. In each of the treatment groups, the levels of these covariates are constant across time and the target and matched groups have very similar levels; see Table 11.3. These constant proportions work against the notion that differential selection of students with these characteristics is responsible for the effects of the reform.

Table 11.3 The levels of the key covariates are stable across time; the target and matched groups have similar levels on these covariates

	Pre SY96–97	SY97–98	SY98–99	Post SY99–20	4 Year average
Proportion free or reduced price lunch					
Target	0.14	0.14	0.14	0.13	0.14
Matched	0.11	0.11	0.11	0.11	0.11
Not Matched	0.31	0.32	0.32	0.30	0.31
Proportion African-American					
Target	0.05	0.06	0.07	0.08	0.07
Matched	0.05	0.05	0.06	0.06	0.06
Not Matched	0.21	0.21	0.22	0.23	0.22
Proportion female					
Target	0.50	0.50	0.49	0.48	0.49
Matched	0.48	0.48	0.48	0.48	0.48
Not Matched	0.48	0.48	0.48	0.48	0.48
Student-to-teacher ratio					
Target	22.5	19.6	19.6	16.5	19.5
Matched	22.7	19.9	20.1	17.5	20.0
Not Matched	22.9	19.1	18.9	16.6	19.4

Additional control variables are the location of the school (urban fringe or rural), the lowest grade of the school (pre-kindergarten or kindergarten), whether or not the school qualifies for Title-1 benefits, and whether or not the school is a targeted poverty school under the auspices of SAFE. On the basis of their number of economically disadvantaged students, SAFE schools just miss qualifying for Title-1 aid. To control for the effects of time, the fixed variables also include mean-centered indicators for school year 1997–1998 and school year 1998–1999. The interactions of these indicators with the three types of schools are not statistically significant and for reasons of parsimony do not appear in the model. All of these covariates have been centered by their means so that their effects appear in the intercept term of the multilevel model. This procedure simplifies the statistics describing the results and facilitates their depictions.

Statistical Models

Explications of the treatment effects, the multilevel model, and SAS code compose this section.

Treatment effects

The preintervention time period (hereafter, pre) is SY 1996–1997; the postintervention time period (hereafter, post) is SY 1999–2000. To quantify the treatment effects of the DID design, the multilevel model includes the following indicators and interactions that are not centered by their means: intercept, target schools, not-matched schools, post, target × post, and not-matched × post. Because all of the indicators for the covariates and for the intervening time periods (SY 1997–1998 and SY 1998–1999) are centered by their means, their effects appear in the intercept. As derived earlier, the estimate of the population average treatment effect of the reforms is the coefficient on target × post, which is the interaction of the indicator for the target schools and the indicator for the post period. So that this effect is assessed relative to change in the matched schools, the model also includes the coefficient on not-matched × post, which is the interaction of the indicator for the not-matched schools and the indicator for post. Those two interactions assess the longitudinal trend in test scores for those two groups from May 1997 to May 2000, relative to change in the matched group.

Given the mean-centering of the covariates, the predicted mean test score of a response variable for a cell of the design equals the sum of the relevant coefficients on the design variables. For example, for target schools in the pre time period, the sum includes only the target parameter and the intercept that includes the effects of the mean-centered covariates. For these schools in the post time period, the relevant parameters are the sum of the coefficients on target × post, post, target, and the intercept. For matched schools in the pre period the sum includes only the intercept; in the post period the coefficient on the post parameter is added to the intercept.

As derived earlier, the estimate of the population average treatment effect $\hat{\delta}_t$ equals the pre to post difference between the means in the target group minus the pre to post difference between the means in the matched comparison group. For the not-matched schools, the parameters are derived analogously: the average null treatment effect $\hat{\delta}_{\mathrm{nm}}$ assesses the difference between the differences in the means in the not-matched group relative to the differences in the means in the matched group. If $\hat{\delta}_t$ is large and statistically significant and $\hat{\delta}_{\mathrm{nm}}$ is small and not significant, then this pattern would substantiate the implied causal effects of the Co-nect reforms.

The multilevel model

To provide estimates of these population average treatment effects and to test their statistical significance, this study applies SAS's Proc Mixed and Proc Glimmix for the analysis of repeated measures data (Littell, Milliken, Stroup, Wolfinger, and Schabenberger 2006, 159–202). In the multilevel model, Treatment (Trt) is a classification variable (a Class variable) that SAS uses to provide the indicators for target (Trt = 1) and not-matched (Trt = 2), both relative to matched (Trt = 3). SAS will use Trt = 3, the highest coded value, as the reference category for the indicator variables. The design factors Trt, Post, and Ttrt × Post are considered fixed as are all of the other covariates that are mean-centered. Schools nested in their treatment category [i.e., schools(Trt)] are the subjects [(Sub)]. Different covariance structures can appropriately model the various response variables, as explicated later.

The multilevel statistical model is:

$$y_{ijk} = \mu + \alpha_i + \tau_k + (\alpha\tau)_{ik} + \mathrm{d}_{j(i)} + \mathrm{e}_{ijk} \tag{7}$$

where:

y_{ijk} is the test score at time k for school j in the Treatment (i.e., Trt) condition i;
μ is the intercept term;
α_i is the Trt cross-sectional difference, $i = 1 =$ target schools, $i = 2 =$ not-matched schools; and $i = 3 =$ matched schools;
τ_k is the post time period indicator, $k = 1$ is the post time period (SY 1999–2000) and $k = 0$ is the pre time period (SY 1996–1997), the other time periods are mean centered, and their effects appear in the intercept;
$(\alpha\tau)_{ik}$ is the interaction Trt × Post time period;
$\mathrm{d}_{j(i)}$ is the between random effect associated with the j^{th} school in Trt $= i$, $\mathrm{d}_{j(i)} \sim iid N(0, \sigma^2_{j(i)})$; and
e_{ijk} is the residual random effect associated with the j^{th} school in Trt $= i$ at time period k, $\mathrm{e}_{ijk} \sim iid\ N(0, \sigma^2_{ijk})$.

All fixed covariates other than the design factors are centered by their overall means, thus putting their effects into the intercept term. However, these mean-centered variables appear in the structural portion of the Model statement of the SAS code.

SAS code

The following SAS code for a run of Proc Mixed using the Repeated statement can provide estimates of the parameters of the multilevel model:

```
Proc Mixed Data = standard ic ratio covtest cl;
  Class Trt period school Post;
  Model msapperform = Trt Post Trt*Post
    year98 year99 lowstgrd locate targett1 safe
    Stratio femprob Blackp forredlp/solution cl DDFM = KR;
  Repeated Period/sub = school(Trt) type = cs r rcorr;
  Weight Impscore;
  LSmeans Trt*Post/pdiff adj = bon;
Run;
```

Proc Mixed statement

The first statement calls Proc Mixed, requests that it use the data set “standard” and provide estimates of such goodness-of-fit parameters (ic) as the AIC, AICC, and BIC, the ratio of the estimates of the covariance parameters, their statistical significance, and confidence limits (cl). Box 11.1 below displays the unweighted estimates of the covariance parameters when the SPI is the response. The notes at the bottom of the box define the parameters.

Box 11.1 Covariance Parameter Estimates Not Weighted by Implementation Scores, SPI is Response

Cov Parm	Subject	Ratio	Estimate	Standard Error	Z Value	Pr Z	Alpha	Lower	Upper
CS	School (Trt)	1.031	31.4078	11.5239	2.73	0.0064	0.05	8.8214	53.9942
Residual		1.000	30.4708	4.7102	6.47	<.0001	0.05	22.9981	42.3057

CS School(Trt) = 31.4078 is the estimate of the covariance between two measures of *Y* on the same school at all pairs of times. Here it is the between-school covariance component $\sigma_s^2 = \rho\sigma^2$ = Cov [*Yijk Yijk'*], where Var[Y_{ijk}] = σ^2 (Littell et al. 2006, 171).

Let σ_b^2 be the estimate of the residual variance component which here is 30.4078. It is the variance of Y_{ijk} conditional on a school: Var[$Y_{ijk}|i$] = σ_b^2. Thus $\sigma^2 = \sigma_s^2 + \sigma_b^2$ and ρ = covariance/variance $= \sigma_s^2/(\sigma^2) = \sigma_s^2/(\sigma_s^2 + \sigma_b^2)$, see Littell et al. (2006), 172.

Null model likelihood-ratio test

SAS determines the significance of σ_s^2 by applying a null model likelihood-ratio test. It compares the –2 × residual log likelihood (hereafter –2RLL) of the model that includes σ_s^2 with an otherwise identical model in which this covariance parameter is absent. That is, it compares the more complex model to the independent errors model that only has a constant variance on the main diagonal and no off-diagonal covariances: $\Sigma = \sigma^2 I$, where I is the identity matrix that has ones on the main diagonal and zeros elsewhere. For these data the difference in degrees of freedom is 1, the $\chi^2 = 29$, and the probability $p < .0001$ decisively rejects the simpler model. An appropriate model for the SPI should include nonzero off-diagonal covariances.

Class statement

The Class statement defines as attributes the treatments, the study period of 4 years, the schools, and the post time period; the latter will compare the last year to the first year of the evaluation period. Class will create indicator variables for Trt and Post.[6]

Model statement

The Model statement specifies that the school performance measure (SPI) is the response variable and that the variables to the right of the equal sign compose the explanatory structure of the model. The variables Trt, Post, and Trt×Post operationalize the DID design and are not centered by their means, but the fixed structural covariates—the controls—are centered by their means. The following options appear after the / : Solution (or "S") requests that SAS display the right-hand-side variables along with their estimated coefficients and their statistical significance; cl requests their confidence limits. The DDFM = KR option, which all of the substantive runs use, requests the KR correction for standard errors, *F*-statistics, and degrees of freedom (Littell et al. 2006, 188).[7] This correction may slightly decrease the size of the p-values of the treatment effects but, because of the increased degrees of freedom, it will reduce to reasonable values the effect sizes that will be calculated by the correlational method.

Repeated statement

The key elements of the Repeated statement specify the four school years as the Period of the times series; these options follow the / : the Subjects are the 31 schools nested within the three treatment groups, the Type of covariance structure for this response variable (but not necessarily for the other response variables) is compound symmetry (CS), r requests a display of estimated R matrix for a school nested by its treatment, and rcorr requests the display of the correlation matrix derived from the R matrix. [8]

The R Matrix display of Box 11.2 illustrates the compound symmetry covariance structure: all on-diagonal variances are equal and all off-diagonal covariances are equal: Littell et al. (2006, 174–177) and Singer and Willett (2003, 250–263) name as Σ the matrix composed of the covariance parameters that are on and above the diagonal in a matrix such as that of Box 11.2. These Σ matrices contain the various covariance structures that are specified by the Type = covariance structure option of the Repeated and Random statements in Proc Mixed and of the _Residual_ and Random statement in Proc Glimmix. In Σ notation, the matrix of Box 11.2 for compound symmetry is:

Box 11.2 Estimated R Matrix for School(Trt) for Darlington, Not Weighted by Impscore

Row	Column 1	Column 2	Column 3	Column 4
1	61.8786	31.4078	31.4078	31.4078
2	31.4078	61.8786	31.4078	31.4078
3	31.4078	31.4078	61.8786	31.4078
4	31.4078	31.4078	31.4078	61.8786

The off-diagonal element σ_s^2 = CS School(Trt) = 31.4078, it is the covariance between observations of Y on the same school at all pairs of times. The Residual = $\sigma_b^2 = 30.4708$, see Box 11.1. Their sum is $\sigma^2 = \sigma_s^2 + \sigma_b^2 = 61.8786$, which is Var[$Y_{ijk}$] = σ^2, the diagonal variance element.

$$\sum = \begin{bmatrix} \sigma^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ & \sigma^2 & \sigma_s^2 & \sigma_s^2 \\ & & \sigma^2 & \sigma_s^2 \\ & & & \sigma^2 \end{bmatrix} \tag{8}$$

For this CS structure there are only two parameters: the main diagonal cells have the same variance σ^2 ($= \sigma_s^2 + \sigma_b^2$), and the off-diagonal cells have the same covariance σ_s^2, which is the covariance between observations on the same unit at all pairs of times.

The division of each parameter in Box 11.2 by the diagonal variance results in this RCORR matrix of Box 11.3:

Box 11.3 Estimated R Correlation Matrix for School(Trt) for Darlington, Not Weighted by Impscore

Row	Column 1	Column 2	Column 3	Column 4
1	1	0.5076	0.5076	0.5076
2	0.5076	1	0.5076	0.5076
3	0.5076	0.5076	1	0.5076
4	0.5076	0.5076	0.5076	1

The correlation ρ = covariance/variance $=\sigma_s^2/(\sigma^2) = \sigma_s^2/(\sigma_s^2+ \sigma_b^2)$, and $\sigma^2/\sigma^2 = 1$.

The resulting Σ matrix, which equals that of expression (11.8) is:

$$\sum = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{bmatrix} \tag{9}$$

When the matrix is multiplied by the constant σ^2, that constant will appear on the main diagonal and the products $\sigma^2\rho$ will equal the covariance, σ_s^2. Recall that correlation ρ = covariance/variance = $\sigma_s^2/(\sigma^2)$, and therefore $\sigma_s^2 = \sigma^2\rho$. Thus, this CS structure depends on only two parameters, σ^2 the common variance, and ρ, the common constant correlation between pairs of repeated measurements on the same unit at different points in time. Because this CS structure is so parsimonious, the goodness-of-fit statistics do not impose a severe penalty; CS is often preferred to more complex covariance structures that use more free parameters.

Weight statement

Implementations of innovations often vary from site to site. To adjust for the different quality of the implementations of the reforms in the target schools, this chapter uses this Weight statement to weight the data by a school's implementation score:

```
Weight Impscore;
```

The illustrative results in Boxes 11.1 through 11.3 are not based on the use of the Weight statement. Consequently, the additive properties of the CS covariance components (i.e., $\sigma^2 = \sigma_s^2 + \sigma_b^2$) hold true. When the data are weighted by Impscore, these simple relationships among the covariance parameters become more complex. However, the subsequent SAS runs in this chapter do implement the Weight statement. After deleting the Weight statement, the interested reader could replicate the analysis using the available data sets and SAS code and compare the results.

Least-squares means

This least-squares means statement requests that SAS display the predicted means for the pre and post periods for each of the three treatment groups:

```
LSmeans trt*post/pdiff bon;
```

Figure 11.1 and Figs. 11.3 through 11.9 report these trends along with estimates of the two treatment effects ($\hat{\delta}_{\mathrm{t}}$ and $\hat{\delta}_{\mathrm{nm}}$) and their statistical significance. The options request comparisons between the various means and the significance of these differences without and with Bonferroni adjustments for multiple comparisons. However, these tests of significance depend in part on the appropriateness of the assumed covariance structure.

Selecting Covariance Structures

Because the covariance structure of a multilevel model influences the size of the treatment effects and their statistical significance, it must be chosen with care. When response variables are normally distributed and the REML procedure estimates the parameters of the multilevel model, such goodness-of-fit statistics as the AIC, AICC (i.e., the AIC corrected for bias), and the BIC can help the modeler select which of several candidate covariance structures best fit the data, given the same set of structural variables in the multilevel models. This section discusses a direct approach that relies on the goodness-of-fit statistics and the Covtest approach that relies on likelihood-ratio tests.

The direct approach

For selecting preferred covariance structures, the modeler requests that Proc Mixed estimate the parameters of a number of models that have different but reasonable covariance structures; the modeler then chooses as preferred a reasonable covariance structure that produces the lowest values of the AIC, AICC, or BIC. Table 11.4 applies this approach to select the covariance structure for the modeling of the SPI; note that the sizes of the treatment effects and their significance vary from model to model. However, the fit statistics point to the CS structure as preferred, compared with the other candidate structures. The first column of Table 11.4 presents the parameter estimates for the saturated unstructured covariance model, the subsequent columns report the parameter estimates for structures that range from simple to more complex.

Unstructured (UN) covariances

The UN covariance model fits the data perfectly but at the expense of parsimony. Because this model imposes no mathematical structure on the covariances, there will be $K(K + 1)/2 = 10$ unique parameters with K equal to the number of time

Table 11.4 Testing covariance structures using goodness-of-fit statistics from Proc Mixed

	Unstructured	Independent	Banded main	Heterogeneous	Compound	VC structured	Autoregressive
Parameters useful	covariance model	Errors	diagonal	compound	symmetry	as CS[a]	of order One
for these tests	UN	IE	UN(1)	symmetry CSH	CS	VC + CS	AR(1)
–2Res LL	737.7	775.3	773.4	745.4	747.7	748.5	760
AIC	757.7	777.3	781.4	755.4	751.7	752.5	764
AICC	760	777.4	781.8	755.9	751.8	752.6	764.1
BIC	772.4	778.8	787.3	762.7	754.6	755.4	766.9
Estimate of δ_t	7.39	7.04	7.12	6.83	6.74	6.78	6.75
Significance of δ_t	$DF = 27$	$DF = 27$	$DF = 27$	$DF = 27$	$DF = 27$	$DF = 81$	$DF = 27$
	$t = 2.98$	$t = 1.67$	$t = 1.80$	$t = 2.28$	$t = 2.12$	$t = 2.13$	$t = 1.79$
	$p = 0.006$	$p = 0.106$	$p = 0.083$	$p = 0.031$	$p = 0.043$	$p = 0.036$	$p = .084$
Estimate of δ_{nm}	–0.43	–0.15	0.31	–0.33	–0.81	–0.81	–2.18
Significance of δ_{nm}	$DF = 27$	$DF = 27$	$DF = 27$	$DF = 27$	$DF = 27$	$DF = 81$	$DF = 27$
	$t = -.20$	$t = -.15$	$t = .09$	$t = -.13$	$t = -.29$	$t = -.29$	$t = -.67$
	$p = 0.84$	$p = 0.97$	$p = 0.93$	$p = 0.90$	$p = 0.77$	$p = 0.77$	$p = .50$
DF of Trt × Post	Num =2,	Num =2,	Num =2,	Num =2	Num =2,	Num =2,	Num =2,
	Den = 27	Den = 27	Den = 27	Den = 27	Den = 27	Den = 81	Den = 27
Significance of	$F = 6.75$	$F = 2.04$	$F = 2.18$	$F = 3.96$	$F = 3.76$	$F = 3.79$	$F = 3.51$
Trt × Post	$\Pr > F =$	$\Pr > F =$	$\Pr > F =$	$\Pr > F =$	$\Pr > F =$	$\Pr > F =$	$\Pr > F =$
	0.004	0.15	0.133	0.031	0.036	0.027	.044
Null Model LL	$DF = 9$	$DF = 1$	$DF = 4$	$DF = 4$	$DF = 1$	$DF = 1$	$DF = 1$
Ratio Test	$\chi2 = 37.6$	$\chi2 = 0$	$\chi2 = 1.93$	$\chi2 = 29.99$	$\chi2 = 27.7$	$\chi2 = 26.9$	$\chi2 = 15.4$
	$p < .0001$	$p = 1.000$	$p = 0.749$	$p < .0001$	$p < .0001$	$p < 0.0001$	$p < .0001$

[a] This VC + CS model is specified by: Random Intercept/Subject = School(Trt) Type = VC S V. It is a CS Model. All of the other models are specified by the Repeated statement in Proc Mixed.

points that define the number of rows and columns of the matrix. Box 11.4 illustrates the unstructured RCORR matrix for the SPI; all of its parameters are unique:

Box 11.4 Estimated UN Matrix for School(Trt) for Darlington, Weighted by Impscore

Row	Column 1	Column 2	Column 3	Column 4
1	36.8196	28.0986	21.4079	17.9514
2	28.0986	58.0047	14.0921	28.9702
3	21.4079	14.0921	35.9985	15.9196
4	17.9514	28.9702	15.9196	39.7485

Variances are in the main diagonal cells; covariances are in the off-diagonal cells.

The UN Σ matrix of variances and covariance looks like this:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2 \\ & \sigma_2^2 & \sigma_{23}^2 & \sigma_{24}^2 \\ & & \sigma_3^2 & \sigma_{34}^2 \\ & & & \sigma_4^2 \end{bmatrix} \tag{10}$$

This UN model fits the data very closely, as indicated by a reading along the row of Table 11.4 that presents the estimates of –2RLL. Even though the UN model has the lowest value of the –2 RLL, the values of all three fit statistics—the AIC, AICC, and the BIC—punish this structure for its lack of parsimony. Except for the over-simplified diagonal models— the VC and UN(1)—the fit statistics for the UN have higher values than those for the other models (low values indicate a better fit). Because this UN model uses ten parameters, these goodness-of-fit statistics severely penalize it and point to the CS model, which uses only two parameters, as preferred.

For a model's use of free parameters, these three goodness-of-fit measures impose penalties of varying severity. The AIC imposes the least severe penalty and usually supports the selection of more complex models as preferred. The BIC imposes the most severe penalty and usually supports the selection of more parsimonious models as preferred. The AICC corrects the AIC for bias; its penalty is intermediate. By selecting complex models, which might include superfluous parameters, the AIC could accept false models, thereby increasing Type II errors. Contrarily, by selecting simpler models that might be incomplete, the BIC could reject true models, thereby increasing Type I errors. Thus, to control Type I error use the AIC; if loss of power is crucial, use the BIC. To balance Type I and Type II errors, the AICC might be the better of these three statistics.

Variance Components (VC)

If the Repeated statement is used to model these repeated measures, and if a random statement is not used to define other covariance parameters, then the VC model is composed of a constant main-diagonal variance, σ^2. Because the covariance parameter $\sigma_s^2 = 0$ and $\sigma^2 = \sigma_b^2$, the resulting model has independent errors; $\Sigma = \sigma^2 I$:

$$\sum = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{bmatrix} \tag{11}$$

Earlier, the null model likelihood-ratio test indicated that the off-diagonal covariances were informative and a preferred model should include them. Because the VC model did not include these off-diagonal covariances, it did not fit the data. This finding is corroborated by the difference between the –2RLL of the VC model, which only includes the residual variance ($\sigma^2 = \sigma_b^2$) and that of the CS model, which includes two covariance parameters (σ_s^2 and σ_b^2). The following likelihood ratio test assesses the null hypothesis H_0 that there is no difference between the CS model and the simpler VC model: –2 log (Residual LL_{VC}/Residual LL_{CS}) = -2RLL_{VC} – -2RLL_{CS} = (775.3 – 747.7) = 28.3; which is clearly statistically significant: $\chi^2 = 28.3$, *df* =1, and the $p < .0001$. The CS model fits the data better than the simpler VC model, and also better than the banded main diagonal structure.

Banded main diagonal, UN(1)

Whereas the VC structure constrains the on-diagonal variances to be the same, the UN(1) allows them to be different, here is its $\mathbf{\Sigma}$ matrix:

$$\sum = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ & \sigma_2^2 & 0 & 0 \\ & & \sigma_3^2 & 0 \\ & & & \sigma_4^2 \end{bmatrix} \tag{12}$$

Because of its zero off-diagonal covariances and its four heterogeneous on-diagonal variances, this UN(1) model produces the worst values of the goodness-of-fit statistics among the structures this chapter considers.

Heterogeneous compound symmetry (CSH)

This model improves upon the UN(1) by combining the homogeneous off-diagonal covariances of the CS structure with the heterogeneous on-diagonal variances of the UN(1); here is its Σ matrix:

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ & \sigma_2^2 & \sigma_s^2 & \sigma_s^2 \\ & & \sigma_3^2 & \sigma_s^2 \\ & & & \sigma_4^2 \end{bmatrix} \tag{13}$$

The value of the –2RLL for this CSH structure is slightly smaller than that of the CS structure, 745.4 compared to 747.7, indicating a slightly better fit. But the CSH achieves this fit by using three more free parameters than the CS, which has eight. Consequently, the fit statistics charge a higher penalty; all three point to the CS structure as preferred.

Variance components plus compound symmetry

Instead of using a Repeated statement, the sixth model in Table 11.4 and that for Fig. 11.2 use this Random statement that displays the random effects for each school:

```
Random Intercept/Subject = School (Trt) Type = VC
Solution V VCORR;
```

Although this statement specifies a VC covariance structure, it produces the random-effects estimates using a structure nearly identical to CS: there are two covariance parameters; the on-diagonal variances are the same, as are the off-diagonal covariances:

$$\sum = \begin{bmatrix} \sigma^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ & \sigma^2 & \sigma_s^2 & \sigma_s^2 \\ & & \sigma^2 & \sigma_s^2 \\ & & & \sigma^2 \end{bmatrix} \tag{14}$$

This model fits the data almost as well as the explicit CS model, but there is one salient difference: there are 81 degrees of freedom for the treatment effects rather than the 27 that characterize the other models. This difference changes the calculated values of the effect sizes when the correlational method is used, as discussed later on. Because the CS fits slightly better and its specification is unambiguous, the CS is preferred to the VC with the CS structure; however, both are preferred to the AR(1).

First order autoregressive, AR(1)

The AR(1) structure assumes that pairs of observations that include measures that span adjacent distances in time exhibit larger correlations than observations on pairs of observations that span greater distances in time. Given r_{12}, it assumes that

$r_{13} = r_{12} \times r_{12} = {r_{12}}^2$ so that $r_{12} > \mathrm{r}_{13}$ and $r_{12} = r_{23}$. When the AR(1) structure is specified for these data, the resulting RCORR matrix of Box 11.5 has this form:

Box 11.5 Estimated AR(1) RCORR Matrix for School(Trt) for Darlington, Weighted by Impscore

Row	Column 1	Column 2	Column 3	Column 4
1	1	0.4366	0.1906	0.0832
2	0.4366	1	0.4366	0.1906
3	0.1906	0.4366	1	0.4366
4	0.0832	0.1906	0.4366	1

${r_{12}}^2 = \mathrm{r}_{13}$ and ${r_{12}}^3 = \mathrm{r}_{14}$ and $r_{12} = r_{23}$.

Although this fitted structure clearly exhibits the AR(1) assumption, comparisons of the goodness-of-fit statistics indicate that the CS structure is more appropriate; the $\mathrm{CS}_{\mathrm{BIC}}$ of 754.6 is considerably less that $\mathrm{AR(1)}_{\mathrm{BIC}}$ of 766.9. Moreover, the pattern of these AR(1) correlations does not correspond visually to the actual correlational pattern produced by the unstructured model. Here is the Σ matrix for the AR(1):

$$\sum = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & P \\ & & & 1 \end{bmatrix} \tag{15}$$

This direct approach can point to the preferred model when the response is normally distributed, as this chapter assumes. However, when the responses are not normal; for example, dichotomies, as they are in the next chapter, the appropriate logistic or probit regression models should be modeled using Proc Glimmix. But its log pseudolikelihood estimation procedures invalidate the use of these standard goodness-of-fit measures. Consequently, SAS developed a series of Covtest statements that can help the modeler choose which of several candidate models are preferred. To help prepare the reader for the complexities of the next chapter, this approach will be introduced here.

The COVTEST approach

SAS's Covtest statements for Proc Glimmix test alternative covariance structures for closeness of fit against the more comprehensive, unstructured covariance model. Modelers can use these tests even when standard goodness-of-fit statistics are inappropriate because the candidate models are estimated by log pseudolikelihoods, as recommended by SAS when response variable are not normally distributed. Proc Mixed cannot process these Covtest statements, but, as the direct approach illustrated, it can implement likelihood-ratio tests that compare the fit of competitive models. Moreover, if the response variable is normally distributed, and if the Model statement for a Proc Glimmix run specifies the distribution as Normal and the link as Identity, then Proc Glimmix will produce the same estimates as Proc Mixed, and Covtest statements can be used to help select the preferred covariance structures. The following pages briefly sketch these procedures for the data at hand, using Proc Glimmix.

SAS Institute (2005) models "G-side" and "R-side" random effects. The former, which are associated with the Random statement of Proc Mixed, appear in the **G** matrix. The "R-side" or "residual" random effects, which are associated with the Repeated statement in Proc Mixed and the _Residual_ key word of Proc Glimmix, appear in the **R** matrix. All of the subsequent Covtests are based on this statement:

```
Random _residual_ /type = unr sub = school(trt) s v vcorr;
```

This endnote displays an example of the SAS code that implements these tests. [9] All of the random effects in these runs are R-side; there are no G-side random effects.

For each of four candidate covariance structures and eight response variables, Table 11.5 reports the results of these covariance tests. For each response variable, along with the decisive BIC goodness-of-fit statistics obtained from the direct approach, the various columns of the table report the degrees of freedom (*df*) of the candidate covariance structure, that is, its number of unused or "free" parameters; the value of the −2 RLL; the difference in χ^2 between the candidate model and the UNR model; and the probability of fit of the candidate covariance structure. For each unique response variable, these statistics influence the choice of the preferred covariance structure.

By applying likelihood-ratio tests, the Covtest statements assess the relevance of the parameters of the full unstructured model; these parameters define the baseline for the comparisons. To this baseline, the tests compare the closeness of fit of the candidate covariance structures, which are simplified special cases of the full model. The column headings of Table 11.5 describe the candidate covariance structures and the Covtest results for four standard models: the unstructured "saturated" model (ZeroG) that has a full complement of on-diagonal variances and off-diagonal covariances ($df = 0$), these parameters can be the same (i.e., homogeneous) or different (i.e., heterogeneous); a banded main diagonal model [UN(1)] derived from the Independence test (Indep), that may have different on-diagonal variances but the off-diagonal covariances are zero ($df = 6$); a heterogeneous compound symmetry (CSH) model that can have the same or different on-diagonal variances but all off-diagonal covariances are the same ($df = 5$); and a compound symmetry (CS) model that has the same on-diagonal variances and the same off-diagonal covariances ($df = 8$).

Unstructured covariances, UNR

Visual displays of the parameters of the candidate models can help the modeler to choose the preferred covariance structures. By specifying the following Covtest statement that eliminates all G-side random effects (ZeroG), Proc Glimmix will display the parameters of the unstructured correlational model (UNR) when SAS implements the code presented earlier in endnote nine.

```
Covtest 'Ho: No G-Side Random Effects (UNR Parameters)'
ZeroG/cl;
```

Table 11.5 Tests of candidate covariance structures using Proc Glimmix's covtest statements and goodness-of-fit statistics from Proc Mixed

Response variable:	Unstructured (UNR) covariances test: Zero G-Side random effects	Banded main diagonal, UN(1), or VC test: diagonal only	Heterogeneous compound symmetry (CSH) test: homogeneous off-diagonal	Compound symmetry (CS) test: equal variances and equal covariances	BICs of the better models	Models selected
School performance						
DF	0	6	5	8	$CS_{BIC} = 754.6$	CS
–2 Residual LL	737.7	773.4	745.3	747.7	$CSH_{BIC} = 762.6$	
χ^2	0	35.69	7.6	10.0		
Probability of Fit	1	<.0001	0.18	0.27		
Average composite						
DF	0	6	5	8	$CS_{BIC} = 695.1$	CS
–2 Residual LL	678.16	713.45	685.71	688.13	$CSH_{BIC} = 703$	
χ^2	0	35.29	7.6	10.0		
Probability of Fit	1	<.0001	0.18	0.27		
Grade 3 composite						
DF	0	6	5	8	$CS_{BIC} = 753.7$	CS
–2 Residual LL	738.34	751.2	742.29	746.79	$CSH_{BIC} = 759.6$	
χ^2	0	12.86	3.95	8.45		
Probability of Fit	1	0.045	0.557	0.39		
Grade 3 reading						
DF	0	6	5	8	$VC_{BIC} = 793.3$	VC
–2 Residual LL	779.26	785.16	784.12	789	$CS_{BIC} = 795.9$	
χ^2	0	5.9	4.86	9.73	$Un(1)_{BIC} = 799$	
Probability of Fit	1	0.43	0.43	0.28		
Grade 3 mathematics [a]						
DF	0	6	5	8	$CS_{BIC} = 837.2$	CS
–2 Residual LL	812.87	830.72	825.57	830.26	$VC_{BIC} = 839.1$	
χ^2	0	17.85	12.7	17.4	$CSH_{BIC} = 843$	
Probability of Fit	1	0.01	0.026	0.026		
Grade 5 Composite						
DF	0	6	5	8	$CS_{BIC} = 710.4$	CS
–2 Residual LL	688.56	737.04	700.63	703.45	$CSH_{BIC} = 718.5$	
χ^2	0	48.48	12.07	14.89	$UN_{BIC} = 723.2$	
Probability of Fit	1	<.0001	0.034	0.061		
Grade 5 Reading						
DF	0	6	5	8	$CSH_{BIC} = 786.5$	CSH
–2 Residual LL	768.09	796.98	769.19	781.05	$CS_{BIC} = 788$	
χ^2	0	28.89	1.1	12.97		
Probability of Fit	1	<.0001	0.95	0.11		
Grade 5 Mathematics	VC	UN(1)	CSH	CS	$CS_{BIC} = 767.9$	CS
–2 Residual LL	790.4	788.6	755	761	$CSH_{BIC} = 772.2$	
AIC	792.4	796.6	765	765		
AICC	792.5	797	765.4	765.1		
BIC	793.9	802.5	772.2	767.9		

[a] For Grade 5 Mathematics the unstructured model does not estimate properly. Here the AICC selects the same preferred models as the BIC.

In its first row of data, Box 11.6 below reports the results produced by this Covtest statement; the other rows are produced by the other Covtest statements that this section explicates.

Box 11.6 Estimates of Covariance Parameters for the UNR, UN(1), CSH, and CS Candidate Covariance Structures, the School Performance Index (SPI) is the Response

Model	V1,1	V2,2	V3,3	V4,4	C2,1	C3,1	C3,2	C4,1	C4,2	C4,3
UNR	55.9	88.3	54.6	60.5	0.608	0.587	0.308	0.469	0.603	0.420
UN(1)	48.2	78.7	52.7	57.6	0.000	0.000	0.000	0.000	0.000	0.000
CSH	53.4	86.2	59.2	62.7	0.502	0.502	0.502	0.502	0.502	0.502
CS	64.8	64.8	64.8	64.8	0.494	0.494	0.494	0.494	0.494	0.494

The column headings indicate that the first four parameters are the estimates of the on-diagonal variances in positions row 1, column 1 through row 4, column 4 of the UNR matrix. The next six parameters are the estimates of the off-diagonal correlations indexed so that they are below the main diagonal in positions row 2, column 1 through row 4, column 3. The Type = UNR specification outputs these parameters in this order, whereas the Type = UN specification outputs the parameters in a different order. The different orders of the data affect the specification of the tests using the Covtest General statement as clarified later.

The visual inspection of the UNR row suggests that, except for the outlying estimate of V2,2, the on-diagonal variances are rather similar as are the substantial off-diagonal correlations. The latter parameters do not exhibit the pattern of declining correlations intrinsic to the AR(1) structure. Because this unstructured model uses the maximum number of covariance parameters, the goodness-of-fit statistics impose a large penalty that most often prevents this model from becoming a preferred choice. Even so, this test provides a check on the code and induces SAS to display the covariance parameters for the candidate models so that the analyst can visually assess their patterns, and the presence or absence of the AR(1) pattern; the Covtest statements cannot directly test for that structure.

For the fifth grade mathematics response variable, the UNR model does not estimate properly. Consequently, for this response the columns of Table 11.5 report the goodness-of-fit statistics for the candidate covariance structures. If the models include the same set of fixed covariates and REML estimation is used, then these fit statistics can also gauge the fit of the AR(1) structure compared with other candidate structures. These statistics indicate that the AR(1) does not fit these data.

Banded main diagonal, UN(1)

The second row of data in Box 11.6 reports the results of the Independence test (IND). This test compares the fit to the UN model of the UN(1) covariance structure, which has on-diagonal variances and zero off-diagonal correlations. This Covtest statement requests this test:

```
Covtest 'Ho: Independence = No G-side, Diagonal R-side'
Indep/cl;
```

The visual comparison of the UN(1) row of parameters with those for UNR indicates that their on-diagonal variances are rather similar, but the zero off-diagonal correlations of the UN(1) destroy its fit. The diagonal model produced by the Independence test can contain either heterogeneous on-diagonal variances as in a UN(1) model or homogeneous on-diagonal variances as in a VC model. Most often these models are too parsimonious to fit the full model closely, as is the case here: the difference in χ^2 is 35.7 and the difference in *df* is 6, producing a probability of fit $p < .0001$ for this UN(1) model.

Heterogeneous compound symmetry (CSH)

The third row of data in Box 11.6 reports the parameters for the heterogeneous compound symmetry (CSH) structure. Although this structure can have heterogeneous variances on the main diagonal, it requires homogeneous correlations in the off-diagonal cells. A visual comparison of its parameter estimates to those for the UNR suggests that the CSH model fits the UNR pattern very closely: the on-diagonal variances are near identical and the off-diagonal correlations are reasonably close. No unique single Covtest key word can directly test this visual observation. Instead, the following Covtest statement and its General specifications can produce the null model for this test; the code is coordinated to the elements of the UNR in Box 11.6:

```
covtest 'Ho:Homogeneous off-diagonal correlations, CSH '
General
   0 0 0 0 1-1,
   0 0 0 0 1 0 -1,
   0 0 0 0 1 0  0 -1,
   0 0 0 0 1 0  0  0 -1,
   0 0 0 0 1 0  0  0  0 -1/estimates;
```

Unlike the specifications for the CS structure that constrains the four on-diagonal variances to be equal, here the four leading zeros allow them to vary. The first 1 of a pair (1... –1) anchors the equality constraints using C2,1 as the base. The 1 –1 tests for equality the first (C2,1) and second (C3,1) correlations. The 10 –1 tests for equality the first (C2, 1) and third (C3,2) correlations. The 100 –1 tests for equality the first (C2, 1) and fourth (C4, 1) correlations. The 1000 –1 tests for equality the first (C2,1) and fifth (C4,2) correlations. Finally, the 10000 –1 tests for equality the first (C2,1) and sixth (C4, C3) correlations. The / estimates option requests the estimates. When the CSH model is compared to the UNR model, the difference in $df = 5$, the difference in χ^2 is 7.59 (745.3 – 737.71), and probability of fit $p = 0.1803$. The CSH fits the data more parsimoniously than the UNR, but less parsimoniously than the CS structure.

Compound symmetry (CS)

Visual comparisons of the parameters for the CS structure in Box 11.6 with the parameters for the other candidate structures suggests that it fits the UNR more closely than the UN(1) and less closely than the CSH. The latter closely reproduces both the heterogeneous variances and the similar off-diagonal correlations. There is no unique single Covtest key word that tests directly for the CS structure. However, the following Covtest statement and its General specifications, which are coordinated to the parameters of the UNR matrix in Box 6, produce the null model for this test:

```
Covtest ‘Ho: compound symmetry’ General
  1 -1,
  1  0 -1,
  1  0  0 -1,
  0  0  0  0 1 -1,
  0  0  0  0 1  0 -1,
  0  0  0  0 1  0  0 -1,
  0  0  0  0 1  0  0  0 -1,
  0  0  0  0 1  0  0  0  0 -1/estimates;
```

The first three lines of code test the equality of the four on-diagonal variances; the remaining lines, which are identical to those for the CSH test, constrain the off-diagonal correlations to be equal. In the first line the first 1 in the pair anchors the comparisons so that V1,1 is the baseline. The codes stipulate that V1,1 = V2,2; V1,1 = V3,3, and V1,1 = V4,4. Because the resulting CS model uses only two parameters, eight parameters are free and the likelihood-ratio test rewards this model for its parsimony: the $df = 8$, $\chi^2 = 9.97$ (747.68 − 737.71) and the probability of fit of $p = .27$ is higher than that for the CSH model. The BIC statistics also favor CS over CSH: $\mathrm{CS}_{\mathrm{BIC}}$ =754.6 and $\mathrm{CS}_{\mathrm{CSH}} = 762.6$; these statistics confirm the choice of the CS model as preferred for this response variable, it is also the preferred choice for five other response variables, see the last column of Table 11.5.[10]

Other preferred covariance structures

The reading tests for third grade and for fifth grade call for the VC and CSH covariance structures, respectively. The UNR model for third grade reading indicates that the off-diagonal covariances are not large and could be zero. Thus, the VC model could fit better than the CS structure that has homogeneous on-diagonal and nonzero off-diagonal parameters, and it could also fit better than the UN(1) that has zero off-diagonal covariances but heterogeneous on diagonal variances. The BIC difference favors VC over CS by –2.6; as does the null model likelihood ratio test that assess the importance of the covariance parameter for the off-diagonal covariances: the difference in –2 × RLL is 0.86 (789.86 − 789), the *df* difference is 1, and the null model fits with a probability of $p = 0.3537$. Moreover, the school

(Trt) covariance parameter is not statistically significant, $p = 0.415$, so the VC model is preferred to the CS model. It is also preferred to the UN(1) model: the BIC difference favors VC over the UN(1), 793.3 to 798.6 and the null model likelihood ratio test suggests that its homogeneous on-diagonal variances are preferred to the nuanced heterogeneous variances of the UN(1): the difference in –2 × RLL is 5.08 (789.9 – 784.8), the *df* difference is 3, and the simpler VC model fits with a $p = 0.1664$; the heterogeneous estimates of the variances are superfluous.

For fifth grade reading, the UNR model indicates that the on-diagonal variances are heterogeneous, ranging from 30.9 to 97.7; the CSH estimates track these variances very closely, ranging from 32.7 to 96.9, whereas the CS estimates are all 79.1. A likelihood-ratio test suggests that CSH model fits better than the CS model. The null H_0 of no difference between the two models is rejected in favor of the H_a that the CSH fits better than the CS: the difference in –2 × RLL is 11.86 and the difference in *df* is 3; consequently, the null hypothesis of equality is rejected at the $p = 0.01$ level of significance. The difference between the BIC statistics confirms the better fit of the CSH model compared to CS, 785.1 to 788.

The careful selection of preferred covariance structures is a prerequisite for the estimation of correct sizes of effects and tests of their significance. However, the multiplicity of response variables can lead to false findings, if not corrected.

Corrections for Multiplicity

Multiplicity arises because the sizeable numbers of response variables in an analysis capitalize on chance: by chance alone a finding may be statistically significant when in fact the underlying mechanism is not efficacious. The risk of accepting false hypotheses—Type II errors—increases with the increased number of response variables. In this study three families of response variables are vulnerable to such errors: The SPI and the CI are essentially the same measure of overall school performance. Because the reading and mathematics tests are components of the comprehensive test, the third grade comprehensive, reading, and mathematics tests are confounded, as are the fifth grade tests.

Meta-analytic techniques adjust for the multiplicity of the SPI and CI by forming one composite measure based on the weighted average of their treatment effects. The weight for each is the reciprocal of its squared standard error divided by the sum of the reciprocals of the squared standard errors (Fleiss 1981, 160–168).

The step-down Bonferroni option of Proc Multtest (SAS Institute Inc 1997) adjusts for the multiplicity of the third and fifth grade tests. This algorithm orders from lowest to highest the *p*-values for the three response variables of the set. It multiplies the lowest (i.e., best) *p*-value by 3, the number of variables in the set, obtaining the adjusted *p*-value. Then it multiplies the second lowest *p*-value by 2, the remaining number of variables in the set, obtaining a tentative estimate of the adjusted *p*-value. It compares this tentative estimate to the adjusted *p*-value for the first response variable, and chooses the larger of these estimates as the adjusted

p-value for the second response variable. Then, it multiplies by 1 the p-value for the remaining response variable in the set, obtaining a tentative estimate of the adjusted p-value for the third response variable. It compares this tentative estimate to the adjusted p-value for the second response variable and chooses the larger of these two estimates as the adjusted p-value for the third response variable. The algorithm adjusts the p-values for larger families of response variables following the same logic, which SAS formalizes as follows:

> Suppose the base test p-values are ordered as $p_1 < p_2 < \ldots < p_R$. The Bonferroni stepdown p-values $s_1, \ldots, s_R$ are obtained from
>
> $$\begin{aligned} s_1 &= Rp_1 \\ s_2 &= \max\,(\mathrm{s}_1, (R - 1)p_2) \\ \mathrm{s}_3 &= \max\,(\mathrm{s}_2, (\mathrm{R} - 2)p_3) \\ &\vdots \end{aligned}$$

These adjusted p-values influence interpretations of the effect sizes.

Effect Sizes

In educational research, the standardized effect size (ES) most often is estimated by the quotient of the difference between the means of the response variable in the treatment and comparison groups (i.e., the average treatment effect) divided by a standard deviation (SD) of the response variable. Opinions differ, however, about which of several SDs should be the devisor: the comparison group's SD, the target group's SD, or their pooled SD. For every response variable in this chapter, the SD of the not-matched group is higher than the SDs of the target and matched groups so that the overall pooled SD is a bit smaller than that for not-matched group. Because division by a larger number reduces the size of the quotient, this variability in the SDs could create problematical effect sizes. Avoiding the choice of which SD to use as the devisor, for estimating effect sizes this chapter applies the correlational approach (Rosenthal 1991, 19–20).[11] Namely,

$$r = \textit{square root}\ (t^2/(t^2 + df) \tag{17}$$

and given r, then

$$d = 2r/\textit{square root}\ (1 - r^2) \tag{18}$$

However, the number of degrees of freedom in the denominator of (17) can vary depending upon the specification of the covariance structure. For the same value of t^2, a large number of df will reduce the value of r, which in turn will reduce the d; a smaller number of df will increases r and increase d.

When the modeling uses the Repeated statement, to produce estimates of effect sizes similar to those of the standardization approach, the subsequent analyses apply the KR corrections for standard errors, F-statistics, and degrees of freedom. Consequently, the coefficients on the treatment effects will have about 87 *df* creating slightly smaller p-values that reduce Type-I errors. But their larger number of *df* produce considerably smaller effect sizes than those from the same repeated-measures models that are not KR-adjusted, here these have 27 *df*. For the same SPI response and the same multilevel model using the Repeated statement, Box 11.7 compares the parameter estimates for the KR-corrected and not-adjusted models, and, for comparison, the structured variance components (SVC) model of Fig. 11.2 and Table 11.4:

Box 11.7 KR-Corrected, Not-Adjusted (NA), and Structured VC (SVC) Effects on the SPI

	$\hat{\delta}$	SE	DF	t	Pr > t	ES r	ES d
KR $\hat{\delta}_t$	6.7391	3.183	86.9	2.12	0.0371	0.222	0.455
KR $\hat{\delta}_{\text{nm}}$	–0.8064	2.7611	88.2	–0.29	0.7709	–0.031	–0.062
NA $\hat{\delta}_t$	6.7391	3.1789	27	2.12	0.0433	0.378	0.816
NA $\hat{\delta}_{\text{nm}}$	–0.8064	2.7563	27	–0.29	0.7721	–0.056	–0.112
SVC $\hat{\delta}_t$	6.7768	3.1837	81	2.13	0.0363	0.230	0.473
SVC$\hat{\delta}_{\text{nm}}$	–0.8054	2.7594	81	–0.29	0.7711	–0.032	–0.064

The effect sizes calculated on the basis of the overall standard deviations are about 0.43 and –0.05 in standard deviation units. The NA effect sizes are out of line with the other estimates.

When the Repeated statement is used, the KR adjustment should be used in order to produce realistic effect sizes from the correlation method. When the structured variance components model is used, these adjustments are not necessary. All three models will produce reasonable effect sizes when their treatment effects are divided by appropriate standard deviations; in educational research reasonable positive effect sizes range from $d = 0.20$ to $d = 0.80$ in standard deviation units.[12]

Results

For the target, matched, and not-matched schools, the subsequent figures present estimates of their average achievement scores pre (SY 1996–1997) to post (SY 1999–2000), using as response variables the SPI, CI, and their reading and mathematics components; PROC MIXED provided the KR-corrected estimates of degrees of freedom and tests of significance. The first three figures depict how the treatment effects influenced overall school performance; the next three, achievement in third grade; and the following three, achievement in fifth grade. The final two figures present the summarizing estimates of effect sizes, with and without the KR corrections for degrees of freedom; Appendix Table A.11.1 and A.11.2 present the calculations of effect sizes in detail.

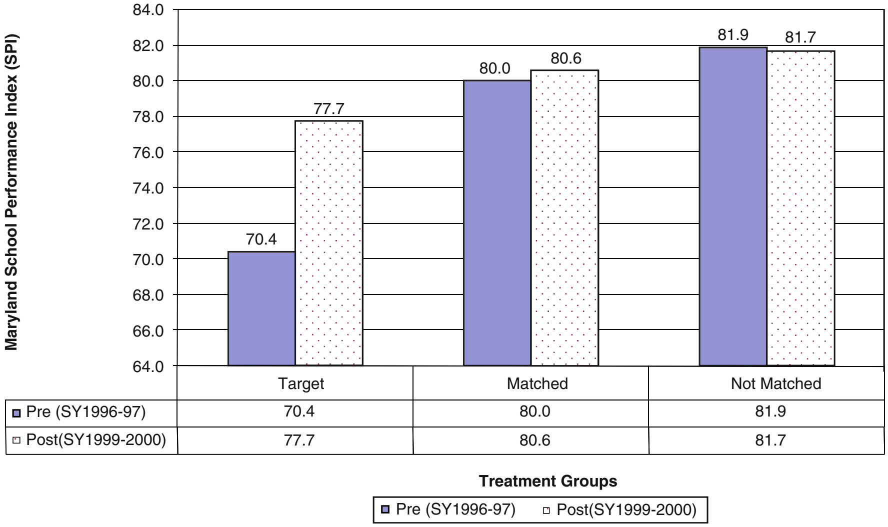
Overall School Performance

The pre-to-post changes in the test scores for the SPI, the CI, and their weighted average indicate the overall performance of the schools.

Figure 11.1 depicts the growth on the SPI in the target schools from an initial value of 70.4 to 77.7; the improvement in the matched and nonmatched schools was minimal. The average treatment effect $\hat{\delta}_t$= + 6.7 (t = 2.12, p = 0.037), resulting in an effect size correlation of r = 0.22 and an effect size d = 0.46 in SD units. Contrarily, the $\hat{\delta}_{\mathrm{nm}}$ = –0.81 (t = –0.29, p = 0.77) for an r = –0.032 and d = –0.064.

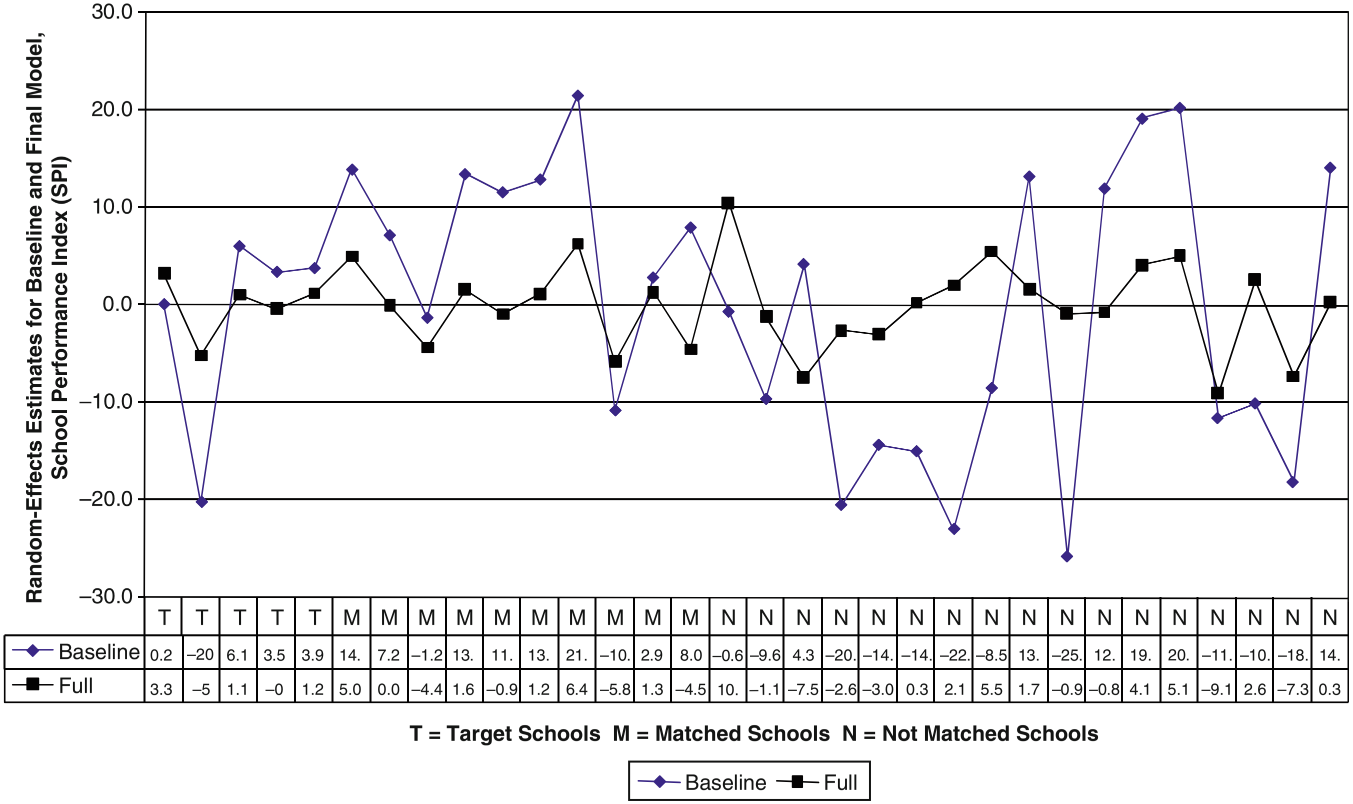
Across all 31 schools the estimates of the covariance parameters are statistically significant: $\hat{\sigma}_s^2$, the estimate of the covariance between two measures of the SPI on the same school at different times, equals 32.7 (p = 0.0069); and $\hat{\sigma}_b^2$, the residual variance component, equals 32.9 ($p < 0.0001$). That $\hat{\sigma}_s^2$ remains significant in the full model works against the idea that the reform treatment causes the improvement in the target schools, but a close inspection of the random-effects estimates for each school reveals the following: for the five target schools none of their random-effects estimates significantly deviate from the mean of zero; similarly, for the ten matched schools none of their random-effects estimates significantly deviate from the mean of zero; but, for the 16 non-matched schools, four random-effects estimates significantly deviate from zero. Apparently, the reform treatment in the target schools compared with the null treatment in the matched schools reduced the between-school random-effects estimates; the significant variability was due to the unmatched schools.[13]

Figure 11.2 depicts the random-effects estimates for the baseline model that includes only the intercept and for the full model that includes all of the structural



	Target	Matched	Not Matched
■ Pre (SY1996-97)	70.4	80.0	81.9
□ Post(SY1999-2000)	77.7	80.6	81.7

Fig. 11.1 Compared to matched schools, target schools significantly improved their School Performance Index (SPI) scores from SY 1996–1997 to SY 1999–2000 (δ_{t} = 6.7, t = 2.12, p = 0.037); not-matched schools did not improve (δ_{nm} = –0.81, t = –0.29, p = 0.77), CS-KR



	T	T	T	T	T	M	M	M	M	M	M	M	M	M	M	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Baseline	0.2	–20	6.1	3.5	3.9	14.	7.2	–1.2	13.	11.	13.	21.	–10.	2.9	8.0	–0.6	–9.6	4.3	–20.	–14.	–14.	–22.	–8.5	13.	–25.	12.	19.	20.	–11.	–10.	–18.	14.
Full	3.3	–5	1.1	–0	1.2	5.0	0.0	–4.4	1.6	–0.9	1.2	6.4	–5.8	1.3	–4.5	10.	–1.1	–7.5	–2.6	–3.0	0.3	2.1	5.5	1.7	–0.9	–0.8	4.1	5.1	–9.1	2.6	–7.3	0.3

Fig. 11.2 In the full model all the random-effects estimates for the target and matched schools do not vary significantly from the mean of zero.The variability in the not-matched schools is higher and some random effects are statistically significant

variables. The trend lines portray the extreme variability of the baseline model's estimates and the dampened variability of the full model's estimates, especially for the target (T) and matched schools (M).[14] The full models for the various response variables exhibit this same pattern: the random-effects estimates for the target and matched schools do not differ significantly from the mean of zero; the not-matched (N) schools most often exhibit significant variability.[15]

A school's overall CI score is the simple average of its fifth grade and the third grade CI scores. On this measure Fig. 11.3 depicts the change in the target schools from their initial value of 47.5 to 53.1; the change in the matched and not-matched schools is minimal. The $\hat{\delta}_t$= +5.24 (t = 2.17, p = 0.033); resulting in an r = 0.23 and a d = 0.46 in SD units. In contrast, the $\hat{\delta}_{\text{nm}}$ = –0.64 (–0.3, p = 0.762) for an r = –0.03 and a d = –0.06 in SD units.

Because the SPI and CI are very highly correlated indicators of overall school performance, this chapter conceptualizes them as two aspects of that concept, and measures the average treatment effects by taking a weighted average of the separate effects (Fleiss 1981, 160–168). The composite $\hat{\delta}_t$= + 5.77 (2, 9.55) for an ES = 0.43 (5.77/13.46). Contrariwise, the composite $\hat{\delta}_{\text{nm}}$ = –0.70 (–3.97, 2.57)—not significant; its ES = –0.052 (–0.70/13.46).[16]

Thus, for all three of these indicators, the reform treatment significantly improves the target group's overall school performance; the differences between the matched and not-matched groups are minimal and not statistically significant.

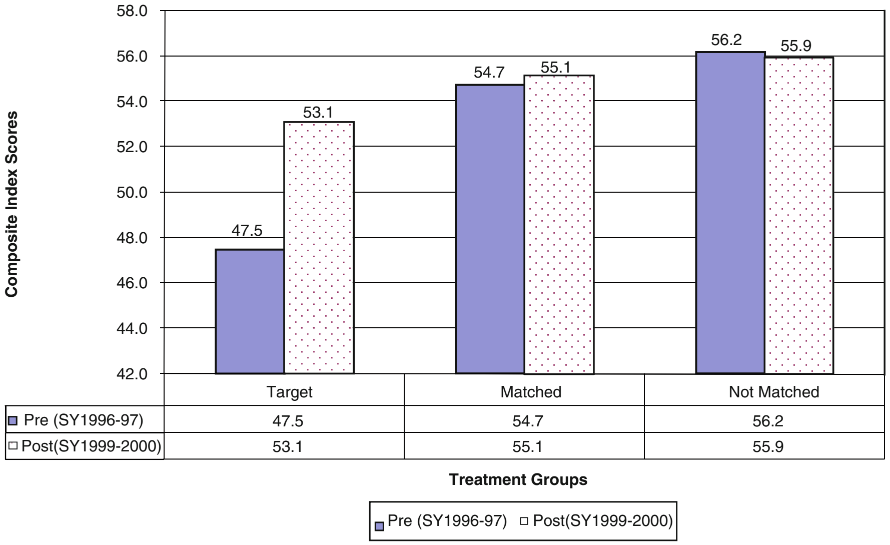
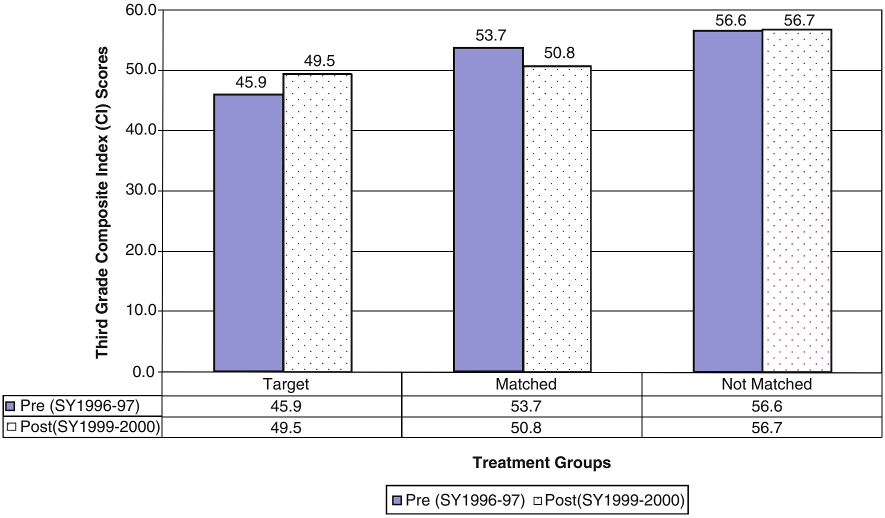


Fig. 11.3 Compared to matched schools, target schools significantly improved their Composite Index (CI) scores from SY 1996–1997 to SY 1999–2000 (δ_{t} = 5.2, t = 2.16, p = 0.034); not-matched schools did not improve (δ_{nm} = –0.64, t = –0.30, p = 0.76), CS-KR

Third Grade Achievement

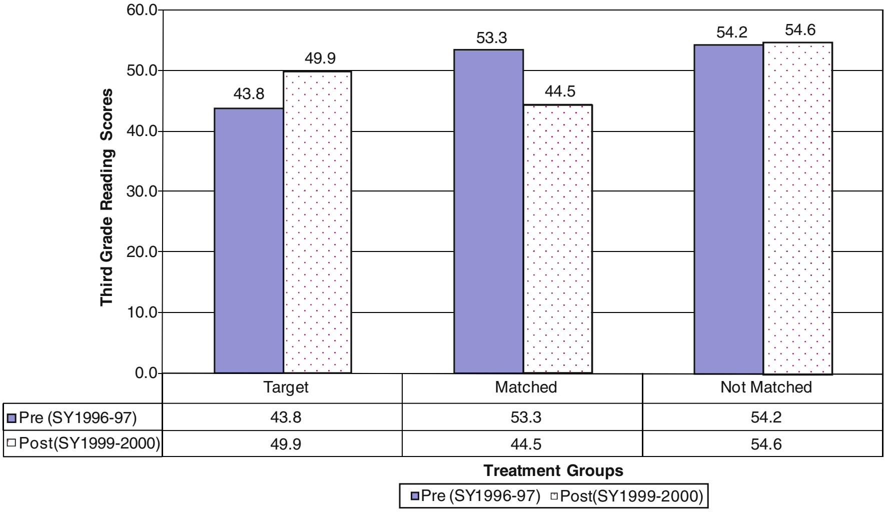
For the third grade CI scores, Fig. 11.4 depicts the positive change in the target schools, the negative change in the matched schools, and the lack of change in the not-matched schools. The target schools improved from 45.9 to 49.5 in third grade CI units, whereas the matched schools declined from 53.7 to 50.8. The difference between these differences results in a $\hat{\delta}_t = +6.46$ ($t = 1.90, p = 0.059$) that misses statistically significance at the $p < 0.05$ level. However, the effect sizes are substantial: $r = 0.20$ and $d = 0.41$ in SD units. Because the scores for the matched schools declined and the scores for the not-matched schools were constant, the statistics for the not-matched schools are not negative. The $\hat{\delta}_{\text{nm}} = +3.06$ ($t = 1.05$, $p = 0.30$)—not statistically significant. The effect size $r = 0.11$ and the $d = 0.22$ are both about one half the size of those estimates for the target schools.

Reading and mathematics are important components of the CI. As Fig. 11.5 indicates, the target schools significantly improved their performance on the third grade reading tests. Their reading scores improved 6.3 reading units, declined 8.9 in the matched schools, and held steady (+.5) in the not-matched schools. The combination of this noticeable increase in the target schools and the large decline in the matched schools generates a strong $\hat{\delta}_t = 15.0$ ($t = 4.5, p = 0.001$) that favors the reforms; the $r = 0.31$ and the $d = 0.64$ in SD units. However, the smaller increase in the not-matched schools coupled with the large decline in the matched schools also produces a noticeable positive effect, the $\hat{\delta}_{\text{nm}} = 9.3$ ($t = 2.41$, $p = 0.018$); the $r = 0.23$, and $d = 0.46$. But, the reform treatment effects are



	Target	Matched	Not Matched
▪ Pre (SY1996-97)	45.9	53.7	56.6
□ Post(SY1999-2000)	49.5	50.8	56.7

Fig. 11.4 Compared to matched schools, target schools improved their third grade Composite Index (CI) scores from SY 1996–1997 to SY 1999–2000 ($\delta_{\text{t}} = 6.5$, $t = 1.91$, $p = 0.059$); not-matched schools did not improve much ($\delta_{\text{nm}} = 3.06$, $t = 1.05$, $p = 0.30$), CS-KR



	Target	Matched	Not Matched
▣Pre (SY1996-97)	43.8	53.3	54.2
□Post(SY1999-2000)	49.9	44.5	54.6

Fig. 11.5 Compared to matched Schools, target schools significantly improved their third grade reading scores from SY 1996–1997 to SY 1999–2000 ($\delta_t = 15$, $t = 3.34$, $p = 0.001$); not-matched schools also improved ($\delta_{nm} = 9.33$ $t = 2.41$, $p = 0.018$), VC-KR

considerably larger than those for the null treatment: the treatment-effect difference = +5.7, the r-difference = +0.08, and the d-difference = +0.18.

Figure 11.6 reports that third grade mathematics scores declined less in target schools than in the matched schools. The resulting $\hat{\delta}_t = +9$ ($t = 1.78$, $p = 0.08$). This produces an $r = 0.19$ and a $d = 0.38$ in SD units. In the not-matched group, the $\hat{\delta}_{nm} = 6.7$ ($t = 1.54$, $p = 0.13$). The effect size correlation $r = 0.17$ and $d = 0.34$.

Without any corrections for multiplicity, the target schools' p-values for improvements on the third grade CI ($p = 0.0588$), reading tests ($p = 0.0012$), and mathematics tests ($p = 0.0788$) are all noticeably more favorable than those for the not-matched schools (respectively; $p = 0.2988$, $p = 0.0178$, and $p = 0.1269$). Moreover, their effect sizes are larger than those for the not-matched schools. But, the Bonferroni step-down procedure for controlling for this multiplicity finds that only the treatment effect of the reforms on the reading test scores retains its statistical significance. For the set of three reform treatment effects, the Bonferroni-corrected p-values are: $\text{CI}_{\text{T}} = 0.1176$; $\text{Reading}_{\text{T}} = 0.0036$, and $\text{Mathematics}_{\text{T}} = 0.1176$. Moreover, for the set of three not-matched null treatment effects, none of their Bonferroni-corrected p-values retain their significance: $\text{CI}_{\text{N}} = 0.2988$; $\text{Reading}_{\text{N}} = 0.0534$, and $\text{Mathematics}_{\text{N}} = 0.2538$. For the set of all six measures, the corrected p-values are $\text{CI}_{\text{T}} = 0.2352$, $\text{Reading}_{\text{T}} = 0.0072$, $\text{Mathematics}_{\text{T}} = 0.2364$, $\text{CI}_{\text{N}} = 0.2988$, $\text{Reading}_{\text{N}} = 0.0890$, and $\text{Mathematics}_{\text{N}} = 0.2538$.

Consequently, the educational reforms only improved the third grade reading tests scores; none of the other target or not-matched treatment effects for third grade is statistically significant at the $p < 0.05$ level.

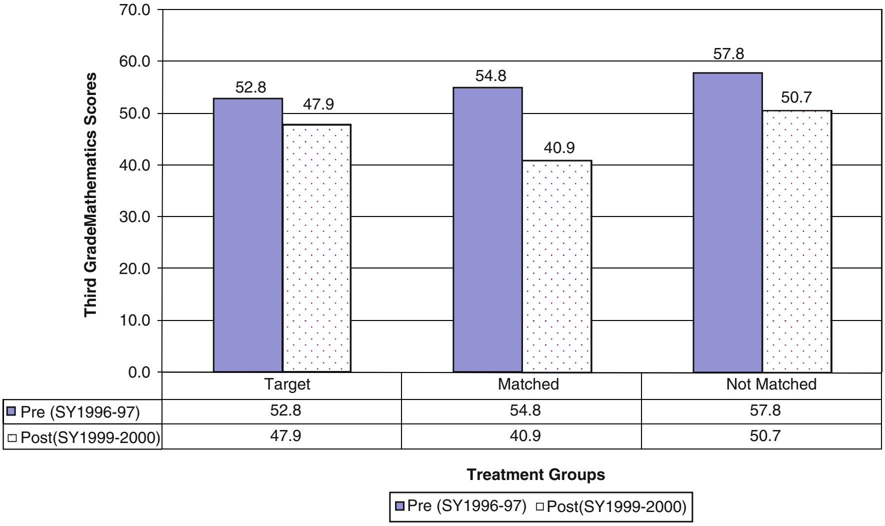


Fig. 11.6 Compared to matched schools, third grade mathematics scores declined less in target schools from SY 1996–1997 to SY 1999–2000 (δ_t = 9, t = 1.78, p = 0.079); and less in not-matched schools (δ_nm = 6.75, t = 1.54, p = 0.127), CS-KR

Fifth Grade Achievement

For the fifth grade CI, Fig. 11.7 depicts the target schools' overall improvement of about 8, which is larger than the matched schools' improvement of 4; the $\hat{\delta}_t$= 4 favors the reforms but not significantly (t = 1.58, p = 0.118), the r = 0.17 and d = 0.34 in SD units. Contrariwise, the scores for the not-matched schools declined slightly; the $\hat{\delta}_\text{nm}$ = −4.26 (t = −1.93, p = 0.057) for an r = –0.20 and a d = –0.41 in SD units.

On the fifth grade reading tests, the schools in all three treatment conditions improved, as Fig. 11.8 clearly shows. Consequently, the $\hat{\delta}_t$ = +4.15 lacks statistical significance (t = 1.25, p = 0.21); the r = 0.12 and d = 0.24. In contrast, the $\hat{\delta}_\text{nm}$ = –0.95 (t = –0.29, p = 0.77) for an r of –0.03 and a d = –0.06.

Figure 11.9 clearly depicts the improvement from 55.9 to 57.7 in the target schools' fifth grade mathematics test scores, and the declines in the matched and not-matched schools. The $\hat{\delta}_t$= 9.2 is statistically significant (t = 2.71, p = 0.008); the effect sizes are substantial: r = 0.29 and d = 0.60 in SD units. Contrarily, the $\hat{\delta}_\text{nm}$ = −3.1 (t = −1.07, p = 0.29), which produces an r = –0.12 and a d = –0.24.

Without any corrections for multiplicity, the target schools' p-values for its positive improvements on the fifth grade CI (p = 0.1179), reading tests (p = 0.2131), and mathematics tests (p = 0.0085) are all noticeably more favorable than those for the negative effects of the not-matched schools (respectively; p = 0.0551, p = 0.7437, and p = 0.2858). Moreover, their effect sizes are positive whereas those for the not-matched schools are all negative. Again, using the

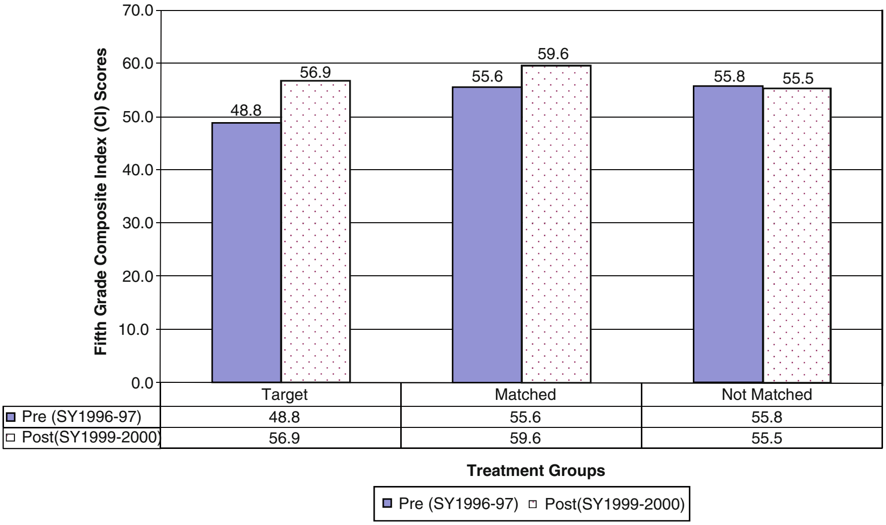


Fig. 11.7 Compared to matched schools, target schools improved their fifth grade Composite Index (CI) scores from SY 1996–1997 to SY 1999–2000 ($\delta_t = 4$, $t = 1.58$, $p = 0.118$); not-matched schools declined ($\delta_{nm} = -4.3$, $t = -1.94$, $p = 0.055$), CS-KR

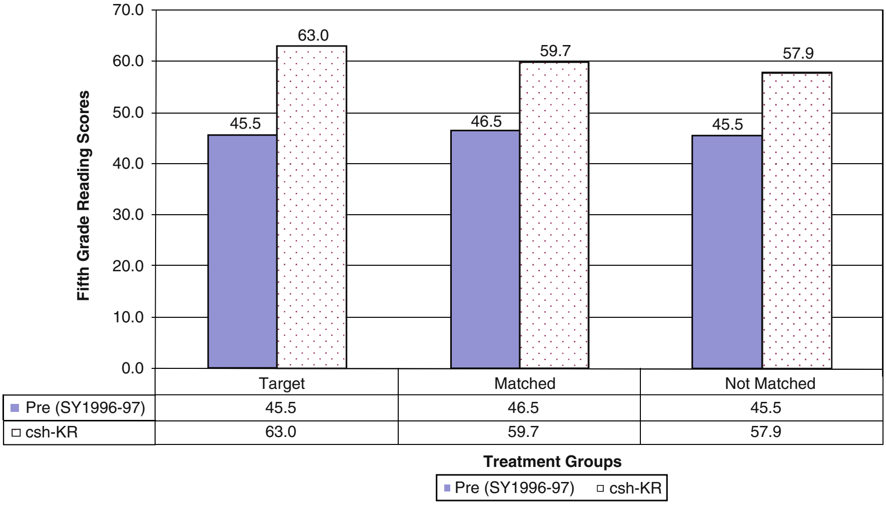


Fig. 11.8 Compared to matched schools, target schools slightly improved their fifth grade reading scores from SY 1996–1997 to SY 1999–2000 ($\delta_t = 4.15$, $t = 1.25$, $p = 0.213$); not-matched schools did not improve ($\delta_{nm} = -0.936$, $t = -0.33$, $p = 0.744$), CSH-KR

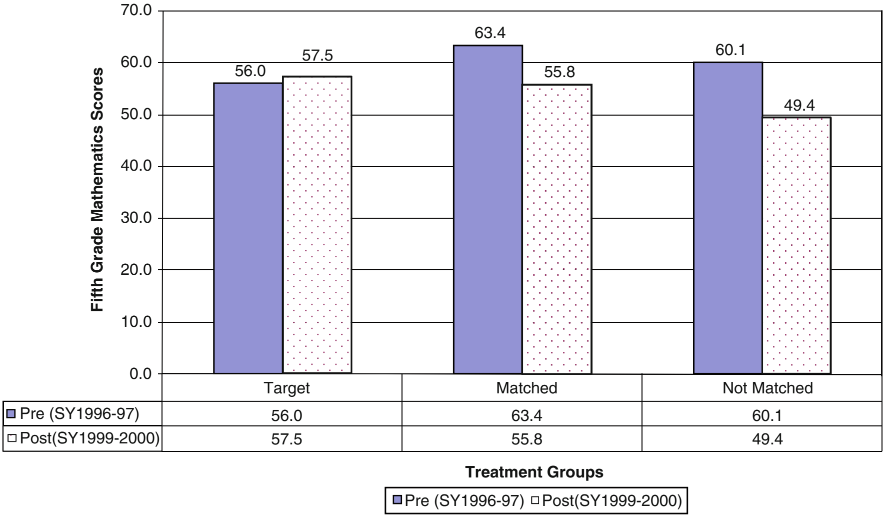


Fig. 11.9 Compared to matched schools, target schools improved their fifth grade mathematics scores from SY 1996–1997 to SY 1999–2000 ($\delta_t = 9.1$, $t = 2.7$, $p = 0.012$); not-matched schools declined ($\delta_{nm} = -3.15$, $t = -1.08$, $p = 0.29$), CS-KR

Bonferroni step-down procedure for the set of three positive treatment effects, the corrected p-values are: $\mathrm{CI}_T = 0.2358$; $\mathrm{Reading}_T = 0.2358$, and $\mathrm{Mathematics}_T = 0.0255$. Compared to the matched schools, the reform treatment improved the fifth grade mathematics test scores.

For the set of three probabilities associated with the negative effects of the null treatment of the not-matched group, the corrected p-values are: $\mathrm{CI}_N = 0.1653$; $\mathrm{Reading}_N = 0.7437$ and $\mathrm{Mathematics}_N = 0.5716$. Compared to the matched schools, the null treatment of the not-matched school did not distinguish between these two groups of schools.[17]

Discussion

Summary

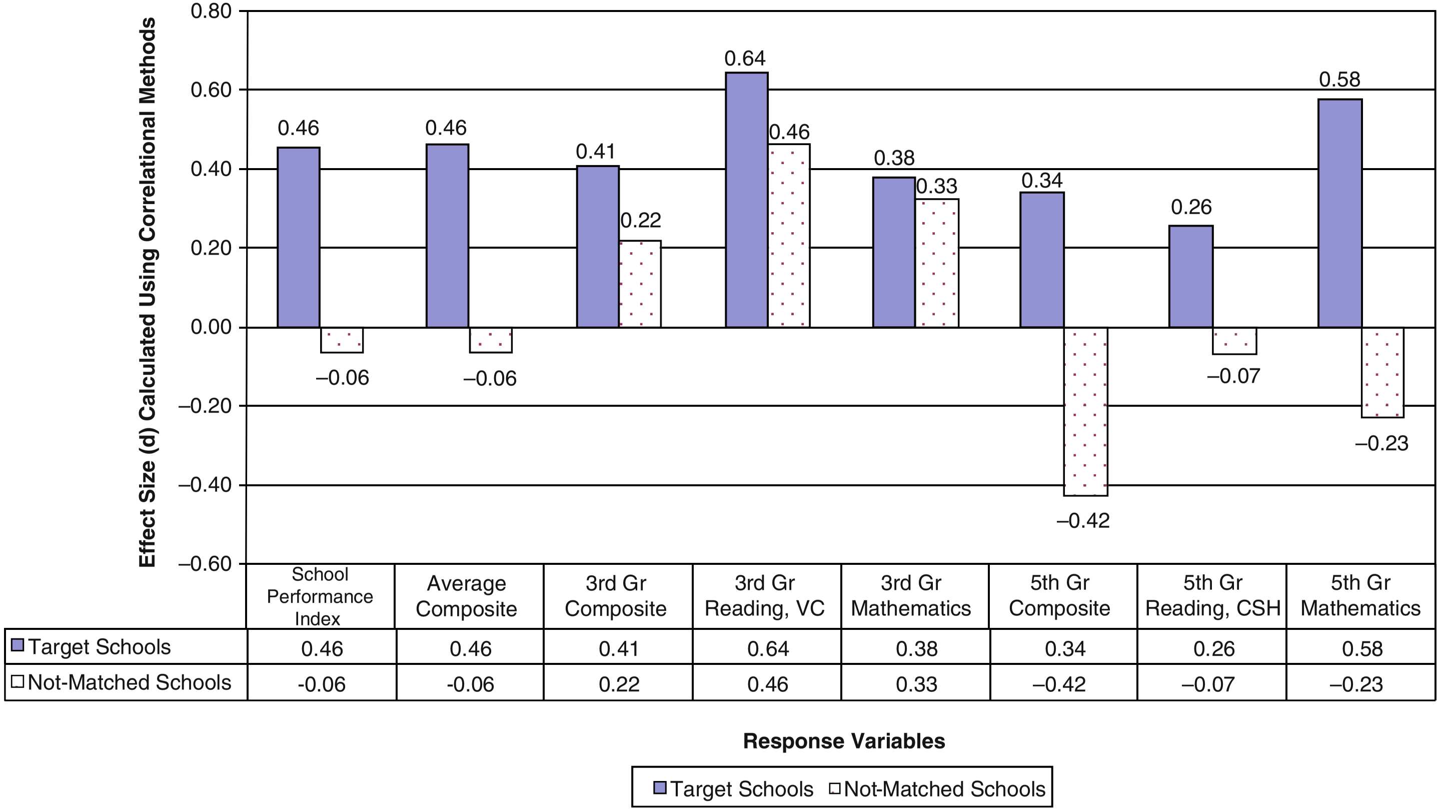
Practitioners of comprehensive school reform premise that underperforming schools can be changed and underperforming students can improve their intellectual achievement. This chapter has tested this policy orientation by studying the effects on repeated measures of student achievement of comprehensive school reforms as provided by external consultants. The reforms emphasized project-based learning, standards-based assessment, small teams of teachers and students, and the effective use of technology. Five elementary schools in Harford County, Maryland, were exposed to these reforms for a three year period of the consultant's maximum engagement; ten closely matched schools were not exposed to these reforms. Although the quality of the implementations varied from school to school, the step-down Bonferroni corrections for the multiplicity of the response variables indicate that the consultants did significantly improve the overall performance of the target schools, and their students' scores on the third grade reading and fifth grade mathematics tests, compared with the change in the matched schools. Other positive effects of the reforms produced substantial effect sizes that are based on noticeable positive treatment effects that did not attain statistical significance.

For all these results, Fig. 11.10 portrays the effect sizes d that were calculated using the correlational approach. For each response variable, the bars depict the effect size of the reform treatment in the target schools relative to the matched group of schools, and also the effect size of the null treatment in a not-matched group of schools relative to the matched group. All but two of these models used the compound symmetry covariance structure but all applied the KR corrections for degrees of freedom. These corrections slightly improved the p-values of the effects and reduced the effect sizes of the reform treatment to reasonable values, ranging in units of standard deviation from a $d = 0.64$ for third grade reading to a $d = 0.26$ for fifth grade reading. In every comparison the treatment effects of the reforms relative to the matched schools are favorable. Moreover, these effects are always more favorable than the effects of the not-matched schools relative to the matched schools, both of these groups received the null treatment.

Implications

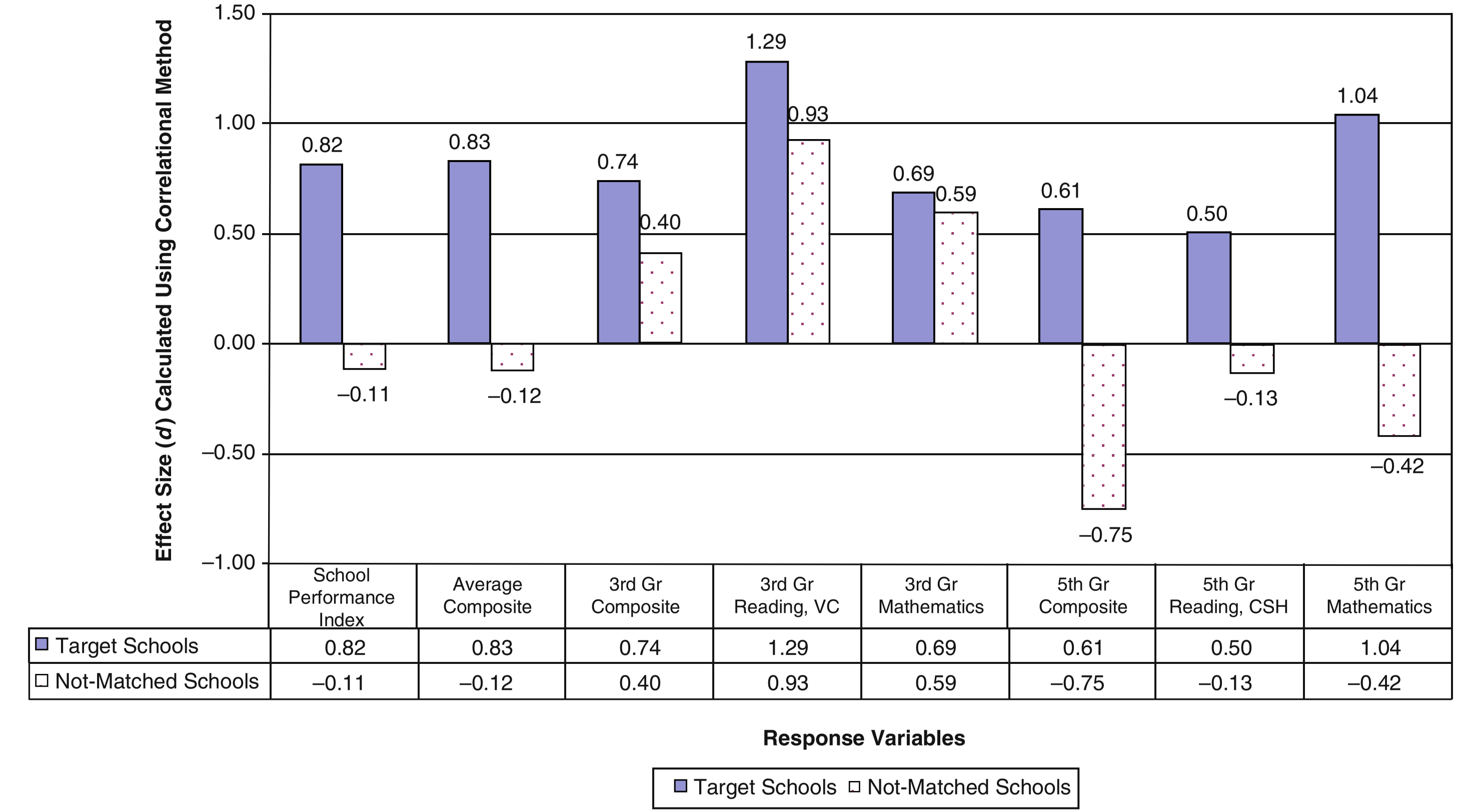
Effect sizes can vary depending upon the calculation method; their magnitudes should not be reified. The equations for the calculation of the effect size r and d include the degrees of freedom. Holding constant the other quantities in the equation, smaller values of the degrees of freedom create larger estimates of effect sizes, as Fig. 11.11 documents. It presents the effect sizes for the same models of Fig. 11.10, but without the KR corrections; in this figure, each treatment effect has 27 degrees of freedom which is much less than the 87 or so when the corrections are applied. Here the effect sizes are twice as large, ranging in units of standard deviation from a $d = 1.29$ for third grade reading to a $d = 0.50$ for fifth grade reading. Using either metric, the activities of the change agents brought about improvements in the schools.

Selection of participants in a study can bias the results of quasiexperimental studies. To minimize such bias this chapter built incrementally on an earlier study. That study closely matched the target schools with other schools and disregarded schools that were not matched; few effects were favorable to the reforms. This chapter applied a difference-in differences study design, added more observations on the target and matched schools, pooled the data including not-matched schools, and borrowed strength by estimating average treatment effects in



	School Performance Index	Average Composite	3rd Gr Composite	3rd Gr Reading, VC	3rd Gr Mathematics	5th Gr Composite	5th Gr Reading, CSH	5th Gr Mathematics
■ Target Schools	0.46	0.46	0.41	0.64	0.38	0.34	0.26	0.58
□ Not-Matched Schools	-0.06	-0.06	0.22	0.46	0.33	–0.42	–0.07	–0.23

Fig. 11.10 When compared to matched schools, in every comparison the effect sizes favor the target schools over the not-matched schools, with Kenward-Roger (KR) corrections for degrees of freedom and standard errors



	School Performance Index	Average Composite	3rd Gr Composite	3rd Gr Reading, VC	3rd Gr Mathematics	5th Gr Composite	5th Gr Reading, CSH	5th Gr Mathematics
■ Target Schools	0.82	0.83	0.74	1.29	0.69	0.61	0.50	1.04
□ Not-Matched Schools	–0.11	–0.12	0.40	0.93	0.59	–0.75	–0.13	–0.42

Fig. 11.11 When compared to matched schools, in every comparison the effect sizes favor the target schools over the not-matched schools, no Kenward-Roger (KR) corrections for degrees of freedom and standard errors

multilevel models with appropriate covariance structures. It compared the effects of the reform treatment to the effects of the null treatment of the matched schools, and the effect of the null treatment of the not-matched schools to the effect of the null treatment in the matched schools. That this null treatment always produced smaller effects than the reform treatment served to strengthen causal inferences. The next chapter advances this approach by applying propensity scores to reduce selection bias in an evaluation of Co-nect's reforms in Houston, Texas, elementary schools.

Acknowledgements The author thanks Sarah Cook and Dr. Bruce Goldberg of Co-nect for facilitating this research and encouraging its publication; Nancy Harkins for sharing information about her Harford County school system; and Philip Gibbs of the SAS Institute for clarifying nuances of PROC MIXED. Dr. Elizabeth Newton critiqued an early version of the research report; a latter version was favorably peer-reviewed by Aladjem et al. (September 2006), who included its results in their meta-analysis of comprehensive school reform. The views expressed in this chapter are the author's own and do not necessarily reflect those of the above mentioned people or organizations.

Appendix

Table A.11.1 Effect size calculations and random-effects estimates using Kenward-Roger (KR) corrections for degrees of freedom

	Steps in calculations		Fig. 11.1 SPI	Fig. 11.3 CI	Average of SPI & CI	Fig. 11.4 3rd Grade CI	Fig. 11.5 3rd Grade Read	Fig. 11.6 3rd Grade Math	Fig. 11.7 5th Grade CI	Fig. 11.8 5th Grade Reading	Fig. 11.9 5th Grade Math
Correlational approach											
Target Effect δt			6.739	5.216	5.978	6.485	15.022	9.014	4.018	4.151	9.112
Standard Error			3.183	2.419	2.801	3.3875	4.5006	5.0685	2.5443	3.3114	3.382
DF			86.9	87	87	88.6	108	88.2	86.7	96.9	87
t Value			2.12	2.16	2.13	1.91	3.34	1.78	1.58	1.25	2.69
Probability			0.037	0.034	0.036	0.059	0.001	0.079	0.118	0.213	0.008
	t^2	1	4.494	4.666	4.555	3.648	11.141	3.168	2.496	1.571	7.260
	t^2 + DF	2	91.39	91.666	91.555	92.248	119.141	91.368	89.196	98.471	94.260
	rsquare	1/2	0.049	0.051	0.050	0.040	0.094	0.035	0.028	0.016	0.077
	r	sqroot	0.222	0.226	0.223	0.199	0.306	0.186	0.167	0.126	0.278
	2r	3	0.444	0.451	0.446	0.398	0.612	0.372	0.335	0.253	0.555
	$1 - r^2$	4	0.951	0.949	0.950	0.960	0.906	0.965	0.972	0.984	0.923
	$\surd(1 - r^2)$	5	0.975	0.974	0.975	0.980	0.952	0.983	0.986	0.992	0.961
	d	3/5	0.455	0.463	0.458	0.406	0.642	0.379	0.339	0.255	0.578
Correlational approach											
Not-Matched Effect δ_{nm}			–0.806	–0.639	–0.723	3.063	9.332	6.752	–4.293	–0.936	–3.150
Standard Error			2.761	2.098	2.429	2.931	3.878	4.38	2.209	2.855	2.934
DF			88.2	88.2	88.2	89.9	108	89.5	87.9	97.2	88.6
t Value			–0.29	–0.3	–0.3	1.05	2.41	1.54	–1.94	–0.33	–1.07
Probability			0.771	0.761	0.767	0.299	0.018	0.127	0.055	0.744	0.286
	t^2	1	0.084	0.090	0.090	1.103	5.791	2.372	3.764	0.107	1.153
	t^2 + DF	2	88.284	88.290	88.290	91.003	113.791	91.872	91.664	97.307	89.753
	rsquare	1/2	0.001	0.001	0.001	0.012	0.051	0.026	0.041	0.001	0.013
	r	sqroot	0.031 (–)	0.032 (–)	0.032 (–)	0.110	0.226	0.161	0.203 (–)	0.033 (–)	0.113 (–)

Table A.11.1 (continued)

	Steps in calculations		Fig. 11.1 SPI	Fig. 11.3 CI	Average of SPI & CI	Fig. 11.4 3rd Grade CI	Fig. 11.5 3rd Grade Read	Fig. 11.6 3rd Grade Math	Fig. 11.7 5th Grade CI	Fig. 11.8 5th Grade Reading	Fig. 11.9 5th Grade Math
	2r	3	0.062 (–)	0.064 (–)	0.064 (–)	0.220	0.451	0.321	0.405 (–)	0.066 (–)	0.227 (–)
	$1 - r^2$	4	0.999	0.999	0.999	0.988	0.949	0.974	0.959	0.999	0.987
	$\sqrt{(1 - r^2)}$	5	1.000	0.999	0.999	0.994	0.974	0.987	0.979	0.999	0.994
	d	3/5	0.062 (–)	0.064 (–)	0.064 (–)	0.221	0.463	0.326	0.414 (–)	0.067 (–)	0.228 (–)
Standardization approach											
Across All Four Time Periods:											
Ungrouped (Overall)	Std Dev		15.33	11.59	13.46	12.19	12.58	14.26	12.06	12.89	12.87
Co-nect	Std Dev		11.24	8.50	9.87	9.30	9.47	11.52	9.32	12.79	8.30
NotMatched	Std Dev		16.76	12.66	14.71	13.13	13.06	14.94	12.89	12.82	13.04
Matched	Std Dev		11.34	8.57	9.96	10.35	11.77	13.68	8.60	11.38	10.06
d = Target δ_t/Overall Std Dev			0.440	0.450	0.44	0.532	1.19	0.63	0.333	0.322	0.71
d = Not-Matched δ_{nm}/Overall Std Dev			–0.053	–0.055	–0.05	0.251	0.74	0.47	–0.356	–0.073	–0.24
d = Target δ_t/ Matched Std Dev			0.59	0.61	0.60	0.63	1.28	0.66	0.47	0.36	0.91
d = Not-Matched δ_{nm}/ Matched Std Dev			–0.07	–0.07	–0.07	0.30	0.79	0.49	–0.50	–0.08	–0.31
Random Effects	σ_s^2		32.7	18.6	25.2	14.3		23.7	26.3	NMLLRTEST χ^2 =38.54 DF = 4	37.4
	Pr *z*		$p = .007$	$p = .007$	$p = .007$	$p = .046$		$p = .0922$	$p = .004$	$p < .0001$	$p = .006$
	σ_b^2		32.9	19.0	25.4	37.7	68.0	84.8	20.9	0.4948	37.1
	Pr *z*		$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$
Covariance Structure			CS	CS	CS	CS	VC	CS	CS	CSH	CS

NMLLRTEST = Null Model Likelihood-Ratio Test.

Table A.11.2 Effect size calculations and random-effects estimates without using Kenward-Roger (KR) corrections for degrees of freedom

	Steps in calculations		Fig. 11. 1 SPI	Fig. 11.3 CI	Average of SPI & CI	Fig. 11.4 3rd Grade CI	Fig. 11.5 3rd Grade Reading	Fig. 11.6 3rd Grade Math	Fig. 11.7 5th Grade CI	Fig. 11.8 5th Grade Reading	Fig. 11.9 5th Grade Math
Correlational approach											
Tar get Effect δ_t			6.739	5.216	5.978	6.485	15.022	9.014	4.018	4.151	9.112
Standard Error			3.178	2.416	2.797	3.384	4.501	5.068	2.544	3.175	3.378
DF			27	27	27	27	27	27	27	27	27
t Value			2.12	2.16	2.14	1.92	3.34	1.78	1.58	1.31	2.698
Probability			0.043	0.04	0.04	0.066	0.002	0.086	0.125	0.213	0.012
	t^2	1	4.494	4.666	4.580	3.686	11.141	3.168	2.496	1.709	7.278
	t^2 + DF	2	31.494	31.666	31.580	30.686	38.141	30.168	29.496	28.709	34.278
	rsquare	1/2	0.143	0.147	0.145	0.120	0.292	0.105	0.085	0.060	0.212
	r	sqroot	0.378	0.384	0.381	0.347	0.540	0.324	0.291	0.244	0.461
	2r	3	0.756	0.768	0.762	0.693	1.081	0.648	0.582	0.488	0.922
	$1 - r^2$	4	0.857	0.853	0.855	0.880	0.708	0.895	0.915	0.940	0.788
	$\sqrt{}(1 - r^2)$	5	0.926	0.923	0.925	0.938	0.841	0.946	0.957	0.970	0.888
	d	3/5	0.816	0.831	0.824	0.739	1.285	0.685	0.608	0.503	1.038
Correlational approach											
Not-Matched Effect δ_{nm}			–0.806	–0.639	–0.723	3.063	9.332	6.752	–4.293	–0.9362	–3.1502
Standard Error			2.756	2.095	2.425	2.926	3.878	4.38	2.2047	2.7384	2.9338
DF			27	27	27	27	27	27	27	27	27
t Value			–0.29	–0.31	–0.3	1.05	2.41	1.54	–1.95	–0.34	–1.07
Probability			0.772	0.763	0.768	0.304	0.023	0.134	0.062	0.7437	0.2916
	t^2	1	0.084	0.096	0.090	1.103	5.791	2.372	3.803	0.117	1.153
	t^2 + DF	2	27.084	27.096	27.090	28.103	32.791	29.372	30.803	27.117	28.153
	rsquare	1/2	0.003	0.004	0.003	0.039	0.177	0.081	0.123	0.004	0.041
	r	sqroot	0.056 (–)	0.060 (–)	0.058 (–)	0.198	0.420	0.284	0.351 (–)	0.066 (–)	0.202 (–)

Table A.11.2 (continued)

	Steps in calculations		Fig. 11. 1 SPI	Fig. 11.3 CI	Average of SPI & CI	Fig. 11.4 3rd Grade CI	Fig. 11.5 3rd Grade Reading	Fig. 11.6 3rd Grade Math	Fig. 11.7 5th Grade CI	Fig. 11.8 5th Grade Reading	Fig. 11.9 5th Grade Math
	2r	3	0.111 (–)	0.119 (–)	0.115 (–)	0.396	0.840	0.568	0.703 (–)	0.131 (–)	0.405 (–)
	$1 - r^2$	4	0.997	0.996	0.997	0.961	0.823	0.919	0.877	0.996	0.959
	$\sqrt{(1 - r^2)}$	5	0.998	0.998	0.998	0.980	0.907	0.959	0.936	0.998	0.979
	d	3/5	0.112 (–)	0.119 (–)	0.115 (–)	0.404	0.926	0.593	0.751 (–)	0.132 (–)	0.413(–)
Standardization approach											
Across All Four Time Periods:											
Ungrouped (Overall)	Std Dev		15.33	11.59	13.46	12.19	12.58	14.26	12.06	12.89	12.87
Co-nect	Std Dev		11.24	8.50	9.87	9.30	9.47	11.52	9.32	12.79	8.30
NotMatched	Std Dev		16.76	12.66	14.71	13.13	13.06	14.94	12.89	12.82	13.04
Matched	Std Dev		11.34	8.57	9.96	10.35	11.77	13.68	8.60	11.38	10.06
d = Target δ_t/Overall Std Dev			0.440	0.450	0.44	0.532	1.19	0.63	0.333	0.322	0.71
d = Not-Matched δ_{nm}/Overall Std Dev			–0.053	–0.055	–0.05	0.251	0.74	0.47	–0.356	–0.073	–0.24
d = Target δ_t/Matched Std Dev			0.59	0.61	0.60	0.63	1.28	0.66	0.47	0.36	0.91
d = Not-Matched δ_{nm}/Matched Std Dev			–0.07	–0.07	–0.07	0.30	0.79	0.49	–0.50	–0.08	–0.31
Random Effects	σ_s^2		32.72	18.6	25.2	14.3		23.7	26.3	NMLLRTEST	37.4
										χ^2 =38.54	
										DF = 4	$p = .0061$
	Pr *z*		$p = .007$	$p = .007$	$p = .007$	$p = .046$		$p = .092$	$p = .004$	$p < .0001$	
	σ_b^2		32.86	18.98	25.4	37.7	68.0	84.8	20.9	0.4948	37.1
	Pr *z*		$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$	$p < .0001$
Covariance Structure			CS	CS	CS	CS	VC	CS	CS	CSH	CS

NMLLRTEST = Null Model Likelihood-Ratio Test.

Endnotes

[1] When selecting the matched schools Russell and Robinson (April 2000, 3) applied these criteria (quoted directly from their report):

a. Grades served had to be identical

b. If enrollment was less than 400, then it should be within 50 students, if enrollment was between 400 and 1000, it should be within 100 students, and if enrollment was above 1000, it should be within 200 students.

c. Differences in ethnic composition and free/reduced lunch should be within 10 percentage points.

d. Differences in LEP and mobility rate should be within 5 points.

The matched schools satisfied these criteria exactly or closely.

[2] A second chart was even more detailed; it separately reported the percentages passing each of the six tests during 1997 and 1999, for the target and its two matched schools, producing a total of 180 data elements (2 time periods times 3 target-matched schools times 6 tests times 5 target-matched comparisons).

[3] These measures of the schools' performance are referred to as *analytical* properties of the schools because they are derived from mathematical operations on the data describing their individual students (Lazarsfeld and Menzel [1961] 1972, 227). The schools are characterized by an average based on the performance on tests of its own students.

[4] By having two confounded measures of school performance and reasonably strong effects of the reform program, the probability of at least one response showing a positive significant effect is higher than if only one measure is used. Let the null hypothesis be H_0: The school reforms enhance school performance. Let the alternative hypothesis be H_a: The school reforms do not enhance school performance. Thus, H_0 is more likely to be accepted (and H_a rejected) even if H_0 is false and H_a is true. This is a Type II error. By combining the two highly correlated treatment effects and testing the composite for significance, this borrowing of strength will minimize the risk of the Type II error of accepting a false hypothesis.

[5] At the time of this study additional information about these measures was available at the School Improvement in Maryland Web site http://www.mdk12.org–and at the Maryland State Department of Education Web site http://www.msde.state.md.us.

[6] In order to accommodate the reversals due to the use of the Class statement, post is referred to in the actual SAS code as ryear20 and its zero (0) value indicates the post time period and its one (1) value indicates the baseline time period.

[7] The DDFM = KR specifies the Denominator Degrees of Freedom Method as Kenward–Roger (KR). This option corrects for unbalanced data, multiple random effects, and correlated errors. SAS (Littell et al. 2006, 188) recommends its use in repeated measures models in part because it reduces Type I errors, the probability of rejecting true hypotheses.

[8] Littell et al. (2006, 200) clarify the difference between R and V options as follows:

> The V option on the RANDOM statement request that the marginal covariance matrix be displayed. The difference between the result of the R option in the REPEATED and the result of the V option in the RANDOM statement is simply that the latter displays Var[Y] = ZGZ' + R, whereas the former displays Var[Y|u] = R.

[9] The following SAS code implements the Covtest statements in Proc Glimmix for the normally distributed response variable, the SPI index. For explication of these tests for logistic regression, see the next chapter or SAS online documentation.

```
Title 'Covariance Tests in Proc Glimmix for Harford Data';
Proc GLIMMIX data = standard ic = Q;
   class school Trt Post;
   model msapperform = Trt Post Trt*Post year98 year99 lowstgrd locate
   targett1 safe stratio femprob Blackp forredlp
```

```
    /solution dist = normal link = identity;
    random _residual_/type = unr sub = school(trt) s v vcorr;
    weight impscore;
    covtest 'Ho: No G-Side Random Effects (UNR Parameters)' ZeroG/cl;
    covtest 'Ho: Independence = No G-side, Diagonal R-side ' Indep/cl;
    covtest 'Ho: Homogeneous off-diagonal correlations, CSH ' General
      0 0 0 0 1 - 1,
      0 0 0 0 1  0 -1,
      0 0 0 0 1  0  0 -1,
      0 0 0 0 1  0  0  0 -1,
      0 0 0 0 1  0  0  0  0 -1/estimates;
    covtest 'Ho: compound symmetry' General
      1 -1,
      1  0 -1,
      1  0  0 -1,
      0  0  0  0 1 -1,
      0  0  0  0 1  0 -1,
      0  0  0  0 1  0  0 -1,
      0  0  0  0 1  0  0  0 -1,
      0  0  0  0 1  0  0  0  0 -1/estimates;
run;
```

[10] However, a likelihood-ratio test indicates that there is little difference between these models. The difference in -2 RLL produces a χ^2 of 2.38 (747.68 – 745.30), the difference in *df* = 3, and the ***p*** = .50. The null hypothesis of no difference is not rejected.

[11] Karney and Bradbury (1997) have applied Rosenthal's formulas to calculate effect sizes based on parameters from multilevel models.

[12] Benchmarks for effect sizes in educational research define a *d* =.20 as small and a *d* =.80 as large. Modest effects ranging from *d* =.10 to *d* =.20 should not be ignored. For further discussion see Borman et al. (2003).

[13] For each response variable the bottom four rows of the Appendix Tables 11.1 and 11.2 report the covariance parameter estimates for the full models and their statistical significance. Apparently, the KR corrections had no discernable effects on the statistical significance of these covariance parameters.

[14] The model composed of all of the predictors except the treatment effect coefficients also exhibits this pattern. But its BIC of 770.5 is higher than the BIC of 755.4 for the full model that includes the treatment effect coefficients—the latter model fits better.

[15] The models that use the Repeated statement do not readily produce the random-effects estimates for each school. This Random statement produced the estimates for Fig. 11.2 and for the other models mentioned here:

Random Intercept/Subject = School (Trt) Type = VC Solution V VCORR;

[16] Both composite effects have been standardized by dividing the treatment effects by the simple average of the overall ungrouped standard deviations of the SPI and CI. The SD of the SPI = 15.33 and that of the CI = 11.59; their simple average is 13.46.

[17] Because all of the effects of the reform treatment are positive and all of the effects of the null treatment of the not-matched group are negative, it is not reasonable to correct the set of six probabilities for multiplicity. If negative signs are attached to the probabilities for the negative effects, then Proc Multtest treats these probabilities as missing and only adjusts the probabilities for the three positive effects.

Chapter 12
Using Propensity Scores

> *Comparing cancer death rates from 1950 to the present… is like visiting base camp at Mount Everest before and after an ascent and concluding that nothing has changed. It ignores the remarkable feat accomplished in the meantime.*
>
> —John R. Seffrin, American Cancer Society (2009)

This chapter examines how both Co-nect's comprehensive school reforms and reduced student-to-teacher ratios engendered change in the schools' proportions of students passing tests of reading, mathematics, and fourth grade writing. Using propensity scores that aim to minimize any bias due to the selection of the target or comparison schools, a difference-in-differences (DID) design compares the change on test outcomes in seven elementary schools that participated in the reforms (the target schools) in the time periods before and after the implementation of the reforms, to the change on those outcomes in six comparable elementary schools (the comparison schools) during those same time periods. Several of the response variables are measured on changing cohorts (i.e., quasi-cohorts) of third through fifth graders in Houston, Texas. Even though the mobility of the students into and out of the schools no doubt changed the original composition of the cohorts, henceforth this chapter refers to these quasi-cohorts simply as cohorts. The tests are components of the Texas Assessment of Academic Skills, hereafter referred to as TAAS. In the spring of each school year (SY) the teachers administered these tests in the English language to all capable students. Although this study probes the effects of the reforms after 1 year, from third to fourth grade, it also studies change from fourth to fifth grade and overall change from third to fifth grade; it often refers to these periods as Time 0, Time 1, and Time 2. Relative to change in the comparison schools, the cohorts of students in target schools generally improved their percent passing from third to fourth grade, but by fifth grade the cohorts in both groups of schools were performing about equally, near the ceiling of 100% passing. This chapter attributes this favorable change in the comparison cohorts to the extra teachers assigned to previously low-performing schools to prepare their students for the fifth grade tests.[1]

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_12, © Springer Science+Business Media B.V. 2011

New Contributions

In addition to this chapter's substantive contributions to educational policy studies via its testing of the Co-nect reform model and its probing of consequences of changed student-to-teacher ratios, it also develops a methodological paradigm for evaluative research that combines propensity scores and generalized multilevel mixed models. To circumvent the statistical problems of separation and non-convergences of estimates, it applies Firth's penalized logistic regression to "Goldbergerized" data, thereby enabling the derivation of propensity scores using logistic regression rather than ordinary least-squares. It derives DID estimators of average treatment effects, relates these to the potential outcomes causal perspective, and estimates the parameters using SAS's Proc Glimmix to model the repeated measures. It validates treatments by discerning their impacts on the standard deviations (SDs) of the response variables; strong treatments induce small, homogeneous values of SDs. It distinguishes *preserving* DID effects from *generating* DID effects: the former combine a modest favorable difference in the treatment group with a substantial unfavorable difference in the comparison group; the latter combine a substantial favorable difference in one group with a modest or negative difference in the other. By sequentially applying SAS's new Covtest statements, it develops and illustrates a decision-logic for the choice of covariance structures in multilevel models based on pseudo-likelihood. It quantifies average treatment effects in terms of odds ratios, proportion differences, and standardized proportion differences (i.e., effect sizes); depicts the key findings graphically; corrects the *p*-values for multiplicity; and calculates the composite effects of the treatments. The SAS code and data sets are available for further study and replications.

The Setting of the Study

Forming feeder systems, the Houston Independent School District (HISD) is composed of a number of subdistricts that include high schools, middle schools, and elementary schools. The elementary schools in this study are components of the feeder system of the John H. Reagan High School of the Central Administrative District.[2] The graduates of these elementary schools usually attend the Hamilton or Hogg Middle Schools, and the graduates of these middle schools usually attend the Reagan High School. Because of reporting irregularities about dropouts, the Texas Education Agency (TEA) downgraded its evaluation of Reagan from Recognized to Low Performing and it also downgraded B. T. Washington, another of the four high schools in this subdistrict.[3] The TEA did not downgrade their ratings of Hamilton and Hogg middle schools, or of any of the elementary schools in this study. Apparently, the aggregated data analyzed here met the standards of the TEA.

Aspects of the Schools

Negating the view that changed distributions of social attributes of a school are responsible for the changes in student achievement, Table 12.1 documents the stability of various correlates of the target and comparison schools for the period of this study. The target schools have a higher proportion of Hispanic students (0.87 compared

Table 12.1 Characteristics of target and comparison elementary schools in the Reagan High School feeder system, Houston Texas, statistically unadjusted rates

Time period	Time 0	Time 1	Time 2	Pooled
School year of test	1999–2000	2000–2001	2001–2002	1999–2002
Ethnicity				
Proportion Hispanic[a]	0.82	0.82	0.82	0.82
Target	0.87	0.87	0.87	0.87
Comparison	0.76	0.76	0.77	0.76
Proportion African-American	0.10	0.10	0.11	0.10
Target	0.03	0.03	0.04	0.03
Comparison	0.19	0.19	0.19	0.19
Economic status				
Free or reduced lunch[a]	0.91	0.91	0.88	0.90
Target	0.88	0.88	0.85	0.87
Comparison	0.94	0.94	0.90	0.93
Title-1 Schools:	0.92	0.92	0.92	0.92
Target	0.86	0.86	0.86	0.86
Comparison	1.00	1.00	1.00	1.00
Preparation for school				
Limited english proficiency	0.47	0.48	0.48	0.48
Target	0.49	0.49	0.50	0.49
Comparison	0.46	0.47	0.46	0.47
Mobility rate[a]	0.18	0.19	0.20	0.19
Target	0.17	0.18	0.20	0.18
Comparison	0.20	0.21	0.20	0.20
School characteristics				
Sixth grade highest grade[a]	0.31	0.23	0.31	0.28
Target	0.14	0.14	0.14	0.14
Comparison	0.50	0.33	0.50	0.44
Success for all[a]	0.46	0.46	0.46	0.46
Target	0.57	0.57	0.57	0.57
Comparison	0.33	0.33	0.33	0.33
Students per teacher[a]	17.0	17.8	15.4	16.7
Target	18.0	18.5	16.0	17.5
Comparison	15.8	17.1	14.7	15.9
School rating	2.4	2.7	3.5	2.9
Target	2.4	2.9	3.7	3.0
Comparison	2.3	2.5	3.3	2.7

[a] Used as control variables. The use of the other items as controls would create problems of collinearity or endogeneity and for that reason they are not used as control variables.

with 0.76) and the comparison schools have a higher proportion of African-American students (0.19 compared with 0.03); at least 90% in each school are minority students. Only the Burrus elementary school is predominantly African American. Economic disadvantage no doubt affects the students' preparation for schooling; on this dimension the target and comparison schools are similar. The comparison schools have slightly higher proportions of economically disadvantaged students: about 0.93 qualified for free or reduced-price lunches compared with 0.86 in the target schools. Not surprisingly, six of the seven target schools and all of the comparison schools are eligible for Title-1 aid. In the target schools 0.49 of the students are classified as Limited English Proficient (LEP) compared to 0.47 in the comparison schools. The mobility rate—the proportion of students transferring to another school during the school year—is 0.18 at the target schools and 0.20 at the comparison schools. In the target schools, fifth grade (rather than sixth) is more likely to be the highest grade.

The highly regarded Success for All (SFA) comprehensive reform program emphasizes developing the students' skills in reading and other core subjects (Borman and Hewes 2002, 245–247).[4] The target schools are slightly more likely to implement SFA than the comparison schools, the proportion difference is 0.24, $p = 0.14$. With the aim of controlling for the effects of SFA, the multilevel models of this chapter treat the presence (1) or absence (0) of SFA in a school as a binary mean-centered fixed covariate.[5]

Smaller class sizes indicated by reductions in the student-to-teacher ratios may ameliorate some of the disparities in educational opportunity of these minority students.[6] As Table 12.1 reports for the period of this study, the ratio of students to teachers is less favorable in the target schools: 17.5 to 1 compared with 15.9 to 1. In both the target and comparison schools these ratios increased from SY 1999–2000 to SY 2000–2001 and declined from SY 2000–2001 to SY 2001–2002: declining from 18.5 to 16 (−2.5) in the target schools and from 17.1 to 14.7 (−2.4) in the comparison schools, the lowest value.

These extra teachers may be responsible for a consistent pattern in these data: namely, in the comparison schools from third to fourth grade, when the ratio of students to teachers became less favorable, the students' test scores declined, whereas from fourth to fifth grade when there were extra teachers, these students generally improved their percentages passing the reading and mathematics tests and the average of these tests. The students in the Co-nect schools exhibited steadier improvement even though their student-to-teacher ratios were less favorable. For the achievement of minority students from low-income families, this pattern suggests that either comprehensive school reforms in the Co-nect manner or the provision of more teachers in underachieving schools may have similar beneficial consequences; this notion is a major interpretive theme of this chapter.[7]

Each school year the Texas Education Agency (TEA) rates each school according to their performance on their standardized tests and on their dropout rate. The categories are exemplary (coded as 4), recognized (coded as 3), academically acceptable (coded as 2), and academically unacceptable (coded as 1). In SY

1998–1999 all of the target schools were rated as acceptable whereas two of the comparison schools were rated as recognized and four as acceptable. In SY 1999–2000 the unadjusted rates indicate very little difference between these two groups of schools. Because these ratings increase monotonically without reversal in both groups of schools, whereas the performance of the comparison schools declined and then recovered, these ratings seem not to track changes in effectiveness and school quality. Consequently, this evaluation uses these ratings sparingly; it studies instead how the test score outcomes are affected by the reforms.

Aspects of the Reforms

The comprehensive school reforms aimed to improve the teachers' manner of instruction, thereby improving the students' achievement—especially that of students with limited proficiency in English. Facilitating these improvements, the Co-nect consultants provided the target schools with consultation, training, professional development, and new instructional strategies; the latter focused on project-based learning, standards-based curriculum alignment, and technology integration. The Co-nect change agents collaborated with school facilitators who were employed by the school district to provide continuous onsite support for the process of change; early on, these school facilitators received six days of intensive training.

The first year of the study (SY 1999–2000 or Time 0) began in the target schools toward the end of the school year; it is the primary baseline period of this study. The consultants introduced the teachers to the available services and to the Exchange, an Internet-based resource, and, to inform the teachers about project-based learning, they conducted workshops that focused on improving the quality of their projects. The teachers also participated in workshops on rubrics and classroom assessment. Within each school the consultants formed design teams to foster school leadership and support for continuous improvement.

During the second year of the study (SY 2000–2001 or Time 1) the consultants focused on improving the teachers' use of project-based learning, new assessment strategies, and technology in teaching; "mini-sabbaticals" taught methods for integrating technology into teaching and learning. Aiming to improve the assessment of students in project-based learning, the consultants also presented workshops on the use of rubrics.

During the third year of the study (SY 2001–2002 or Time 2) the consultants encouraged each school to develop an individualized implementation plan by selecting its own combination of workshops. Most of the schools remained highly focused on project-based learning and technology integration, while also working on grade-level planning and assessment. Milam Elementary School, for example, focused on planning curriculum alignments and data analysis in mathematics; it had the best benchmarking score of these Co-nect schools.[8]

Possible Impacts of the Reforms

For the target and comparison schools, Table 12.2 reports the statistically unadjusted trends on the outcome variables. On reading tests, the target schools always outperformed the comparison schools; the latter schools started out slightly lower, then lost ground, and then recovered. On mathematics tests, both groups improved with the target schools experiencing their improvement earlier than the comparison schools. On the average of mathematics and reading tests, the target schools generally had higher scores; the scores for the comparison schools dipped and then recovered. On the fourth grade writing tests, the comparison schools initially had higher test scores but then declined, whereas the target schools improved. On average, the target schools had higher test scores than the comparison schools on all four of these outcome measures.

Evaluation studies often report that the standard deviations (SDs) of the means on the outcomes are smaller in treatment groups than in comparison groups; robust treatments may create more homogeneous responses than null treatments (Rosenthal 1991, 16). If the Co-nect reform treatments are robust, then the SDs of the means on the outcomes in the treatment schools should be smaller and more stable than those in the comparison schools, and they usually are: Table 12.3 reports

Table 12.2 For target and comparison groups of schools, the change across time in measures of outcomes; statistically unadjusted means

Time Period	Time 0	Time 1	Time 2	Pooled
School year of test:	1999–2000	2000–2001	2001–2002	1999–2002
Proportion passing tests				
Reading test				
Both groups	0.903	0.881	0.932	0.905
Target	0.916	0.950	0.956	0.940
Comparison	0.888	0.800	0.905	0.864
Mathematics test				
Both groups	0.782	0.868	0.966	0.872
Target	0.794	0.957	0.953	0.901
Comparison	0.767	0.763	0.982	0.837
Average of reading and mathematics tests				
Both groups	0.843	0.875	0.948	0.889
Target	0.856	0.956	0.953	0.921
Comparison	0.828	0.781	0.942	0.851
Fourth grade writing test				
Both groups	0.928	0.886	0.925	0.913
Target	0.911	0.899	0.963	0.924
Comparison	0.947	0.872	0.880	0.899
***				
School quality ratings by TEA				
Both groups	2.38	2.69	3.54	2.87
Target	2.43	2.86	3.71	3.00
Comparison	2.33	2.50	3.33	2.72

Table 12.3 For target and comparison schools, the change across time in the standard deviations of the measures of outcomes, statistically unadjusted

Time period	Time 0	Time 1	Time 2	Pooled
School year of test:	1999–2000	2000–2001	2001–2002	1999–2002
Standard deviations (SDs)				
Reading test				
Both groups	0.069	0.131	0.056	0.092
Target	0.066	0.061	0.049	0.059
Comparison	0.075	0.149	0.055	0.106
Mathematics test				
Both groups	0.136	0.174	0.052	0.149
Target	0.057	0.054	0.066	0.096
Comparison	0.200	0.212	0.029	0.190
Average of reading and mathematics tests				
Both groups	0.095	0.145	0.047	0.110
Target	0.055	0.056	0.057	0.071
Comparison	0.132	0.165	0.036	0.135
Fourth grade writing test				
Both groups	0.099	0.082	0.087	0.089
Target	0.113	0.086	0.040	0.086
Comparison	0.086	0.083	0.109	0.094
***				
School quality ratings by TEA				
Both groups	0.506	0.630	0.660	0.767
Target	0.535	0.378	0.488	0.707
Comparison	0.516	0.837	0.817	0.826

the SDs of the unadjusted means reported above in Table 12.2; these SDs are generally smaller and more stable in the target schools than in the comparison schools. But note: because of the extra teachers assigned to low performing schools to prepare their fifth graders for the fifth grade achievement tests, in the comparison schools the student-to-teacher ratio declines from 17.1 at Time 1 to 14.7 at Time 2; then the SDs drop to 0.055 for reading, 0.029 for mathematics, and 0.036 for the average of the reading and mathematics tests. There is no drop in the SDs of the fourth grade writing test due to the extra teachers; in fact the SDs increase from 0.083 to 0.109. Apparently, in the comparison schools the extra teachers have an effect on the SDs of the fifth grade reading and mathematics tests but not on the SD of the fourth grade writing test.

Conjectures

What conjectures may explain these patterns? This chapter suggests that from third to fourth grade, the target schools may have benefited from project-based learning and other key components of the Co-nect design for comprehensive school reform.

Because the comparison schools were not exposed to these reforms, their performance dropped off. From fourth to fifth grade, the reforms either maintained or enhanced the target schools' performance, whereas the comparison schools' performance improved during this period. Why this improvement? It may have been due to the additional school personnel the District assigned to the low performing schools. These additional teachers focused on ameliorating the students' weaknesses and on preparing them for the fifth grade tests. Additional staff in the fifth grade of the comparison schools improved their student-to-teacher ratios to the lowest value; as a consequence the students' test scores in fifth grade may have improved.

The first alternative conjecture posits that these patterns represent random fluctuation. Testing this conjecture, the subsequent sections present the statistically adjusted results. If correctly specified models reject the random fluctuation explanation—that is, at least some of the treatment effects are statistically significant—then the above conjectures about the favorable effects of the reforms and the additional teachers in the comparison schools would garner some support.

The second alternative conjecture posits that any favorable effects of the Co-nect reforms are confounded with the favorable effects of Success for All (SFA), even though the statistical models control for this factor. Further assessing this alternative explanation, a separate analysis examines four types of schools: those exposed to Co-nect reforms and SFA, only the former, only the latter, and neither. If the schools exposed only to the Co-nect reforms perform better than the schools exposed only to SFA, then any favorable effects of the reforms would not be due to the confounding of the effects of these two reform programs, and this is the case empirically.

Methods

The study design articulates DID, propensity scores, repeated measures, and covariance structures; these components restrict the choice of which statistical models to apply. SAS's Proc Glimmix (Generalized Linear Mixed Models) provides solutions to these constraints. It models the test outcomes by estimating the average treatment effects via multilevel logistic regression models that incorporate appropriate covariance structures. For cohorts of third grade students it implements a DID design that compares the change in test outcomes in seven target schools with the change in six comparison schools. It uses the students' aggregated test scores for the spring term of SY 1999–2000 as the primary baseline for the comparisons and tracks these cohorts using their aggregated test scores for fourth grade (SY 2000–2001) and fifth grade (SY 2001–2002). For each of the schools in each school year, SAS treats the outcomes for reading, mathematics, their average, and fourth grade writing as binomial variables based on the proportions passing the tests. The target and comparison schools group these repeated measures.

The Difference-In-Differences (DID) Design

The six cells of the DID design are produced by the cross-tabulation of the dichotomous treatment group difference, target (1) versus comparison (0) schools, with a classification typology (i.e., a Class variable) that codes the time periods as baseline = Time 0 (i.e., third grade or SY 1999–2000), middle = Time 1 (i.e., fourth grade or SY 2000–2001), and final = Time 2 (i.e., fifth grade or SY 2001–2002).[9] SAS creates binary (0,1) indicator variables from this classification using the category with the highest coded value as the omitted category. It estimates a treatment effect by taking the difference between these differences: the difference in means for the target schools and the difference in means for the comparison schools; and, equivalently, by estimating the coefficient on an interaction effect, which is the coefficient on the product of the treatment group and time period indicator variables. Because of the nuances of the indicator variable coding, the DID estimates are easier to interpret than the interaction effects.

DID estimators

This explication assumes that the means of a response variable are available for each of the six cells of the design; SAS's least-squares means and estimate statements can obtain these means and their differences, and test them for statistical significance. These means can be viewed conceptually as a result of the centering of the covariates by their overall sample means, thereby incorporating their effects into the intercept.[10] The variables that form the 2 × 3 design are not mean-centered and their effects are not included in the intercept. Thus, for each cell of design, the mean μ of the response variable depends on the sum of the intercept and the effects of the categories of the design variables coded 1 rather than 0 that operate on the units in the cell.

Table 12.4 presents illustrative derivations of the relevant DID estimators based on this equation:

$$\mu_{\mathrm{trttime}} = \beta_0 + \beta_1 \mathrm{Trt} + \tau_1 \mathrm{Time}_{01} + \delta_{11}(\mathrm{Trt} \times \mathrm{Time}_{01}) + \tau_2 \mathrm{Time}_{02} + \delta_{12}(\mathrm{Trt} \times \mathrm{Time}_{02}) \quad (1)$$

It states that the mean μ_{ji} ($i = 0, 1, 2$ time periods; $j = 0{,}1$ treatments) of a cell of the design depends upon the intercept β_0 plus the effects of different combinations of the coefficients on these indicator variables when they are coded 1 rather than 0: the effect of β_1 due to the difference between target (1) and comparison groups (0), plus τ_1, the effect of time from the baseline (0) to the middle time period (1), plus the effect δ_{11} of the treatment × time_{01} interaction (coded 0 or 1), plus τ_2 the effect of time from the baseline (0) to the final time period (2) (coded 0 or 1), and plus the effect δ_{12} of the treatment × time_{02} interaction (coded 0 or 1). Because all of the fixed covariates are centered by their overall sample means, the intercept β_0 captures their effects. It represents the baseline mean amount of the response

Table 12.4 Derivations of the Difference-in-Differences (DID) estimators

$\mu_{\text{trttime}} = \beta_0 + \beta_1\text{Trt} + \tau_1\text{Time}_{01} + \delta_{11}(\text{TrtTime}_{01}) + \tau_2\text{Time}_{02} + \delta_{12}(\text{TrtTime}_{02})$

Treatment (Trt):	Comparison group = 0	Target group = 1	Target – Comparison group
DID estimators, Time 0 to Time 1			
Time 0 (Baseline)	$\mu_{00} = \beta_0$	$\mu_{10} = \beta_0 + \beta_1$	β_1
Time 1 (Middle)	$\mu_{01} = \beta_0 + \tau_1$	$\mu_{11} = \beta_0 + \beta_1 + \tau_1 + \delta_{11}$	$\beta_1 + \delta_{11}$
Difference Time 1 – Time 0	$d(0)_m = \mu_{01} - \mu_{00}$ $= \tau_1$	$d(1)_m = \mu_{11} - \mu_{10}$ $= \tau_1 + \delta_{11}$	$\hat{\delta}_{\text{mid}} = \text{d}(1)_m - \text{d}(0)_m$ $= \delta_{11}$
DID estimators, Time 0 to Time 2			
Time 0 (Baseline)	$\mu_{00} = \beta_0$	$\mu_{10} = \beta_0 + \beta_1$	β_1
Time 2 (Final)	$\mu_{02} = \beta_0 + \tau_2$	$\mu_{12} = \beta_0 + \beta_1 + \tau_2 + \delta_{12}$	$\beta_1 + \delta_{12}$
Difference Time 2 – Time 0	$d(0)_f = \mu_{02} - \mu_{00}$ $= \tau_2$	$d(1)_f = \mu_{12} - \mu_{10}$ $= \tau_2 + \delta_{12}$	$\hat{\delta}_{\text{fin}} = d(1)_f - d(0)_f$ $= \delta_{12}$
DID estimators, Time 1 to Time 2			
Time 1	$\mu_{01} = \beta_{0\,+}\tau_1$	$\mu_{11} = \beta_0 + \beta_1 + \tau_1 + \delta_{11}$	$\beta_1 + \delta_{11}$
Time 2	$\mu_{02} = \beta_0 + \tau_2$	$\mu_{12} = \beta_0 + \beta_1 + \tau_2 + \delta_{12}$	$\beta_1 + \delta_{12}$
Difference, Time 2 – Time 1	$d(0)_{fm} = \mu_{02} - \mu_{01}$ $= \tau_2 - \tau_1$	$d(1)_{fm} = \mu_{12} - \mu_{11}$ $= \tau_2 - \tau_1 + \delta_{12} - \delta_{11}$	$\hat{\delta}_{\text{finmid}} = \delta_{12} - \delta_{11}$

This table assumes that the lowest coded value of a Class variable is the reference category for the binary (0,1) indicator variables. However, SAS uses the highest coded value of a Class variable as the reference category and this complicates the interpretation of these effects.

variable when none of the design variables are operating; it appears in each cell of the design.

Derivation of DID estimators

From this setup the DID estimators $\hat{\delta}_{\text{middle}} = \delta_{11}$ and $\hat{\delta}_{\text{final}} = \delta_{12}$ can be derived as follows:[11] In the cell for comparison schools (coded 0) at Time 0 the mean test score μ_{00} equals the intercept, β_0. In the cell for comparison schools at Time 1 the mean test score μ_{01} equals the intercept $\beta_0 + \tau_1$, the latter is the coefficient on the indicator variable Time_{01} that gauges the effect of time from Time 0 to Time 1. Let the difference between these means in the comparison group be $d(0)_m = \mu_{01} - \mu_{00} = (\beta_0 + \tau_1) - (\beta_0) = \tau_1$, which is the effect of time.

Similarly, in the cell for target schools (coded 1) at Time 0 the mean test score $\mu_{10} = \beta_0 + \beta_1$, the sum of the intercept plus the coefficient on the indicator variable Trt that compares the target schools (1) with the comparison schools (0). In the cell for target schools at Time 1 the mean test score $\mu_{11} = (\beta_0 + \beta_1) + (\tau_1 + \delta_{11)}$; τ_1 is the coefficient on Time_{01}, and δ_{11} is the coefficient on the interaction Trt × Time_{01}.

Let $d(1)_{\mathrm{m}}$ equal the difference between these target group means: $d(1)_{\mathrm{m}} = \mu_{11} - \mu_{10} = (\beta_0 + \beta_1) + (\tau_1 + \delta_{11}) - (\beta_0 + \beta_1) = (\tau_1 + \delta_{11})$. Then the DID estimator $\hat{\delta}_{\mathrm{middle}} = d(1)_{\mathrm{m}} - d(0)_{\mathrm{m}} = (\tau_1 + \delta_{11}) - (\tau_1) = \delta_{11} = \mathrm{Trt} \times \mathrm{Time}_{01}$; this interaction quantifies the average effect of the reform treatment from Time 0 through Time 1.

The average treatment effect for the final period (Time 2) compared with the initial period (Time 0) is calculated analogously by taking the difference between these two differences: $d(0)_{\mathrm{f}} = \mu_{02} - \mu_{00} = (\beta_0 + \tau_2) - (\beta_0) = \tau_2$; and $d(1)_{\mathrm{f}} = \mu_{12} - \mu_{10} = (\beta_0 + \beta_1) + (\tau_2 + \delta_{12}) - (\beta_0 + \beta_1) = (\tau_2 + \delta_{12})$. Then $\hat{\delta}_{\mathrm{final}} = d(1)_{\mathrm{f}} - d(0)_{\mathrm{f}} = (\tau_2 + \delta_{12}) - (\tau_2) = \delta_{12} = \mathrm{Trt} \times \mathrm{Time}_{02}$; this interaction quantifies the average effect of the reform treatment from Time 0 through Time 2.

Thus far, this section has derived estimators for the average treatment effect from Time 0 to Time 2 and the average treatment effect from Time 0 to Time 1. Logically, the average treatment effect from Time 1 to Time 2 should equal the difference between these estimators; that is, $\hat{\delta}_{\mathrm{finmid}} = \hat{\delta}_{\mathrm{final}} - \hat{\delta}_{\mathrm{mid}} = \delta_{12} - \delta_{11} = \mathrm{Trt} \times \mathrm{Time}_{02} - \mathrm{Trt} \times \mathrm{Time}_{01}$. The third derivation of Table 12.3 confirms this intuition.

These estimators of average treatment effects are consistent with the potential outcomes causal perspective: Prior to assignment to a target or comparison group, a school has potential outcomes under each treatment, reform or null. After assignment to one of these treatments, the school has realized outcomes under that treatment and counterfactual outcomes under the other treatment. For that school the unit-level causal effect of the treatments cannot be calculated because data are missing about the outcomes under the treatment that the school did not receive. However, given closely matched schools in each treatment condition, an estimate of the average causal effect can be calculated as the difference between the differences in means, as was done above. This model assumes that if the target schools had instead received the null treatment, then their pattern of outcomes would reproduce that of the comparison schools; and if those comparison schools had instead received the reform treatment, then their pattern of outcomes would reproduce that of the target schools.

Even though this DID design assesses change on a unit before and after a treatment, thereby using the unit as its own control, because of the absence of random assignment there may be systematic differences between the target and comparison schools that affect the validity of the estimates. With the aim of reducing such biases this chapter applies propensity scores.

Propensity Scores

Although the target and comparison schools are in the same feeder system, are physically proximate, and have very similar if not identical characteristics, there are some relevant differences: The faculty of a target schools voted to have their school participate in the reforms, and at least 70% supported participation. Because the comparison schools were not offered these reforms and did not directly receive any of its potential benefits, a biased selection of the target schools could affect the results. There could be a confounding of the selection of the target schools with the eventual outcomes in student achievement. Although the target and comparison schools had similar socioeconomic and ethnic compositions, there are some noticeable differences (see Table 12.1). Moreover, the target schools exhibited higher initial test scores on reading, mathematics, and the average of reading and mathematics; the comparison schools exhibited higher initial test scores on the fourth grade writing test (see Table 12.2). A ceiling effect could limit the improvement of these schools with the higher initial values.

Addressing these possible biases, this chapter creates propensity scores and uses them as a mean-centered covariate in the multilevel modeling. A control for robust propensity scores may reduce any potentially spurious effects on the outcomes due to the variables in the assignment equation that created the propensity scores. Here, a school's propensity score is its predicted probability of its being a target school (Rosenbaum and Rubin 1985, 34–35). Ideally, the predicted values are calculated from statistical models that are most appropriate for dichotomous response variables, namely logistic regression or probit models. Initially, for the data set of about 200 Houston elementary schools, because of the statistical problem of separation, the estimates from these models would not converge; perhaps there were too few target elementary schools—seven—compared with the 180 nontarget schools with complete information. Circumventing this problem, a "Goldbergerized" ordinary least-squares (OLS) regression analysis obtained the initial probabilities that a school was a target school. Then, given the Goldbergerized data, a Firth-corrected logistic regression obtained the probabilities used in the analysis, which are very similar to the OLS estimates.[12]

Goldberger and Firth estimates

Achen (1986, 40–41) outlines the first three steps of Goldberger's procedure; this chapter adds the fourth step, the Firth-corrected logistic regression:

1. Apply ordinary regression with dichotomous treatment as the dependent variable.
2. For each observation, let the forecast value be p and adjust forecasts above 0.99 to 0.99 and those below 0.01 to 0.01. Set $q = 1 - p$ and $s = \sqrt{pq}$.
3. Divide each variable in the regression by s. The intercept term becomes $1/s$. Apply ordinary regression to this new set of variables, including the variable $1/s$ but suppressing the conventional intercept. The resulting coefficients and standard errors are the linear probability estimates.
4. Apply logistic or probit regression to the Goldbergerized variables. If the estimates do not converge, then use Firth's method of penalized maximum likelihood to obtain the propensity scores, coefficients, and standard errors.

Surprisingly, when applied to the unmodified data the Firth (1993) model did not provide appropriate estimates.

Predictors of treatment group membership

The assignment equation predicting membership in the treatment group includes these 16 prior characteristics of a Houston elementary school ($R^2 = 0.33$, adjusted $R^2 = 0.27$): Project Grad feeder system;[13] total number of students; number of students with limited English propensity; proportion of white students; count of migrant students; proportion of male students; mobility rate; location (i.e., inner city, outer fringe, etc.); count of Native Americans; proportion of students eligible for free or reduced-price lunches; lowest grade (i.e., kindergarten, first grade, etc.); count of Asians or Islanders in the school; and dummy variables for these SY 1998–1999 school ratings; Exemplary, Recognized, Unacceptable, an Unusable (Acceptable is the omitted category).

The OLS and Firth logistic regressions using the Goldbergerized data produce nearly identical estimates of the propensity scores (r = 0.999); see Table 12.5. The seven target schools have a slightly higher average probability (0.33) of being in the target group than do the comparison schools (0.29); their propensity scores also have a narrower range (0.31 to 0.36) than the comparison schools (0.22 to 0.38). No schools exhibit extremely large or extremely small probabilities that would limit their usefulness as a mean-centered control in the logistic regressions.[14] The analysis uses the Firth scores.

Table 12.5 Propensity scores derived from "Goldbergerized" data, estimates from ordinary least-squares (OLS) and Firth logistic regressions

Target	Estimates:		Comparison	Estimates:	
Schools:	OLS	Firth	Schools:	OLS	Firth
Milam	0.36	0.36	Eighth Ave.	0.38	0.38
Helms	0.35	0.36	Memorial	0.36	0.36
Travis	0.34	0.34	Burrus[a]	0.25	0.27
Love	0.33	0.33	Brock	0.24	0.26
Stevenson	0.31	0.32	Field	0.24	0.26
Browning	0.31	0.32	Harvard	0.19	0.22
Crockett	0.30	0.31			
Average	0.33	0.33		0.28	0.29

[a] Burrus is mostly African-American; the other schools, mostly Hispanic. The two measures are correlated . 999; the analyses use the Firth estimates. The 16 predictors of membership in the treatment group are the following: Project Grad feeder system, total students, limited English proficient, white proportion, count of migrants, male proportion, mobility rate, locale, count of native Americans, free or reduced-price lunch proportion, lowest grade of school, count of Asians and islanders, prior school rating indicators: exemplary, recognized, unacceptable, and other.

Robustness of propensities

These propensity scores, when used in conjunction with the SFA indicator variable, do exhibit some statistically significant negative effects on the outcomes, see Table 12.6. Given that these variables are positively associated with the treatment and negatively associated with the outcomes, rather than reducing the effect of the treatment, controls for these variables can be expected to enhance the treatment effects, albeit very slightly, and they do; see Fig. 12.1.[15] The treatment effects from Time 0 to Time 1 are slightly smaller in the baseline models than in the models that also include SFA and the propensity scores (i.e., all four signs of the difference between these more complex models and the baseline models are positive). Of these test-score outcomes, the enhancement (1.35 compared with 0.83) of the average treatment effect in the full model is strongest for fourth grade writing for which the negative effects of SFA and the propensity scores are the largest and most significant.

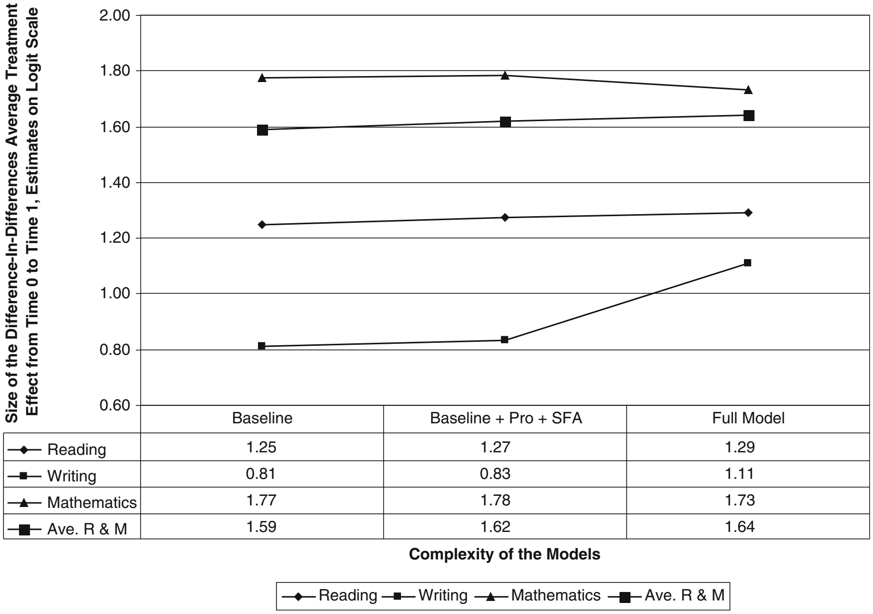
Corroborating these findings, Fig. 12.2 depicts the change in the goodness of fit of the models using as the metric their value of the generalized χ^2 divided by the degrees of freedom (*df*) of the model; smaller values indicate a better fit.[16] Compared with the baseline models, the additional controls for the propensity scores and SFA improve the fits slightly (i.e., all four signs of the differences between these models and the baseline models are negative). Moreover, for the maximum likelihood estimates of the school quality ratings, the BIC statistics indicate that

Table 12.6 The propensity scores (PRO) and Success for ALL (SFA) exhibit noticeable negative direct effects on the outcomes, when the design variables are controlled

Logistic Multilevel Model = Design Variables + PRO + SFA

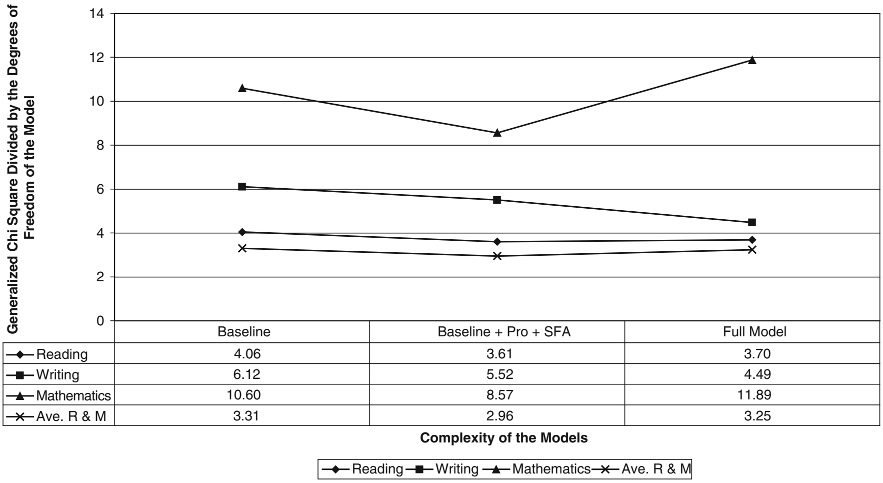
Outcome	Parameter estimates	*t*	Sig. Prob.	Lower	Upper
Reading					
PRO	−7.38	−1.61	0.14	−17.73	2.98
SFA	−1.045	−2.39	0.04	−2.04	−0.056
Mathematics					
PRO	−4.72	1.03	0.33	−15.07	5.64
SFA	−0.92	−2.41	0.04	−1.79	−0.06
Ave. of reading & mathematics					
PRO	−6.13	−1.44	0.18	−15.74	3.48
SFA	−1.07	−2.73	0.02	−1.96	−0.18
Fourth grade writing					
PRO	−14.65	−3.28	0.01	−24.76	−4.54
SFA	−1.63	−3.78	0.004	−2.6	−0.65
***					
School quality ratings					
PRO	−37.10	−2.78	0.009	−64.33	−9.86
SFA	−4.13	−3.28	0.003	−6.70	−1.55

These parameter estimates are on the logistic scale and are taken from multilevel models with appropriate covariance structures. The School Quality Ratings are on the cumulative logit scale.



	Baseline	Baseline + Pro + SFA	Full Model
◆ Reading	1.25	1.27	1.29
■ Writing	0.81	0.83	1.11
▲ Mathematics	1.77	1.78	1.73
■ Ave. R & M	1.59	1.62	1.64

Fig. 12.1 Compared with the Baseline Models, from Time 0 to Time 1 the models that include success for all and the propensity scores very slightly improve the DID treatment effects; In 3 or 4 comparisons the full model continues the improvement



	Baseline	Baseline + Pro + SFA	Full Model
◆ Reading	4.06	3.61	3.70
■ Writing	6.12	5.52	4.49
▲ Mathematics	10.60	8.57	11.89
× Ave. R & M	3.31	2.96	3.25

Fig. 12.2 When added to the baseline models, the propensity scores and success for all slightly improve the fits of the models (i.e., smaller χ^2/df is better). The additional controls of the full models do not necessarily improve these fits

compared with the baseline model, the inclusion of SFA and the propensity scores produces the smallest (i.e., best) value, respectively, 87.1, 71.5, and 79.2.

In sum, the controls for propensity scores and SFA do have some consequences; these variables should be included in the subsequent models of the repeated measures because they do no harm and may do some good.[17]

The Repeated Measures

Scores on standardized tests of achievement gauge the intellectual potential of students narrowly. Because the No Child Left Behind act and TEA require their use, this evaluation primarily focuses on achievement outcomes rather than on such measures as self-esteem and self-direction, which, in any case, were not available. Toward the end of each of the three school years teachers administered in English the TAAS standardized tests of reading, mathematics, and fourth grade writing to the students capable of taking an English-language test. The average of the reading and mathematics test provides an overall summary measure of outcomes as does the school quality rating, a metric that may be questionable because the yearly improvement in the scores of both target and comparison schools may indicate an unpremeditated inflation of the ratings and not improved school quality. The reading and mathematics tests track the same cohort of students from third to fifth grade; the teachers administer the writing tests to the fourth grade students of different cohorts; and TEA rates the school quality each school year.[18] The multilevel models easily accommodate these different repeated measures.

For a cohort of students in a grade of these schools, a typical aggregated test score is the percent passing a TAAS test; for example, 90% pass the test. The statistical analysis conceptualizes this percentage as 90 successes (i.e., events) out of 100 trials and Proc Glimmix uses its events/trials format. Since there are 13 schools observed for 3 years, the total number of observations is 39 and the number of events is 3,900 (39×100). Valuing simplicity, the analyses neither weight these observations by the number of students nor by the implementation scores; the unweighted data suffice.

Glimmix requires the specification of a probability distribution and a link function. For these achievement outcomes the distribution is binomial and the link is the logit; thus the model is logistic regression. The logit is the log to the base e of the odds, that is, the probability of success (p) divided by the probability of not success ($1 - p$); in this example the logit $= \ln(0.90/0.10) = 2.197$. The SAS program transforms the test scores into logits and these transformed scores become the measures on the response variables in the modeling. For each of these tests, the program reports the DID effects and their statistical significance on the logit scale (i.e., the log of the odds scale). It exponentiates these DID effects (i.e., e^{DID}) producing odds and odds ratios, and converts these odds ratios into proportions using the inverse logit function, which is logit $^{-1}(DID) = e^{DID}/\,(1 + e^{DID})$; all are based on the specifications of the multilevel models.

The Multilevel Models

For modeling the available data, the results are influenced by the form of the multilevel statistical model, the covariates, the hypothesized covariance structure, and the metrics of the effect sizes. The multilevel data structure defines the aggregated test scores of students in a school at different time points as the level-1 unit; the school is the level-2 unit. That treatment typology groups the schools conceptually and appears in the model as a fixed effect along with the covariates; the latter are all properties of the schools at different time points and are structural variables. A two-level statistical model for this data structure is:

$$\mathrm{Log}[\pi_{ijk}/(1-\pi_{ijk})] = \mu + \alpha_i + \tau_k + (\alpha\tau)_{ik} + \mathrm{d}_{j(i)} + \mathrm{e}_{ijk} \tag{2}$$

where

π_{ijk} is the proportion of students who pass a test at school j at time period k for a school in the treatment condition i;

μ is the intercept term that includes the effects of the mean-centered covariates;

α_i is the treatment cross-sectional difference (Trt): $i = 1 =$ the target group of schools; $i = 0 =$ the comparison group of schools;

τ_k is the time period classification variable (Time): $k = 0$ is third grade in SY 1999–2000, $k = 1$ is fourth grade in SY 2000–2001, and $\mathrm{k} = 2$ is fifth grade in SY 2001–2002;

$(\alpha\tau)_{ik}$ is the interaction of the treatment and the time periods, $(\mathrm{Trt} \times \mathrm{Time})_{ik}$;

$\mathrm{d}_{j(i)}$ is the covariance parameter of the j^{th} school grouped in treatment i, $\mathrm{d}_{j(i)} \sim iid\,N(0, \sigma^2_{j(i)})$; and

e_{ijk} is the residual covariance parameter of the j^{th} school grouped in treatment i at time period k, $\mathrm{e}_{ijk} \sim iid\,N(0, \sigma^2_{ijk})$.

The model for the writing test in fourth grade is analogous. The Glimmix instructions specify the events/trial format for the test scores, an appropriate covariance structure, the binomial error, and the logit link. Because these runs do not use the repeated statement, the Kenward–Roger corrections for degrees of freedom are not needed here. All of the covariates are centered by the value of their own overall sample mean; the intercept includes these effects. For each cell of the design Proc Glimmix calculates the DID estimates and the least-squares means on the logit scale, and then transforms these values to produce the odds ratios, the predicted proportions, and their upper and lower bounds. All of these models use the same set of fixed covariates.

Covariates

The assignment equation that predicts membership in the target group of schools includes a comprehensive set of available predictors; some of these are highly correlated. Given the propensity scores, which take into account the effects of the predictors implicitly, risking redundancy the subsequent analyses explicitly control for the effects of some of these predictors by including them as fixed covariates in the multilevel models. By eliminating for use as covariates those variables that produced high values of variance inflation and low values of tolerance, a series of ordinary least-squares (OLS) regressions pruned the predictor set of collinear variables. The variables that survived this pruning and that are included in the multilevel models as fixed covariates are the Firth-corrected propensity scores from the logistic regression applied to the Goldbergerized data (GPropensity3), the indicator for Success for All (SFAL), the indicator for highest grade (sixth or fifth grade) of the school (`higraddm`), the proportion of students in the school eligible for free or reduced-price lunches (`frrdprob`), the ratio of students to teachers (`stratio`), the proportion of mobile students (mobility), and the proportion Hispanic (Hispprob). All of these covariates are mean-centered.

SAS code

Here is the SAS code for the basic multilevel model for the reading outcomes:

```
Proc Glimmix Data = Standard Method = rspl Noitprint IC = pq;
   Class school treatment period;
   Model reading/denom = treatment period treatment*period
   GPropensity3 SFAL
   higraddm frrdprob stratio mobility Hispprob
   /solution cl dist = binomial link = logit;
   NLoptions maxiter = 50;
   Random period/sub = school(treatment) type = CS
   s residual vcorr;
run;
```

The first statement calls Proc Glimmix, tells it to use the data from the data set Standard, which contains the mean-centered covariates and also the design variables that are not mean-centered. It then specifies the default rspl method of estimation: r (=residual), s (=expansion locus is the vector of random effects solutions), pl (=pseudo-likelihood); the iteration history table should not be displayed (i.e., No iteration print). The IC = pq option requests that the penalties applied to the fit statistics include the number of fixed effects. The SAS Institute (2005, November 18–32) provides more information about these and other options for the Proc Glimmix statement.

The Class statement specifies that the school, treatment, and time period variables are treated as categorical attributes. Indicator variables will be created for the reform (1) and null (0) treatments and for time period categories, 0, 1, and 2. The ordering of treatment and period on this class statement is important; here treatment comes before time period. Once this ordering has been established on the Class statement, to minimize confusion this ordering should not be changed on the other statements.

The Model statement specifies the structural aspects of the generalized linear multilevel model: the fixed effects, the probability distribution, and the link. The events/trials format for the response variable is indicated by reading/denom (e.g., 90 successes divided by 100 trials). The response variable is modeled as a function of the design variables, the controls for the propensity scores and Success for All, and the covariates. The options requested follow the /. Solutions requests the display of the fixed effects parameters and their measures of statistical significance—the standard errors, *t*-statistics, and probabilities; cl requests the confidence limits. For the events/trials format of the response variable, a multilevel logistic regression model is appropriate, thus the distribution is the binomial and the link is logit. The SAS Institute (2005, November, 62–75) discusses further options for the Model statement.

The statement, NLoptions maxiter = 50, specifies that 50 is the maximum number of iterations for this run.

The Random statement specifies the stochastic aspects of this model; it is analogous to the Repeated statement in Proc Mixed. Here the time period effect for each school is considered to be a residual (i.e., R-Side) random effect with a compound symmetry (CS) covariance structure. The schools are thought to be nested by the treatment, reform or null. (But here the nesting by treatment has no discernable consequences on the parameter estimates and their statistical significance. It does not change the degrees of freedom of the un-nested model.)[19] The s requests that the solutions for any G-side random effects be displayed; Residual specifies that the random effects listed in the statement are R-side random effects; and vcorr requests the display of the correlation matrix for the covariance structure of a school nested by treatment. The Run statement concludes this program. The SAS Institute (2005, November, 7–9, 94–107) further explicates Random statements.

Covariance Structures

The precision (i.e., accuracy) of a parameter is estimated by the inverse of its variance; when the variance is estimated from empirical data, then the inverse of its squared standard error defines the estimate of precision. To enable a multilevel model to produce precise estimates of the statistical significance, standard errors, and confidence intervals of its fixed components, the covariance structure for its random effects should appropriately fit the data being modeled. When the response variable is normally distributed, such goodness-of-fit measures as the deviance, AIC, and BIC can be used for this purpose. Most simply, the analyst specifies the same fixed (i.e., structural) component for the various models, applies restricted maximum likelihood (REML) to estimate the alternative covariances structures (i.e., the stochastic components), and then compares the relative fits of the models using goodness-of-fit measures; REML produces unbiased estimates of the covariance parameters, maximum likelihood (ML) does not. Typically, the analyst chooses as best the model with a unique covariance structure that produces the smallest value of the BIC; Littell et al. (2006, 183–186) and Singer and Willett (2003, 264–265) provide examples of selecting covariance structures based on these measures, as does the previous chapter.

But, when Proc Glimmix is used to estimate generalized linear mixed models with such endpoints as counts, binomial variables, and so forth, and the estimates are based on pseudo-likelihoods, then the standard goodness-of-fit measures should not be used to compare models or to assess the appropriateness of the covariance structures: When estimation based on linearization and pseudo-likelihood is applied to a pseudo-response variable this creates a pseudo-model (Littell et al. 2006, 541). This approach allows generalized linear mixed models to be estimated, but it invalidates the use of the standard goodness-of-fit measures because each model is based on different pseudo-data. Glimmix warns the user about this limitation by printing out this statement: “REML information criteria are adjusted for fixed effects and covariance parameters. Fit statistics based on pseudo-likelihoods are not useful for comparing models that differ in their pseudo-data.” However, the estimates of the generalized chi-square (χ^2) and the generalized chi-square divided by the degrees of freedom of the model (χ^2/df) are useful for comparing some models, as are the new Covtest statements available in SAS 9.2.

A Covtest statement implements a likelihood-ratio test that enables the analyst to ascertain if a more parsimonious, nested covariance structure fits the data just as well as the more comprehensive model that includes it as a special case. Most often, an unstructured covariance model defines the more comprehensive model, and the various covariance tests determine which of its elements, if any, are superfluous. Table 12.7 tabulates the results of the sequential application of these tests to the data of this study. By testing for G-side or R-side random effects, the homogeneity of estimates across groups, the significance of off-diagonal correlations, complete independence, homogeneous off-diagonal correlations, and compound symmetry (CS); these procedures identify nuanced and alternative reasonable choices of covariance structures—CS is the alternative reasonable choice.

This section explicates the logic of these tests using as an example the selection of the CS covariance structure for the reading outcomes. It then discusses the selection of CS for the average of mathematics and reading; the banded main diagonal, UN(1), for the fourth grade writing outcomes; and the two-banded Toeplitz, TOEP(2), for the mathematics outcomes.

G-Side or R-Side random effects

Proc Glimmix estimates G-side and R-side (i.e., residual) random effects; the former appear in the matrix G and the latter in matrix R (SAS Institute 2005 November, 8–9). The variance components (VC) structure, which may include both G-side and R-side parameters, is the default covariance structure of Proc

Table 12.7 Choice of covariance structures for the different achievement test outcomes

		Measures of achievement test outcomes:			
Covariance test parameters:		Reading	Mathematics	Ave. of reading and mathematics	Fourth grade writing
Diagnostic covariance tests					
G-Side random effects[a]		None	None	None	None
Homogeneity across groups[b]		Yes	Yes	Yes	Yes
Significance of off-diagonal Correlations	Corr (2,1)	0.40	0.45	0.53	0.14
	Probabilty	0.167	0.066	0.025	0.649
	Corr (3,1)	0.70	−0.12	0.41	0.20
	Probabilty	0.001	0.711	0.200	0.513
	Corr (3,2)	0.70	0.53	0.75	0.45
	Probabilty	<.0001	0.025	<.0001	0.096
Complete independence: no G-Side, diagonal R-Side	DF	3	3	3	3
	−2Res. Log P-L[c]	88.03	102.76	91.64	112.48
	χ^2	9.19	8.04	10.29	2.52
	Probability	0.027	0.045	0.016	0.472
Homogeneous off-diagonal correlations (Heterogeneous compound symmetry, CSH)	DF	2	2	2	2
	−2Res. Log P-L	81.55	100.54	83.58	110.76
	χ^2	2.71	5.83	2.23	0.8
	Probability	0.258	0.054	0.328	0.669
Compound symmetry	DF	4	4	4	4
	−2Res. Log P-L	83.73	106.74	84.90	112.88
	χ^2	4.89	12.02	2.23	2.92
	Probability	0.299	0.017	0.470	0.57

(continued)

Table 12.7 (continued)

		Measures of achievement test outcomes:			
Covariance test parameters:		Reading	Mathematics	Ave. of reading and mathematics	Fourth grade writing
Nuanced choice		Compound symmetry	Two-banded Toeplitz (2)	Compound symmetry	Banded main diagonal, UN(1)
Alternative choice (i.e., A reasonable choice)		Heterogeneous CS[d]	Heterogeneous CS	Heterogeneous CS	Heterogeneous CS
Predictors in test	Baseline	Yes	Yes	Yes	Yes
	Propensity + SFAL	Yes	Yes	Neither	Yes
	Covariates	All	None	All	None
Fit of Nuanced model	Generalized χ^2	96.25	309.03	84.46	26
	χ^2/DF	3.7	11.89	3.25	1

[a] There are no G-Side random effects by design. [b] The tests for homegenity across groups do not converge; consequently, such homogenity is assumed. [c] – 2Res. Log P-L is an appreviation for −2 residual log pseudo-likelihood. [d] Heterogeneous Compound Symmetry (CSH).

Glimmix. The key word Intercept on this Random statement will produce G-side random effects:

```
Random Intercept / sub = school(treatment) type = unr s vcorr;
```

The following Random statements with different key words produce identical R-side effects for these data:

```
Random _residual_ / sub = school(treatment) type = unr s vcorr;

Random Period / sub = school(treatment) type = unr Residual s vcorr;
```

R-side random effects are closely analogous to those produced by Proc Mixed for normally distributed repeated measures when the analyst specifies a Repeated statement like the following (Littell et al. 2006, 166–167):

```
Repeated Period / sub = school(treatment) type = un r rcorr;
```

For the analysis of these repeated measures using Proc Glimmix, models with the R-side random effects are preferred here because they closely parallel those obtained from the use of a Repeated statement in Proc Mixed. Consequently, all of these analyses estimate R-side and not G-side random effects.

By design, here there should be no G-side random effects and there are none. By testing whether the G matrix can be reduced to a zero matrix, this Covtest statement verifies their absence and can provide benchmark values of the R-side random effects:

```
Covtest 'zerog = no G-side' zerog/cl Wald estimates;
```

Homogeneity across groups

For the schools grouped according to the reform or null treatments, this test can determine if the same variance-covariance structure fits. This test requires a change in the earlier Random statements and this new Covtest statement:

```
Random Period/sub = school group = treatment type = unr
Residual s vcorr;
Covtest 'common variance across groups' Homogeneity;
```

On the Random statement, the group = treatment option tells SAS to create two distinct treatment groups, reform (1) and null (0), and type = unr requests estimates of the unstructured covariance structure (with on-diagonal variances and off-diagonal correlations), one for each group. The Covtest statement requests a test for homogeneity across these two groups. However, for each of the outcome variables, this test does not converge to a solution. Consequently, for each outcome this chapter assumes that the same covariance structure fits both the reform and null treatment groups of schools.

Significance of off-diagonal correlations

The output display from the UNR option for the reading outcomes looks like this:

Box 12.1 Unstructured (UNR) Covariance Parameter Estimates

Covariance Parameter	Subject	Estimate	Standard Error	Z Value	Pr Z	Wald 95% Confidence Limits	
Var(1)	School(Trt)	5.326	2.889	1.84	0.0326	2.306	22.701
Var(2)	School(Trt)	13.902	6.901	2.01	0.0220	6.372	50.401
Var(3)	School(Trt)	7.182	3.901	1.84	0.0328	3.107	30.695
Corr(2,1)	School(Trt)	0.401	0.290	1.38	0.1670	0.168	0.969
Corr(3,1)	School(Trt)	0.700	0.215	3.26	0.0011	0.279	1.121
Corr(3,2)	School(Trt)	0.698	0.167	4.17	<.0001	0.370	1.026

In a 3-row by 3-column matrix of these data, the variances are on the main diagonal in positions (1,1), (2,2), and (3,3) respectively, and the correlations are off-diagonal elements in positions (2,1), (3,1), and (3,2). Two of these correlations are about 0.70 and are clearly statistically significant; Corr(2,1) is substantial but not significant. The significance of such off-diagonal correlations influences whether or not complete independence holds.

Complete independence

By eliminating any G-side random effects and reducing the R-side random effects to a diagonal structure, the following Covtest statement tests for complete independence; that is, all off-diagonal correlations are set to zero:

```
Covtest ‘independent = no G-side,diagonal R-side’ INDEP/cl
Wald;
```

For the reading outcomes, this test compares these parameters:

Box 12.2 Covariance Parameters Based on the Residual Pseudo-Likelihood

Test Data	DF	−2 Res. P-L	χ^2	Pr > χ^2	Var (1,1)	Var (2,2)	Var (3,3)	Corr (2,1)	Corr (3,1)	Corr (3,2)
UNR, (ZeroG)	0	78.84	0.00	1.00	5.33	13.9	7.18	0.401	0.700	0.698
INDEP	3	88.03	9.19	0.027	5.52	10.8	4.49	0.000	0.000	0.000

DF: *p*-value based on a chi-square with DF degrees of freedom.

Because there are no G-side random effects, the use here of the ZeroG test simply reproduces the parameter values for the unstructured variance-correlation matrix specified by the UNR option. Given the absence of G-side random effects, here the

Independence test sets the off-diagonal R-side correlations to zero, recalculates the variances, and tests whether or not this reduced model fits the more inclusive UNR model. The difference in pseudo-likelihood between the two models is 9.19 and the difference in degrees of freedom is 3; consequently, the probability of fit of the independence model is 0.027. This probability decisively rejects the null hypothesis H_0: of no difference between the two models. Inspections of these data confirm this result, as do the significance of two of the three off-diagonal correlations and the substantial size of the third correlation; these correlations may be homogeneous, the same size.

Homogeneous correlations

This test determines whether or not the UNR variance-correlation parameters are consistent with this null hypothesis, H_0: the off-diagonal correlation coefficients are the same. For the data of Box 12.1: Unstructured (UNR) Covariance Parameter Estimates, the following SAS code implements the test for homogeneous correlations:

```
Covtest 'homogeneous off diagonal correlations' General
  0 0 0 1 -1,
  0 0 0 1  0 -1
  /estimates;
```

The key word General on this Covtest statement enables the testing of various combinations of covariance parameters; this code is coordinated to the UNR output table of Box 12.1.[20] The three leading zeros represent the three variances; the zeros allow these variances to vary. In the first numeric line, the first 1 represents corr (2,1), the −1 represents corr(3,1), and that line tests whether corr(2,1) = corr(3,1). The second numeric line again allows the three variances to vary and it tests whether corr(2,1) = corr(3,2). The/Estimates option requests the display of these covariance parameters that define a heterogeneous compound symmetry (CSH) structure:

Box 12.3 Covariance Parameters Based on the Residual Pseudo-Likelihood

Test Data	DF	−2 Res. P-L	χ^2	Pr > χ^2	Var (1,1)	Var (2,2)	Var (3,3)	Corr (2,1)	Corr (3,1)	Corr (3,2)
UNR, (ZeroG)	0	78.84	0.00	1.00	5.33	13.9	7.18	0.401	0.700	0.698
Homogeneous r's	2	81.55	2.71	0.26	6.01	13.4	6.15	0.574	0.574	0.574

DF: *p-value based on a chi-square with DF degrees of freedom.*

The difference of two degrees of freedom and the $\chi^2 = 2.71$ produce a goodness-of-fit probability of 0.26; this test does not reject the null hypothesis of homogeneous off-diagonal correlations. But, inspection of the values of the parameters suggests that the fit of this model is not exceptionally close; perhaps the compound symmetry model will fit better.

Compound symmetry

By specifying equality among the on-diagonal variances, and equality among the off-diagonal correlations, this code tests for compound symmetry (CS):

```
Covtest ‘CS’ General
  1 -1,
  1  0 -1,
  0  0  0  1 -1,
  0  0  0  1  0  -1/estimates;
```

The first numeric line tests that var(1,1) = var (2,2); the second, var(1,1) = var (3,3); the third, corr(2,1) = corr(3,1), and fourth, corr(2,1) = corr(3,2). If these equalities are true, then a viable covariance structure may be CS, as these data test:

Box 12.4 Covariance Parameters Based on the Residual Pseudo-Likelihood

Test Data	DF	−2Res P-L	χ^2	Pr > χ^2	Var (1,1)	Var (2,2)	Var (3,3)	Corr (2,1)	Corr (3,1)	Corr (3,2)
UNR, (ZeroG)	0	78.84	0.00	1.00	5.33	13.9	7.18	0.401	0.700	0.698
Compound Symmetry	4	83.73	4.89	0.30	8.52	8.52	8.52	0.559	0.559	0.559

DF: *p-value based on a chi-square with DF degrees of freedom.*

Inspection of these data coupled with the goodness-of-fit probability = 0.30 indicate that this covariance structure fits the data less well than the UNR structure. But, invoking parsimony, CS is a better covariance structure than UNR for these reading outcomes; it is the nuanced choice, followed by the CSH.

For the average of the reading and mathematics tests, CS also offers the best choice of covariance structure; the pattern of the results is very similar to that for the reading outcomes: By design there are no G-side random effects; by assumption the covariance structures for the treatment and null groups are the same; two of the three off-diagonal correlations are significant and the third is substantial; the complete independence model does not fit ($p = 0.016$); the model composed of diagonal variances and homogeneous off-diagonal correlations fits rather closely ($p = 0.328$), but CS fits even better ($p = 0.470$).

CSH closely fits the data for the writing tests ($p = 0.669$); but, invoking parsimony, the banded main diagonal UN(1) model is more appropriate for these outcomes. This model is composed of different on-diagonal variances and all off-diagonal correlations are zero. These covariance test results support this choice: all off-diagonal correlations are not statistically significant; the complete independence model holds ($p = 0.472$); and the small off-diagonal correlations are homogeneous ($p = 0.669$). Thus, for fourth grade writing outcomes a nuanced choice is UN(1), and a reasonable alternative choice is CSH followed by CS.

The mathematics outcomes engender a unique pattern of covariance test results because there is a very small negative corr(3,1) = –0.12, and the other two correlations are substantial. Consequently, the covariance tests reject both the independence model ($p = 0.045$) that posits off-diagonal correlations of zero, and the homogeneous correlation model ($p = 0.054$) that posits identical off-diagonal correlations. Because the two-banded Toeplitz covariance structure, TOEP(2), is consistent with these results, it is the nuanced choice; CSH is the alternative choice.

Treatment Effects

The analyses of the effects of the reforms focus on three time periods: baseline (Time 0) to the end of the first full year of implementation (Time 1); overall change from the baseline (Time 0) to the end of the second full year of implementation (Time 2); and change from the first full year of implementation (Time 1) to the second full year (Time 2). Initially, the effects are captured by the indicator variables for the interactions between the treatment group (reform = 1 and null = 0) and the time periods (0, 1, 2). Then, these effects are captured more directly by the DID estimators. Because the effects are quantified first on the logit scale, the effects will be interpreted more intuitively as odds, odds ratios, proportions, and differences in proportions. Because educational researchers report their findings using standardized effects sizes, the effects will be interpreted on that metric as well. Illustrating these measures, this section explicates the effects of the reforms on the reading test scores.

Indicator variables

Box 12.5 below presents the coefficients on the indicator variables that quantify the effects on the logit scale of the Period and Treatment × Period interactions. Because SAS creates indicator variables using the highest coded value as the reference category, the meanings of these effects are unclear. Although the following interpretations may appear strained, they are consistent with the more clear-cut interpretations of the DID approach.

Box 12.5 Logit-Scale Estimates of Period and Treatment*Period Effects

Effect	Treatment	Period	Estimate	Standard Error	DF	t value	Pr > |t|
Period		0	−0.5828	0.5687	17	−1.02	0.3198
Period		1	−0.04934	0.5768	17	−0.09	0.9328
Period		2	0	.	.	.	.
Trt*Period	0	0	0.3436	0.6496	17	0.53	0.6037
Trt*Period	0	1	−0.9465	0.6405	17	−1.48	0.1578
Trt*Period	0	2	0	.	.	.	.
Trt*Period	1	0	0	.	.	.	.
Trt*Period	1	1	0	.	.	.	.
Trt*Period	1	2	0	.	.	.	.

Change in the target group from Time 0 to Time 1

In the comparison group the initial level of the response is 0.3436 higher than that at Time 2. At Time 1 it is lower than at Time 2 by −0.9465. Therefore, the overall change from Time 0 to Time 1 in the comparison group is −0.9465 − (+0.3436) = −1.2901. Since the decline in the comparison group implies an equivalent increase in the target group, the latter's increase in the reading test scores is +1.2901 from Time 0 to Time 1. SAS did not calculate the statistical significance of this effect.

Overall change in the target group from Time 0 to Time 2

In the comparison group the amount at Time 0 is 0.3436 higher than it is at Time 2. Therefore, the decline in the comparison group across this period is −0.3436. Since the change in the comparison group implies an equal but opposite change in the target group, the increase across these periods for the target group equals +0.3436. This change is not statistically significant ($p = 0.6037$).

Change in the target group from Time 1 to Time 2

In the comparison group the level at Time 1 is −0.9465 lower than at Time 2. Thus, the level at Time 2 is +0.9456 higher than it is at Time 1. Consequently, the change in the comparison group from Time 1 to Time 2 is +0.9456 and in the target group it is −0.9456; this effect is not significant ($p = 0.1578$).

Corroborating these interpretations of the treatment effects, the following sections show how to derive the DID effects using the Estimate and LSMEstimate statements:

Estimate statements

SAS can calculate the DID estimators for the comparison group and for the target group by taking the difference between the means in later and earlier time periods, and then taking the difference between these differences, as specified by Estimate statements. For the parameters of the reading outcomes tabulated above in Box 12.5, Estimate statements like the following provide the DID estimates:

```
Title ‘Difference in differences estimator for first two time
periods’;
*1 for the comparison group this estimates the difference
between the means from Time 0 to Time 1;
  estimate ‘0,1 - 0,0’period -1 1 0 treatment*period -1 1 0;
*2 for the reform group this estimates the difference between
the means from Time 0 to Time 1;
  estimate ‘1,1 - 1,0’period -1 1 0 treatment*period 0 0 0 -1
  1 0;
```

```
*3 this estimates the difference between those differences,
the treatment effect from Time 0 to Time 1;
  estimate ‘(1,1 - 1,0)vs(0,1 - 0,0)’treatment*period 1 -1
  0 -1 1 0;
```

The first Estimate statement takes the difference in the comparison group between the mean at time 0 and the mean at time 1. For these periods it subtracts the value with the lower code (which here is 0) from the value with the later code (which here is 1), and ignores the final time period (which is 2). Referring to the tabulation above in Box 12.5 for the coefficients on the time periods, it calculates (−1)(−0.5828) + (1)(−0.04934) + 0(0) = 0.53346. Then, for the treatment*period effects, it subtracts the value with the lower coded value (which is 0) from the value with the higher coded value (which is 1) and ignores the final treatment*period. Referring again to Box 12.5., it calculates (−1)(0.3436) + (1)(−0.9465) + 0(0) = −1.2901. Now, collecting terms, +0.53346 + −1.2901 = −0.75664, which is the difference between the two comparison-group means. This quantity can also be calculated by pre-multiplying the six values by the coded values –1 1 0 −1 1 0 and collecting terms: (−1)(−0.5828) + (1)(−0.04934) + 0(0) + (−1)(0.3436) + (1)(−0.9465) + (0)(0) = −0.75664.

Inspection of the code for the difference in means for the reform treatment group indicates that the contribution of the time period to this difference between the means will be the same as for the comparison group, namely 0.53346, and the pattern 0 0 0 −1 1 0 in effect multiplies the treatment*period coefficients by zero and the zero coefficients by −1 1 0. Consequently, the difference between the means in the reform group is simply +0.53346.

Then, the estimate of the DID effect of the reform treatment is the difference between the differences in means: +0.53346 – (−0.75664) = 1.2901, which was found earlier by interpreting the indicator variable coefficients. This value can be obtained more directly by pre-multiplying the values of the treatment × period coefficients in Box 12.5 by these codes 1 −1 0 −1 1 0; that is, (difference in treatment group) – (difference in comparison group), and collecting terms: (+1)(0.3436) + (−1)(−0.9465) + 0(0) + (−1)(0) + (1)(0) + 0(0) = +1.2901. The DID estimators for the overall effect of the reform treatment from Time 0 to Time 2 and the effect from Time 1 to Time 2 can be calculated by analogy, or by taking differences between the least-squares means.[21]

Least-squares means

The logit-scale DID estimators can be calculated more easily using the least-squares means provided by this statement:

```
lsmeans treatment*period/odds ilink cl;
```

The options after the / provide exponentiated values (odds), proportions (from ilink) and confidence intervals (cl). For the reading outcomes on the logit scale, Box 12.6 below tabulates the least-squares means and their retransformed values:

Box 12.6 Treatment × Period Least-Squares Means

Trt	Period	Estimate on logit scale	Standard Error	DF	*t Value*	Pr > |t|	Odds	Proportions (Means)
0	0	1.8702	0.5190	17	3.60	0.0022	6.4895	0.8665
0	1	1.1136	0.3952	17	2.82	0.0119	3.0452	0.7528
0	2	2.1094	0.5607	17	3.76	0.0016	8.2435	0.8918
1	0	2.8366	0.5655	17	5.02	0.0001	17.0581	0.9446
1	1	3.3701	0.6391	17	5.27	<.0001	29.0816	0.9668
1	2	3.4194	0.6471	17	5.28	<.0001	30.5524	0.9683

Referring to Box 12.6, the effect of the reform treatment on the logit scale at Time 1 compared with Time 0 is based on the these differences between the means: $d(0)_m = \mu_{01} - \mu_{00} = 1.1136 - 1.8702 = -0.7566$; $d(1)_m = \mu_{11} - \mu_{10} = 3.3701 - 2.8366 = +0.5335$; and $\hat{\delta}_{mid} = d(1)_m - d(0)_m = 0.5355 - (-0.7566) = +1.2901$. These three "least-squares means estimate" statements can implement these comparisons and can produce optional odds ratios, proportion differences, and confidence levels:

```
lsmestimate treatment*period ‘difference in comp group’ -1
1 0;
lsmestimate treatment*period ‘difference in trt group’ 0 0
0 -1 1 0;
lsmestimate treatment*period ‘(1,1-1,0) versus (0,1 - 0,0)’
   1 -1 0 -1 1 0/or ilink cl;
```

The code for the third statement specifies (treatment group parameters) – (comparison group parameters) as follows: (3.3701 – 2.8366 + 0) – (1.1136 – 1.8702 + 0) = 1.2901. Or, equivalently pre-multiply the parameters respectively by the code 1 −1 0 −1 1 0 and collect terms: (+1)(1.8702) + (−1)(1.1136) + (0)(2.1094) + (−1)(2.8366) + (+1)(3.3701) + (0)(3.4194) = 1.2901, which is the DID estimate on the logit scale of the treatment effect through Time 1; it is statistically significant ($t = 2.25$; $p = 0.0381$; confidence limits = 0.07949, 2.5007).

The DID effects on the logit scale for the effects of the reforms from Time 0 to Time 1 and from Time 1 to Time 2 are calculated analogously.[22] These effects are 0.3436 ($p = 0.6037$) and –0.9465 ($p = 0.1578$); the same as the coefficients on the indicator variables reported earlier in Box 12.5. Because these effects are on the logit scale, more intuitive interpretations as odds, odds ratios, proportion differences; and standardized effect sizes would be helpful.

Odds and odds ratios

Box 12.6 tabulates the odds for the reading outcomes estimated by the multilevel analysis. These are simply the exponentiated values of the estimates on the logit scale; for example, for schools in the comparison group at Time 0, the odds are exp(1.8702) = $e^{1.8702}$ = 6.4896, which implies that for these schools at that time point the ratio of the proportion passing a reading test to the proportion failing the test is 6.4896 to 1. However, for these schools at Time 1, the odds are smaller: exp(1.1136) = $e^{1.1136}$ = 3.0453, which implies that for these schools at Time 1 the ratio of the proportion passing a reading test to the proportion failing the test is only 3.0453 to 1. The ratio of these odds, the odds ratio, is 3.0453/6.4896 = 0.4693; this number is less than 1 indicating a decline in performance. The natural log of this odds ratio is ln(0.469258) = −0.7566, which, as found earlier, equals the difference on the logit scale between the comparison group means at Time 1 versus Time 0. Thus, exponentiation of the difference between these means on the logit scale equals the odds ratio that compares the odds for the comparison group at Time 1 to those odds at Time 0: exp(−0.75664) = $e^{-0.75664}$ = 0.46924 = 3.0453/6.4896.

A similar logic holds for the schools composing the target group. At Time 0 the odds equal exp(2.8366) = $e^{2.8366}$ = 17.0577, which implies that at baseline for these students the ratio of the proportions passing the reading test to those failing this test was 17.058 to 1. At Time 1, after a full year of the reforms, these odds improved: exp (3.3701) = $e^{3.3701}$ = 29.08143, which implies that at Time 1 for these students the ratio of the proportions passing the reading test to those failing this test was 29.08143 to 1. The ratio of these odds is 29.08143/17.0577 = 1.7049; because this number is greater than 1 it indicates improvement in the schools' performance. The natural log of this odds ratio is ln(1.7049) = 0.5335, which, as found earlier, equals the difference on the logit scale between the target group means at Time 1 versus those at Time 0. Thus, exponentiation of the difference between these means (3.3701 – 2.8366) is exp(0.5335) = $e^{.5335}$ = 1.7049 = 29.0816/17.0581.

To review: For the comparison schools the odds ratio that compares the odds of passing a test at Time 1 to the odds of passing a test at Time 0 is 0.46924. For the target schools the odds ratio that compares the odds of passing a test at Time 1 to the odds of passing a test at Time 0 is 1.7049. The DID average treatment effect is the ratio of these odds ratios: 1.7049/0.46924 = 3.6332. This ratio of odds ratios equals the exponentiated value of the DID on the logit scale; namely exp (1.2901) = $e^{1.2901}$ = 3.6332. For the parameters on the logit scale and their odds ratios, SAS calculates the significance level and confidence intervals for the difference between the means in the comparison group, the difference in the means in the treatment group, and the DID treatment effect; it also transform these effects creating proportions.

Proportions and proportion differences

The last column of Box 12.6 presents the proportions of the students at a school passing the reading test for the two treatment groups at each time point. These means can be derived from the odds using this expression: $u_{ij} = \text{odds}_{ij}/(1 + \text{odds}_{ij})$;

here *i* indexes the two treatment groups and *j* the three time periods; SAS's ilink option provides the estimates. Thus, for the comparison group, $\hat{\mu}_{00} = 6.4895/7.4895 = 0.8665$; $\hat{\mu}_{01} = 3.0452/4.0452 = 0.7528$. The difference between these proportions gauges the decline in the proportion passing the reading tests in the comparison group: $\hat{\mu}_{01} - \hat{\mu}_{00} = 0.7528 - 0.8665 = -0.1137$.

For the reform group, $\hat{\mu}_{10} = 17.0581/18.0581 = 0.9446$, and $\hat{\mu}_{11} = 29.0816/30.0816 = 0.9668$. The difference between these proportions gauges the increase in the proportion passing the reading tests in the treatment group: $\hat{\mu}_{11} - \hat{\mu}_{10} = 0.9668 - 0.9446 = +0.0222$. No doubt, a ceiling effect limits this increase. But, the DID estimate compares this proportion difference to that for the comparison group: DID $= 0.0222 - (-0.1137) = 0.1359$. Even though this effect is largely due to the decline in the comparison group, the logic of the DID and the potential outcomes perspective point to the reform treatment as improving the reading scores in the target group from Time 0 to Time 1. This proportion difference does not equal the value obtained from the inverse link function applied to the logit-scale DID for the treatment effect: $e^{1.2901}/(1 + e^{1.2901}) = 0.784$, which does not equal DID $=$ 0.1359. Because of the nonlinearity of the inverse link (i.e., the ilink), it does not usefully quantify the DID treatment effect (Littell et al. 2006, 551–552). Instead, this effect on the proportion difference scale is best quantified by using means on the logit scale, the odds and odds ratios, and the proportions based on the odds ratios, and then taking the difference between the latter proportions.

Standardized effect sizes

Educational researchers usually estimate an effect size by taking the difference between the means of the response variable in the treatment and comparison groups and then dividing this difference by the standard deviation (SD) of the response variable in the comparison group; alternatively, they can use correlational methods that do not require a choice of which SD to use. For these data the SD of the comparison group and the SD of both groups pooled are generally larger than that of the target group; the use of the smaller SD as the devisor could result in an overly optimistic estimate of treatment effects; similarly, the use of a larger SD could result in an overly pessimistic estimate of treatment effects. Because the pooled standard deviation here is intermediate in size between the smaller SD of the target group and the larger SD of the treatment group for the relevant time periods, and because the pooled estimate tends to provide a better estimate of the population standard deviation in the long run (Rosenthal 1991, 16), this chapter uses the pooled estimate to standardize the DID average treatment effects, as illustrated next for the reading tests.

As noted earlier, the DID_{01} average treatment effect on the proportions scale for the reading tests from Time 0 to Time 1 $= 0.0222 - (-0.1137) = 0.1359$. The standard deviation pooled (i.e., subscript *p*) across the two groups and these two time periods is $\mathrm{SD}_{01\mathrm{p}} = 0.10323$. Thus, the effects size $\mathrm{ES}_{01} = \mathrm{DID}_{01}/\mathrm{SD}_{01\mathrm{p}} =$

0.1359/0.10323 = 1.316 = 1.3; that is, the reform treatment is responsible for change of 1.3 SD in the proportion passing the reading tests in the target schools from Time 0 to Time 1. Similarly, for the overall change $\mathrm{ES}_{02} = \mathrm{DID}_{02}/\ \mathrm{SD}_{02\mathrm{p}} = -0.002$ /0.09153 = −0.022 = −0.02; in the target schools this DID effect results in a reduction of −0.02 SD in the proportion passing reading tests scores from Time 0 to Time 2. This reduction is not due to a decline in reading skills in those schools; rather, it is due to the improvement in the comparison schools from Time 1 to Time 2.

The multiplication of the DID for the target group from Time 1 to Time 2 by −1 results in the DID for the comparison group for this period: $\mathrm{ES}_{12\mathrm{c}} = (-1)(\mathrm{ES}_{12\mathrm{t}}) = (-1)\ (\mathrm{DID}_{12})/\mathrm{SD}_{12\mathrm{p}} = (-1)(-0.138)/0.102 = 1.35.$[23] From Time 1 to Time 2 the comparison schools generated a change of 1.35 SD in their proportion of students passing the reading tests; this could be due to their improved student-to-teacher ratios.

Cross-checking the above estimates, the correlational approach, which does not require the choice of a standard deviation, calculates effect sizes using these equations (Rosenthal 1991, 19–20):

$$r = square\ root\ (t^2/(t^2 + df)) \tag{3}$$

and given r, then

$$d = 2r/square\ root\ (1 - \mathrm{r}^2) \tag{4}$$

The values of t and the degrees of freedom (df) are those for a DID effect on the logit scale. For the DID effect of the reading outcome from Time 0 to Time 1, the $t = 2.25$ and the $df = 17$. Thus, $r = (5.0625/(5.0625 + 17))$ ½ = $\sqrt{.22946} = 0.4790$. Then, $d_{01} = 2(0.479)/(1 - 0.22946)$ ½ = 0.958 / $\sqrt{.7705}$ = 0.958/0.8779 = 1.091. = 1.1, which is a little less than the value of 1.3 found above. These estimates are derived from a statistically significant logit-scale effect.

For the overall change from Time 0 to Time 2, $t = 0.53$ and $df = 17$. Then $r = (0.2809/17.2809)$½ = 0.12749, and $d_{02} = 0.25499 / (1 - 0.01625)$ ½ = 0.25499/0.991840 = 0.257 = 0.26, which is more favorable to the target schools than the standardized DID of −0.02 found above. However, these estimates are derived from a statistically insignificant logit-scale effect and are best interpreted as no noticeable difference in overall effect between the target and comparison schools.

For the change from Time 1 to Time 2 in the comparison group, $t = 1.48$ and $df = 17$. Then $r = (1.48^2 / (1.48^2 + 17))$ ½ = (2.1904 / (19.1904)) ½ = 0.3378, and $d_{12} = 2(0.3378) / (1 - 0.114)$ ½ = 0.6756 / 0.9418 = 0.72, which is less than the value of 1.35 found above. Both are derived from a logit-scale estimate that somewhat approaches statistical significance.

Consequently, the small insignificant difference in the overall change in the proportions passing the reading tests between the target and comparison schools masks different dynamics in these schools: the target group change was significant from Time 0 to Time 1 and the effect-size estimates ranged from 1.1 to 1.3 SD units.

From Time 1 to Time 2 the change in the comparison group approached significance and the effect-size estimates ranged from 0.7 to 1.35 SD units. The outcomes for the other tests tend to corroborate these different patterns of change in the two groups of schools, even when the results are corrected for multiplicity, as explicated later in this chapter.

The Results

This section depicts the findings graphically, interprets the patterns, and then takes into account the multiple test outcomes. The tests of reading, mathematics, and their average track changes in cohorts of students in target and comparison schools, as they move from third grade in SY 1999–2000 to fifth grade in SY 2001–2002 (from Time 0 to Time 2). The writing test takes place in fourth grade and the data describe the achievement of fourth graders during each of the three school years. Figs. 12.3 through 12.5 depict the change in the proportions passing reading, mathematics, and the average of the reading and mathematics tests; Fig. 12.6 depicts the change in the proportion passing the fourth-grade writing tests.

The brief title of these figures mentions the outcome, the covariance structure, and the scale of the estimates. After this, the caption reports the DID effects, their significance, and their odds ratios and confidence limits. Each figure displays the values of the depicted proportions below the *x*-axis. From these proportions the model-based DID effects on the untransformed scale can be calculated easily and then standardized by the relevant SDs, these can be calculated from those presented in Table 12.3. As explained later, because of the multiplicity of the outcomes, the step-down Bonferroni and the false discovery rate test key results; see the three panels of Table 12.8 later on below.

The text distinguishes the various effects using these acronyms: DIDL_{01} symbolizes the DID on the logit scale (L) from Time 0 to Time 1; OR_{12} refers to the odds ratio for the target group from Time 1 to Time 2: $\mathrm{OR}_{12\mathrm{C}}$ refers to the odds ratio from Time 1 to Time 2 for the comparison (C) schools; DIDP_{02} symbolizes the DID effect on the proportions scale from Time 0 to Time 2; ES_{02} refers to the effect size from Time 0 to Time 2 for which the DIDP is divided by a standard deviation (SD); and d_{02} equals the effect size calculated using the correlational approach.

In sum, these results will assert that: (1) the reforms in the target schools worked to *preserve* their cohorts' achievement in reading, mathematics, and their average against the declines exhibited by the cohorts in the comparison schools from Time 0 to Time 1; (2) the reduction in the student-to-teacher ratio due to the extra fifth grade teachers *generated* improved scores on these measures by the cohorts in the comparison schools from Time 1 to Time 2, while the target schools held steady as a consequence of the reforms and perhaps because of the extra teachers; and (3) the extra teachers did not affect scores on the fourth grade writing test and, therefore, on this test the overall DID average treatment effect favors the target schools. Although the overall difference by Time 2 between the target and comparison schools on the reading and mathematics tests are not large, these schools reached that point following different paths; this fact crucially influences correct interpretations of these findings.

Reading Tests

Fig. 12.3 clearly depicts the constant high performance in reading (about 0.96 correct) in the cohorts in the target schools and the pattern of dip and recovery in the cohorts in the comparison schools; these proportions are derived from the logit-scale effects estimated by a model with a compound symmetry (CS) covariance structure.[24] Relative to the change in the comparison schools, from Time 0 (third grade) to Time 1 (fourth grade), the target schools significantly improved their cohorts' percentage passing reading ($DIDL_{01} = 1.29$, $p = 0.038$). This favorable change is expressed by the odds ratio and its confidence limits, $OR_{01} = 3.6$ (1.1, 12.2). The proportion differences clearly show that the reforms preserve the achievement of the cohorts in the target schools ($0.967 - 0.945 = 0.022$), while the achievement of the cohorts in the comparison schools decline precipitously ($0.753 - 0.867 = -0.114$); the difference in these differences is $DIDP_{01} = 0.136$. The $ES_{01} = 0.136/0.103 = 1.3$ and the correlational $d_{01} = 1.1$. The differences in the reading SDs (as reported earlier in Table 12.3) for the target and comparison schools point to the homogenizing effect of the school reform treatment: in the target schools the SDs are lower (0.066 and 0.061) than those of the comparison schools (0.075,0.149) across these first two time periods.

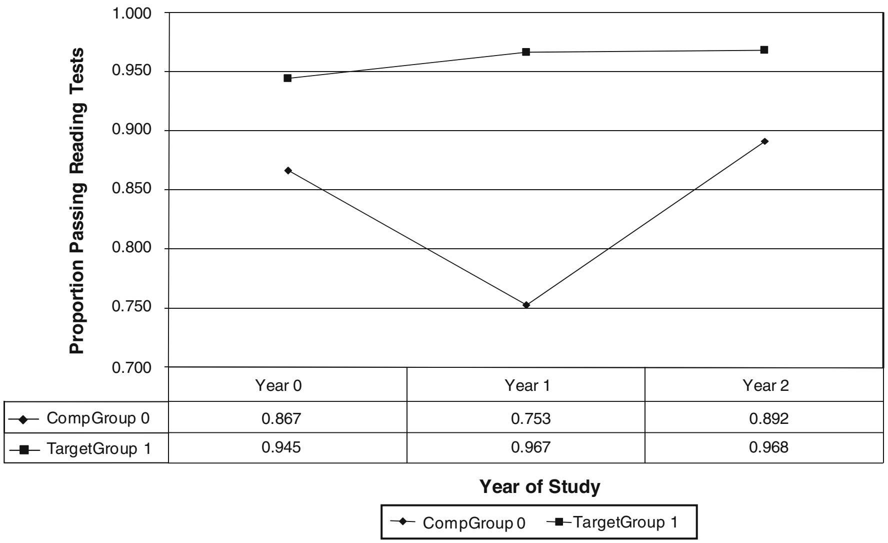


Fig. 12.3 Reading (CS, Logit). Target $DID_{01} = 1.29$ ($df = 17$, $t = 2.25$, $p = 0.038$). $DID_{02} = 0.34$ ($df = 17$, $t = 0.53$, $p = 0.6$). $DID_{12} = -0.95$ ($df = 17$, $t = -1.48$, $p = 0.16$). $OR_{01} = 3.6$ (1.1, 12.2), $OR_{02} = 1.4$ (0.36, 5.6), $OR_{12} = 0.39$ (0.1, 1.5), $OR_{12c} = 2.6$

For the second two time periods, the average treatment effect of the reforms assessed on the logit scale is negative for the cohorts in the target schools ($\text{DIDL}_{12} = -0.95$, $p = 0.16$), even though their predicted proportions passing are stable: The odds ratios for the target schools express the growth in achievement of the cohorts in the comparison schools, and not the decline in achievement of the cohorts in the target schools. The $\text{OR}_{12} = 0.39$ (0.1, 1.5); taking the reciprocal of these estimates produces these favorable odds ratios for the comparison cohorts: $\text{OR}_{12\text{C}} = 2.56$ (0.67, 10); since the lower value is less than unity, these confidence limits indicate that this noticeable effect is not statistically significant.

This favorable change in the comparison schools may be due to the much improved student-to-teacher ratios, which can be viewed as a second treatment effect from Time 1 to Time 2. Changes in standard deviations of the response variable substantiate this notion. Paralleling the improvement in the comparison schools from Time 1 to Time 2, the SDs of the proportions of the cohorts passing the reading test declined, from 0.15 to 0.06; the "treatment" due to the extra teachers reduced the SD of the reading tests. Paralleling the steady high test scores of the cohorts in the target schools from Time 1 to Time 2, the SDs of the proportions of the cohorts passing the reading test are about constant, 0.06 and 0.05

From Time 0 through Time 2 the odds ratios for the cohorts in the target schools are positive, noticeable, but not statistically significant: $\text{OR}_{02} = 1.4$ (0.36, 5.6); which suggests that the overall impacts of the reforms are beneficial. The reforms improved the reading test scores of the cohorts in the target schools from Time 0 to Time 1 by preserving their performance while the performance of the cohorts in the comparison schools declined. Moreover, the cohorts in the comparison schools improved their performance from Time 1 to Time 2; apparently, this improvement was generated by the extra teachers who helped prepare the students for taking the fifth grade tests. In the target schools the reforms (and perhaps the extra teachers) held their cohorts' test scores steady during this period; a ceiling effect may have limited further growth.

Mathematics Tests

The proportions in Fig. 12.4 depict the early significant improvement of the cohorts in the target schools presumably due to the reforms, and the later significant improvement of the cohorts in the comparison schools, presumably due to the extra teachers. These proportions are derived from logit-scale effects estimated by models with a two-banded Toeplitz, TOEP(2), covariance structure.[25] From Time 0 to Time 1 the $\text{DIDL}_{01} = 1.73$ ($p = 0.01$); the $\text{OR}_{01} = 5.66$ (1.6, 20.6); $\text{DIDP}_{01} = 0.126$; $\text{ES}_{01} = 0.126/0.159 = 0.79$ and $d_{01} = 1.4$. The SDs for the target schools indicate the homogenization effect of the reform treatment: the SDs are low and steady (0.057 and 0.054) whereas in the comparison schools the SDs are higher (0.2 and 0.212).

From Time 1 to Time 2 the cohorts in the target schools held steady, but the improvement of the cohorts in the comparison schools accelerated, creating a negative differences-in-differences average treatment effect for the reforms: $\text{DIDL}_{12} = -2.8$

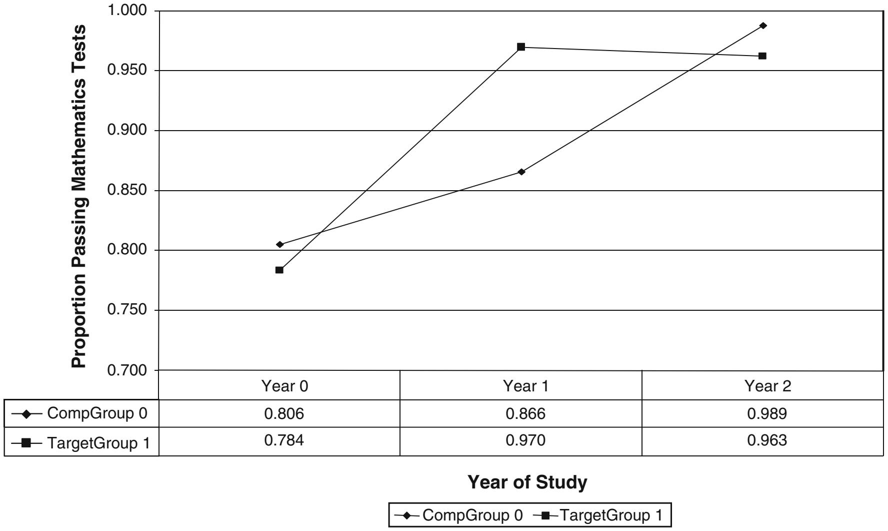


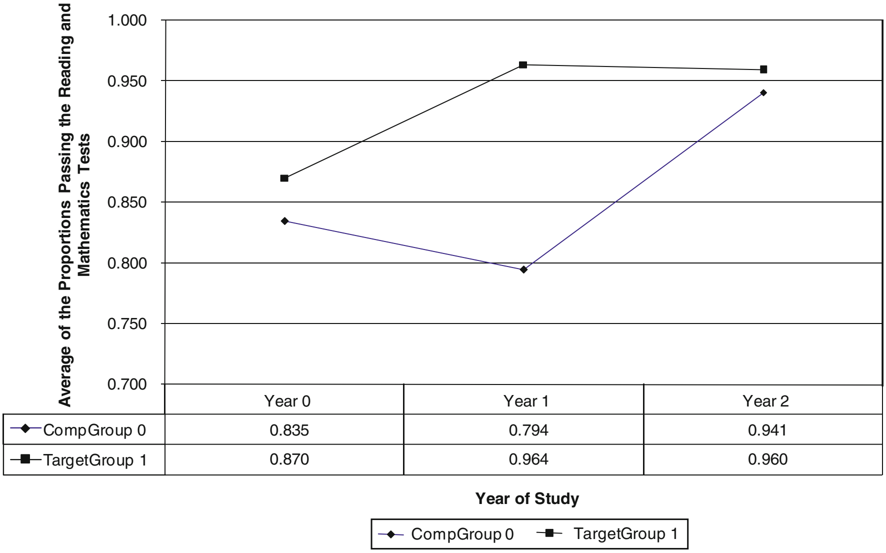
Fig. 12.4 Mathematics (Toep(2), Logit). $\mathrm{DID}_{01} = 1.73$ ($df = 17$, $t = 2.8$, $p = 0.01$). $\mathrm{DID}_{02} = -1.06$ ($df = 17$, $t = -0.8$, $p = 0.45$). $\mathrm{DID}_{12} = -2.8$ ($df = 17$, $t = -2.5$, $p = 0.02$). $\mathrm{OR}_{01} = 5.66$ (1.6, 20.6), $\mathrm{OR}_{02} = 0.34$ (0.02, 6.3), $\mathrm{OR}_{12} = 0.06$ (0.006, 0.65), $\mathrm{OR}_{12c} = 16.4$

($p = 0.02$); the $\mathrm{OR}_{12} = 0.061$ (0.0057, 0.646); $\mathrm{DIDP}_{12} = -0.13$; $\mathrm{ES}_{12} = -0.13/0.135 = -0.96$ and $d_{12} = -1.4$. The positive effects of the teachers in the comparison schools are indicated by reversing the negative signs and by the reciprocals of the odds ratios; the latter are very clearly significant: $\mathrm{OR}_{12C} = 1 / \mathrm{OR}_{12} = 16.4$ (1.55, 175.4!). If these effects are due to the extra teachers in fifth grade in the comparison schools, then the SD in these schools should drop precipitously because of the homogenization effect of the second treatment due to the extra teachers. The SDs do drop precipitously in the comparison schools, from 0.212 to 0.029, whereas in the target schools the SDs do not change much, from 0.054 to 0.066.

By Time 2 the comparison schools exhibit a slightly higher proportion passing than the target schools,0.989 to 0.963. These signal no gross overall difference between of the two types of schools across the three periods: the $\mathrm{DIDL}_{02} = -1.06$ ($p = 0.45$) and the $\mathrm{OR}_{12} = 0.34$ (0.02, 6.3), both measures are not statistically significant. It would be erroneous to infer from these DID statistics that the reforms failed. A more correct interpretation holds that the reforms improved the mathematics achievement of the targeted cohorts and then preserved these gains, and the extra teachers generated the improvements in the comparison cohorts.

Average of Reading and Mathematics Tests

School administrators use the average of the reading and mathematics test scores as a summary measure of the verbal and quantitative development of these elementary



	Year 0	Year 1	Year 2
CompGroup 0	0.835	0.794	0.941
TargetGroup 1	0.870	0.964	0.960

Fig. 12.5 Ave. R. & M. (CS, Logit). Target $\text{DID}_{01} = 1.64$ (*df* = 17, $t = 3.2, p = 0.005$). $\text{DID}_{02} = 0.107$ (*df* = 17, $t = 0.18$, $p = 0.86$). $\text{DID}_{12} = -1.54$ (*df* = 17, $t = -2.45$, $p = 0.026$). $\text{OR}_{01} = 5.2$ (1.8, 15.2), $\text{OR}_{02} = 1.1$ (0.3, 4), $\text{OR}_{12} = 0.22$ (0.06, 0.81), $\text{OR}_{12c} = 4.5$

school children. Because it averages the earlier patterns, which are similar to each other, this measure can be expected to produce findings similar to those for the separate tests, and it does. Inspection of the proportions depicted in Fig. 12.5 clearly shows the early growth in achievement in the target schools and their preservation of this change, whereas the average of these test scores drop and then recover in the comparison schools. The following proportions are derived from logit-scale effects estimated by models with a CS covariance structure.[26] From Time 0 to Time 1 the DIDL = 1.64 and it is statistically significant ($p = 0.005$); $\text{OR}_{01} = 5.2$ (1.8, 15.2); $\text{DIDP}_{01} = 0.135$; $\text{ES}_{01} = 0.135 / 0.121 = 1.1$; and $d_{01} = 1.6$. As expected, the SDs of the response measures in the target schools are small and homogeneous (0.055 and 0.056), and in the comparison schools the SDs are much larger and more diverse (0.132 and 0.165).

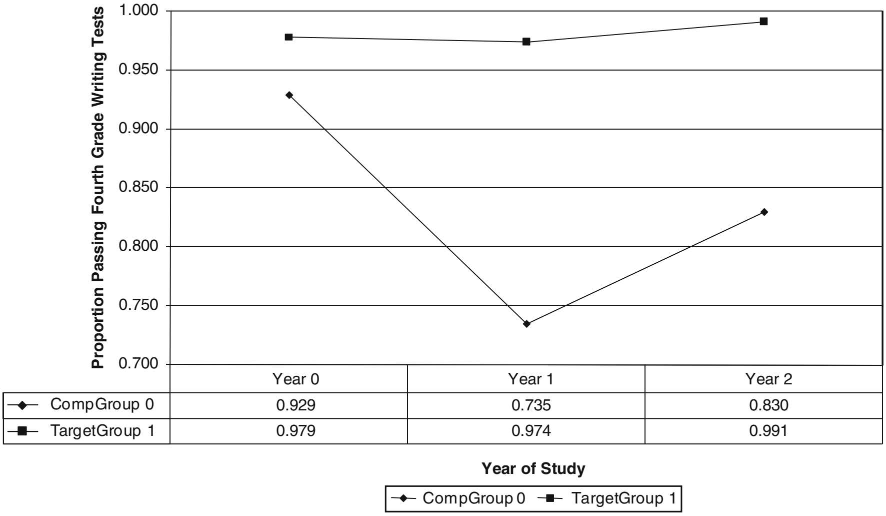
From Time 1 to Time 2 the following DID effect parameters reflect the growth in achievement in the comparison schools and not the decline in achievement of the target schools: $\text{DIDL}_{12} = -1.54$ ($p = 0.026$); the $\text{OR}_{12} = 0.22$ (0.06, 0.81); $\text{DIDP}_{12} = -0.15$; $\text{ES}_{12} = -0.151 / 0.112 = -1.35$ and $d_{12} = -1.19$. Once again, the positive effects of the teachers in the comparison schools are indicated by the reversal of the negative signs and by the reciprocals of the odds ratios; the latter are very clearly significant: $\text{OR}_{12\text{C}} = 1/\text{OR}_{12} = 4.5$ (1.23, 16.7). The SDs again confirm the homogenization effect in the comparison schools due to the extra teachers: from Time 1 to Time 2 these SDs drop from 0.165 to 0.036 while in the target schools they are the same, 0.056 and 0.057.

Masking the dynamics of the change due to the reforms and the extra teachers, across the three periods the overall effects are not statistically significant: $\text{DIDL}_{02} = 0.107$ ($p = 0.86$) and $\text{OR}_{12} = 1.1$ (0.3, 4).

Fourth Grade Writing

These writing tests provide an opportunity to test the extent to which the extra teachers in fifth grade are a causal factor. If, given the earlier evidence that they seem to cause the improvement of the cohorts in the comparison schools on their fifth grade reading and mathematics test scores and their average because of their focus on improving these tests, then these teachers should not affect performance on the fourth grade writing tests. Consequently, the fourth grade writing test scores should not express their effects: the school reforms will be the only treatment; the standard deviations in the target schools will be lower than in the comparison schools, especially at Time 2; the overall DID effects for the target schools should be large and significant; and there should be no significant spurt in achievement in the comparison schools from Time 1 to Time 2. The following data tend to support these stipulations.

For the fourth grade writing tests, Fig. 12.6 presents the proportions passing, which are derived from the logit-scale effects estimated by a model with a banded main diagonal UN(1) structure that includes elements on the main diagonal only.[27] Once again, the target schools maintained a high proportion passing the test, about



	Year 0	Year 1	Year 2
CompGroup 0	0.929	0.735	0.830
TargetGroup 1	0.979	0.974	0.991

Fig. 12.6 Writing (UN(1), Logit). Target $\text{DID}_{01} = 1.35$ ($df = 17$, $p = 1.63$, $p = 0.12$). $\text{DID}_{02} = 1.89$ ($df = 17$, $t = 2.7$, $p = 0.02$). $\text{DID}_{12} = 0.54$ ($df = 17$, $t = 0.55$, $p = 0.59$). $\text{OR}_{01} = 3.86$ (0.67, 22.1), $\text{OR}_{02} = 6.6$ (1.5, 29.6), $\text{OR}_{12} = 1.71$ (0.22, 13.2), $\text{OR}_{12c} = 0.58$

0.98, whereas the comparison schools declined and then slightly improved, averaging 0.83. Thus, from Time 0 to Time 1 the target schools experienced a positive, but not statistically significant, preservative average treatment effect of $\text{DIDL}_{01} = 1.35$ ($p = 0.12$); the $\text{OR}_{01} = 3.9$ (0.67, 22.1) is substantial but also not statistically significant. These positive effects were mostly due to the decline in the comparison schools, as these proportion differences highlight: $(974 - 0.979) - (0.735 -0.929) = -0.005 - (-0.194) = 0.189$. The $\text{ES}_{01} = 0.189/0.092 = 2.05$; the correlational estimate = 0.79.

The following overall effects of the reforms on the fourth grade writing tests will signal the changes in the SDs across the three periods: The $\text{OR}_{02} = 6.6$ (1.5, 29.6) is clearly statistically significant; the $\text{DIDP}_{02} = (0.991 -0.979) - (0.830 -0.929) = 0.012 - (-0.099) = 0.111$; $\text{ES}_{02} = 0.111/0.089 = 1.25$; and the correlational $d_{02} = 1.29$. As expected, the reform treatment homogenizes the SDs of the writing tests in the target schools across the three periods; these SDs decline from 0.113 to 0.086, and then to 0.04 at Time 2. Because the comparison schools experience neither the effects of the reforms nor of the improved student-to-teacher ratios, their SDs on the writing tests are at first constant,0.086 to 0.083, and then increase to 0.109 at Time 2.

The change in the comparison schools from Time 1 to Time 2 is neither large nor statistically significant: $\text{OR}_{12\text{C}} = 0.58$ (0.14, 2.38) and the two effect-size estimators are equivocal: the $\text{ES}_{12\text{C}} = 0.078/0.085 = 0.92$, whereas the correlational estimate $d_{12\text{c}} = (-1)(0.27) = -0.27$; both are based in insignificant logit-scale effects. In the comparison schools the lack of significant growth in fourth grade writing when there were no extra teachers operating indirectly supports the notion that it was the beneficial effects of the extra teachers that improved the proportions passing the reading and mathematics tests in the fifth grade when they were present. The summarizing analysis of the multiple outcomes supports this view.

Multiplicity and Composite Effects

The findings thus far could be vulnerable to inflated probability values (i.e., p-values) stemming from the several correlated response variables. Assessing the range of these effects, the three panels of Table 12.8 present adjusted p-values derived from step-down Bonferroni tests that are stringent, and from the false discovery rates that are less stringent.[28]

These panels also present various χ^2 tests and summary effect sizes that test the key interpretations of the data. The meta-analytic procedures of Fleiss (1981, 160–168) and DerSimonian and Laird (1986) provide the composite effect size across these tests; Chapter 14 explicates these procedures in some detail. It suffices now to say that a composite effect size is calculated by weighting each separate effect by the reciprocal of its squared standard error divided by the sum of the reciprocals of the squared standard errors of all of the effects included in the composite, and summing these products; if the components of the composite are

Table 12.8 Multiplicity and composite effects

Outcome	DID Logit scale	Standard error	t	Prob > |t|	Step-down Bonferroni p	False discovery rate
12.8.1: Baseline (Time 0) to Middle Time Period (Time 1)[a]						
Reading	1.29	0.574	2.25	0.038	0.076	0.051
Mathematics	1.73	0.613	2.83	0.012	0.036	0.024
Ave. R and M	1.64	0.510	3.22	0.005	0.020	0.020
4th Grade writing	1.35	0.828	1.63	0.121	0.121	0.121
12.8.2: Middle (Time 1) to Final Time Period (Time 2)[b]						
Reading	−0.95	0.64	−1.48	0.158	0.158	0.158
Mathematics	−2.8	1.12	−2.5	0.023	0.069	0.039
Ave R and M	−1.54	0.63	−2.45	0.026	0.069	0.039
4th Grade Writing	0.54	0.97	0.55	0.587		
12.8.3: Baseline (Time 0) to Final Time Period (Time 2)[c]						
Reading	0.34	0.65	0.53	0.604	1	0.805
Mathematics	−1.07	1.38	−0.77	0.45	1	0.805
Ave. R and M	0.11	0.61	0.18	0.863	1	0.863
4th Grade writing	1.89	0.71	2.65	0.017	0.068	0.068

[a] χ^2 homogeneous for three basic tests = 0.3, $df = 2$, $0.90 > p > 0.75$; F.E. DIDL = 1.47 (0.73, 2.2). χ^2 homogeneous for all four tests = 0.375, $df = 3$, $0.95 > p > 0.90$; F.E. DIDL = 1.53 (0.94, 2.12). [b] χ^2 homogeneous for three R & M tests = 2.1, $df = 2$, $0.50 > p > 0.25$; F.E. DIDL = −1.46 (−2.28, −0.65). χ^2 homogeneous for R&M composite vs.writing = 3.78, $df = 1$, $p \sim 0.0525$. χ^2 homogeneous for Ave. R&M vs.writing = 3.37, $df = 1$, $p \sim 0.071$. [c] The Bonferroni p-value for 4th grade writing is .051 when the three basic tests are considered. χ^2 homogeneous for composite of R&M and ave. = 0.86, $df = 2$, $p > 0.063$; F.E. DIDL = 0.097 (-0.73, 0.93).

not homogeneous, then random-effects estimates are calculated. The χ^2 statistics provide estimates of the confidence limits of the composite effect and also estimates of the total χ^2 (i.e., χ^2_{total}) and its two components: the associational χ^2 (i.e., χ^2_{assoc}), which gauges the strengths of the effects, and the homogeneous χ^2 (i.e., χ^2_{homog}), which gauges the variability of the components. The percentage of the χ^2_{total} due to χ^2_{assoc} should be much larger than that due to χ^2_{homog}, which gauges the lack of homogeneity of the component effects. If χ^2_{homog} is high, then the null hypothesis of homogeneity will be rejected and the algorithm will calculate the random-effects estimates of the composite. If χ^2_{homog} is small, then homogeneous association may be inferred; if χ^2_{assoc} is large, then real overall association may be inferred (Fleiss 1981, 164).

Change from Time 0 to Time 1

The earlier results suggest that during this period the reforms caused the improvements in the test scores in the preservative sense, the favorable DID effects of the reforms were in large part a consequence of the difference between a small positive difference between the means in the target schools and a large negative difference between the means in the comparison schools. The absence of a large “inter-ocular” positive difference in the treatment group does not necessarily work against a solid average causal effect of the treatment. The governing potential outcomes causal perspective holds that if the target schools had received the null treatment instead of the actual reforms, then their responses would be the same as those of the comparison schools, and, contrarily, if the comparison schools had received the reform treatment, then their responses would be the same as the actual responses of the target schools. The DID estimators reflect this point of view and thus can quantify average causal effects.

Inspection of the uncorrected p-values based on the logit-scale DIDL estimates of the change from Time 0 to Time 1 indicates that the improvements in the target schools are statistically significant in reading ($p = 0.038$), mathematics ($p = 0.012$), and their average ($p = 0.005$), but not in fourth grade writing ($p = 0.121$). Because the average of the two test scores is confounded with the scores on the unique tests that compose it, and the set of measures also includes the writing test, these response variables exhibit multiplicity. As expected, the step-down Bonferroni correction reduces these probabilities: the effect on reading is no longer statistically significant ($p = 0.076$) but the effects on mathematics ($p = 0.036$) and the average of mathematics and reading ($p = 0.02$) maintain their significance; the p-value for writing is unchanged. However, the false discovery p-value for the reading test is at the threshold of significance ($p = 0.051$). Consequently, these corrections for multiplicity do not change the overall intuition that all four effects are favorable during this period, and, with the exception of fourth grade writing, these effects are statistically significant or very close to significance.

This insight can be further tested by applying χ^2 tests to the components of a composite effect as follows: To test that the four response variables exhibit a similar pattern of growth in achievement from Time 0 to Time 1, the null hypothesis H_0 of homogeneous effects should not be rejected; that is, given the degrees of freedom (the number of outcomes in the composite minus 1) and the value of χ^2_{homog}, then the rejection probability should be less than 0.05; unequivocal homogeneous effects will exhibit probabilities much greater than 0.05, circa 0.50 or so.

Confirming the homogeneity of the average treatment effects for the various outcomes, the Notes to panel 12.8.1 presents the relevant χ^2 statistics. When the composite average treatment effect is composed of all four outcomes, then the null hypothesis of homogeneity is decisively not rejected, the goodness-of-fit probability of the null model is between $p = 0.95$ and $p = 0.90$. The percent of the $\chi^2_{\text{total}} = 26$ that is associational is 98.6% versus the 1.4% that is due to the lack of homogeneity. The fixed-effect composite average treatment effect on the logit scale is 1.53 (0.94, 2.12) and the odds ratios are 4.62 (2.56, 8.33), clearly statistically significant. Given the miniscule non-homogeneity component of the χ^2_{total}, the random-effects estimates are not calculated.

Because of the multiplicity induced by the average of the mathematics and reading tests, more prudent estimates result when only the three basic tests—reading, writing, and mathematics—compose the composite; the results are similar to those above: the null hypothesis of homogeneity is decisively not rejected, the goodness-of-fit probability of the null model is between $p = 0.90$ and $p = 0.75$. The percent of $\chi^2_{\text{total}} = 15.7$ that is associational is 98.1% versus the miniscule 1.9% that is due to the lack of homogeneity; the random-effects estimates are not required. The fixed-effect composite average treatment effect on the logit scale rounds to the same value as above: 1.47 (0.73, 2.2), and the odds ratios are similar, 4.35 (0.79, 2.1), clearly statistically significant.

In sum, for the change from Time 0 to Time 1, it is appropriate to say that the reforms in the target schools caused favorable average treatment effects on the outcomes. This pattern of change differs from the other patterns, which also tend to support the notion of favorable effects due to the reform treatment and to the extra teachers assigned in fifth grade.

Change from Time 1 to Time 2

For this period the governing conjecture holds that two treatments were in fact operating: the target schools experienced the effects of the reforms and limited effects of the extra teachers, whereas the comparison schools experienced the effects of the extra teachers but not those of the reforms. In the comparison schools the extra teachers reduced the student-to-teacher ratios to their lowest values. Because of the logic of the DID estimators, the improved performance of the cohorts in the comparison schools induces a negative sign on the average treatment effects for the cohorts in the target schools, even though the performance of these schools held steady near the ceiling of possible achievement. Panel 12.8.2 summarizes the logit-scale effects for the target schools for this period: reading = –0.95 ($p = 0.158$); mathematics = –2.8 ($p = 0.023$), and the average of reading and mathematics = –1.54 ($p = 0.026$). The fourth grade writing tests reflect the impact of the reforms free of the offset due to the extra fifth grade teachers in the comparison schools. On this test the target schools exhibit a positive but not statistically significant effect of the reforms: writing = +0.54 ($p = 0.587$). The estimated effects in the comparison schools can be obtained by multiplying the logit-scale effects for the treatment schools by minus 1 or by taking the reciprocal of the odds ratios.

Because the target and comparison schools experienced different treatments during this period, it is reasonable to correct for multiplicity the p-values for reading, mathematics, and their average, while keeping separate fourth grade writing; panel 12.8.2 displays these results. For these tests the step-down Bonferroni procedure reduces the probabilities: the lack of significance of the effect on reading remains $p = 0.158$ and the probabilities of the effects on mathematics ($p = 0.069$) and the average of mathematics and reading ($p = 0.069$) lose their significance at the 0.05 level. However, the false discovery p-values for the mathematics and the average of reading and mathematics remain significant ($p = 0.039$). Consequently, these corrections for multiplicity do not change the inference that on these test scores all three assumed effects of the extra teachers in the comparison schools are favorable during this period. When there are no extra teachers, as in fourth grade, then the comparison schools exhibit a decline in the fourth grade writing tests that is not statistically significant.

The interpretation that the reforms and the extra teachers—two treatments—operated from Time 1 to Time 2 would be further supported if: (1) the χ^2 tests indicate that the effects on the reading and mathematics test scores and their average are homogeneous; and (2) given a homogeneous composite treatment effect and a separate treatment effect for fourth grade writing, a χ^2 test indicates that these two effects are not homogeneous. The Notes for panel 12.8.2 test these two conjectures. Regarding (1), for the reading and mathematics test and their average, the $\chi^2_{\text{homog}} = 2.1$ and the probability bounds are $0.50 > p > 0.25$; the null hypothesis of homogeneous effects is not rejected. The percentage of the $\chi^2_{\text{total}} = 14.2$, which is due to a lack of homogeneity is 14.4%, and that which is associational is 85.6%. Because of the growth in achievement of the cohorts in the comparison schools, the composite DID fixed effect on the logit scale of the reform treatment is negative, −1.46 (−2.28, −0.65); the odds ratios are 0.23 (0.10,0.52). The reciprocal of those odds ratios measures the effects of the extra teachers in the comparison schools, these are: 4.35 (1.92, 10), statistically significant.

Regarding (2), the χ^2_{homog} for the writing test put together with the composite of the reading and mathematics tests indicates that these effects are disparate: the $\chi^2_{\text{homog}} = 3.78$, $df = 1$, and $p = 0.0525 > 0.05$. Moreover, when the average effect for the mathematics and reading tests is juxtaposed with the writing tests, the disparate nature of these effects is highlighted: the $\chi^2_{\text{homog}} = 3.37$, df = 1, and $p = 0.071 > 0.05$. Given that these goodness-of-fit probabilities are close to the critical value of 0.05 and much smaller than a goodness-of-fit probability of 0.50 ±0.30, there is little evidence to support the view that these effects are not disparate.

In sum, because of the comparison schools' earlier poor performance, extra teachers were assigned to the comparison schools in fifth grade. These teachers reduced the student-to-teacher ratio and induced favorable changes in the students' test scores from Time 1 to Time 2 for reading and mathematics and their average, but not for the fourth grade writing. Because of the logic of the DID estimators, the positive improvement in the comparison schools created negative treatment effects in the target schools on the basic reading and mathematics tests and their average, whereas the absence of the extra teachers in fourth grade allowed the reforms to create a positive DID effect in the target schools for fourth grade writing.

Overall Change from Time 0 to Time 2

Because of the two treatments the overall effects of the reforms in the target schools on the mathematics and reading tests are null; see panel 12.8.3. Contrarily, because of the absence of extra teachers in fourth grade, the overall effect of the reforms in the target schools on the fourth grade writing tests is positive and statistically significant ($p = 0.017$). When the probabilities for all four tests are corrected for multiplicity, then the step-down Bonferroni reduces the significance of the effect on writing to $p = 0.068$; however, when only the three basic tests (reading, writing, and mathematics) are considered, then the Bonferroni corrected p-values for reading and mathematics both are $p = 0.90$ and that for writing is $p = 0.051$. When the writing test is considered separately from the other three measures, then the composite treatment effect for the latter is based on homogeneous component effects ($\chi^2_{\text{homog}} = 0.86$, $df = 2$, $p > 0.63$) and the composite logit-scale fixed treatment effect of 0.097 (−0.733,0.926) is not statistically significant; the odds ratios are 1.1 (0.48, 2.52). Only the overall DID effect of the reforms on fourth grade writing is statistically significant: on the logit scale it is 1.89 (0.385, 3.388); the odds ratios are 6.60 (1.47, 29.6).

Discussion

Summary

This chapter has documented that differences in achievement between the cohorts of students in the target and comparison schools were not due to random fluctuations. Relative to the comparison group, the comprehensive school reforms brought about by Co-nect consultants produced significant improvement in the percentages of students passing tests of mathematics and reading from third to fourth grade, as well as significant improvement from Time 0 to Time 2 in the fourth grade writing test. In the target schools the consultants' reforms enhanced or maintained the students' high proportions passing the achievement tests during the period of their activities.

The performance of the comparison cohorts, which were not exposed to the reform program, generally worsened from third to fourth grade. Then, under the influence of extra teachers, who focused on the students' weaknesses and prepared them for the fifth grade tests, these cohorts regained or even improved upon their earlier performance. Given the logic of the DID estimators, these positive gains obscured the overall effects of the reforms in the target schools.

Policy Implications

In sum, the findings of this chapter and those of the previous chapter imply that comprehensive school reforms and smaller class sizes (indicated by more favorable student-to-teacher ratios) can improve the performance of underachieving ethnic minority students, African American and Hispanic, at least for the period they are exposed to these causes.

Acknowledgments The author produced the initial evaluation report (Smith, September 19, 2003c) as a "third-party evaluator." Dr. Bruce Goldberg of Co-nect was the contracting officer and monitored the project; Sarah Cook of Co-nect provided research assistance; and Dr. Henry May of the University of Pennsylvania critiqued the evaluation report. That report was included in a meta-analysis conducted by Aladjem et al. (September 2006). Because the meta-analysts focused only on the overall effects of the reforms and not on the dynamics of the change, they under-reported the effects of the Co-nect reforms, which this chapter explicates. Vivek Pradhan clarified aspects of separation in logistic regression models and pointed to Firth's solution to this problem. Philip Gibbs and Jill Tao of SAS answered my numerous questions about Proc Glimmix, covariance structures, and the programming of the difference-in-differences estimators. However, the author alone is responsible for the views expressed here, and not Co-nect (which is now a subsidiary of Pearson) or any other organization or person.

Endnotes

[1] Cindy Martin, a Co-nect consultant to the Houston target schools, mentioned these extra teachers in a note (8/19/2003):

> During the 2001–2002 school year North Central District in HISD [Houston Independent School District] concentrated significant district personnel resources to assist schools who were having difficulty in raising their test scores. Additional support was given mostly by North Central supervisors but other available personnel were also used. Assisting these schools became the primary duty of the supervisors assigned [to the] schools with weak test scores. Support personnel worked with teachers and students by demonstrating test taking skills, providing instructional support through modeling lessons, tutoring, working with school leadership in action planning, etc.

The reductions in the student-to-teacher ratios corroborate Martin's observations. Since the time of her note, the North Central District has been eliminated; Reagan is now located in the Central District. The other districts are North, South, East, and West.

[2] The website circa July 2009 for the Reagan High School presents an optimistic image; the school is undergoing comprehensive reform and has a new focus on the teaching of computer programming.

[3] In 2002 the Houston Independent School District (HISD) awarded bonuses to school personnel for improvements in student achievement as measured by standardized tests. Some teachers' aides and janitors received $200 whereas the school superintendent received $25,000, four subdistrict superintendents received $20,000, and other administrators and teachers received awards of $2,500 to $15,000. Because these incentive payments were based on the students' test scores, a number of high schools cheated: the schools' administrators arranged that low performing students not take the definitive tenth grade test. Some students were held back a grade, others were erroneously classified as requiring special education or having limited proficiency in English, still others were encouraged to drop out—at that time dropout rates were not calculated stringently. The net result was a pruning of the pool of potential test-takers so that only the relatively good students would take the tests. For further details see the investigative reporting by Peabody (2003) and Schemo (2003, July 11)

[4] SFA aims to improve achievement outcomes in reading, writing, mathematics, science, and social studies, as well as to improve student attendance, promotion, and discipline. Although it is generally thought to be one of the most successful CSR programs, this chapter reports some negative effects.

[5] Information provided by Co-nect consultants in Houston classified the schools as implementing Success for All or not. This information was based on their personal knowledge and telephone conversations with school administrators.

[6] In their summary of research findings indicating favorable impacts of smaller class size and learning, Murphy and Bella Rosenberg (1998, June 1) state:

> Pupil–teacher ratio and class size technically are not the same thing. Pupil–teacher ratio refers to the number of students divided by the number of staff classified as teachers. Class size refers to the actual number of students in a classroom with a teacher. Pupil–teacher ratio is typically lower than average class size but often used as an approximate measure.

This chapter uses the student-to-teacher ratio as an approximate measure of class size.

[7] The research literature on class size is vast and equivocal: The comprehensive meta-analyses of Glass and Smith (1978, 1979) found that smaller classes not only lead to higher quality schooling and achievement, but also to more positive attitudes. Reporting findings from the Project STAR (Student Teacher Achievement Ratio) studies, conducted in Tennessee from 1989 to 1989, Wood et al. (1990), Frederick Mosteller (1995), and Finn and Achilles (1999) find that smaller class size is especially beneficial for minority children from economically disadvantage families. Jepson and Rivkin (2009) underscore the trade-off between teacher quality and class size. Hoxby (2000) finds that class size does not have a statistically significant effect on the achievement of students in each grade of 649 elementary schools; the author rules out even modest effect sizes of 2–4% of a standard deviation in achievement scores for a 10% reduction in class size. Imbens (2009, 6–9) comments on Hoxby's study and others.

[8] Co-nect's benchmarking scores are derived from surveys of the school's faculty and can be used to assess implementation quality. There are five content areas: *community accountability* (e.g., high expectations, beneficent school climate, shared school improvement plan, community review process, family and community engagement); *high-quality teaching and learning* (e.g., projects and project-based learning, thoughtful discourse, coherent conceptual curriculum, reading and writing, and quantitative reasoning across the curriculum); *comprehensive assessment* (e.g., regular classroom assessment, multiple outcome measures, use of school-level data, community reporting system); *team-based school organization* (participatory instructional leadership, cohesive school community, appropriate grouping, supportive and flexible schedule, collaborative school community); sensible use of technology (e.g., shared vision, access, skillful use in teaching, quality technology training and support, continuous assessment of technology effectiveness). The scores for each of the five dimensions are averaged, leading to an overall percentage score for the school. For 2001 and 2002, respectively, the scores expressed as proportions for these Co-nect schools are: Browning (0.492, 0.426); Crockett (0.418, 0.369); Helms (0.392, 0.370); Love (0.354, 0.477); Milam (0.456, 0.583) Stevenson (0.423, 0.423), and Travis (0, 0). From these proportions implementation scores can be derived by setting all observations on the schools to 1, and then adding these proportions to the values of Co-nect schools in the appropriate time periods. These implementation scores could be used as weights in the mixed modeling, but they to not closely correspond to qualitative observations that implementation of Co-nect reforms was strongest in Helms, Browning, and Love. For the data of this evaluation the unweighted estimates are preferred because there is no information about the Travis school and no information about the implementation of Success for All. Moreover (SFA), the data are not weighted by the number of students in the school, a more obvious weight. Such weightings have very little effect on the results.

[9] Other comparison groups were not added to the design because it was not known which school reforms, if any, were present in the schools that were candidates for these groups.

[10] The author will provide a proof of these relationships upon request.

[11] Wooldridge (2002, 128–132; 2006, 454–459) explicates the difference-in-differences estimator for pooled cross sections: that is, treatment versus control before and after an intervention.

[12] The /Firth option on the model statement of Proc Logistic will provide these estimates.

[13] Project Grad aims to improve educational outcomes in a feeder system through a combination of curriculum change, persuading students to achieve, and financial incentives that offset some of costs of college for those who graduate.

[14] Imbens and Wooldridge (2009, 33) note that: "individuals with propensity scores of 0.45 and 0.50 are likely to be much more similar than individuals with propensity scores equal to 0.01 and

0.06.” The propensity scores in this analysis are not extreme and this supports their use in a regression model.

[15] Given the decomposition rules of path-analytic models, the zero-order effect of the treatment equals the conditioned effect of the treatment plus the spurious component due to antecedent variables like the propensity scores and SFA. When the spurious components are positively related to both assignment to the treatment and the outcome, then the conditioned effect of the treatment equals the zero-order effect minus the spurious component; the conditioned effect of the treatment will be smaller than the zero-order effect. In this analysis SFA and the propensity scores both have negative effects on the outcomes and positive effects on assignment to the treatment. Thus, the spurious component has a negative sign and the conditioned effect equals the zero-order effect plus the spurious component; the conditioned effect is larger than the zero-order effect by the amount of spuriousness. The classic literature refers to this as “suppression” of the zero-order effect by the absence of controls for the other variables.

[16] In Figure 12.2 for the models bearing on the fourth grade writing tests, an alternative covariance structure, compound symmetry, is applied rather than the UN(1), which is used subsequently, because the latter’s value of the generalized χ^2 divided by the degrees of freedom is always 1.

[17] When SFA and the propensity scores are deleted from the full models and the reduced models are re-estimated, then the fits of the reading and writing models worsen: the χ^2/df increase to 5.02 and 5.31; respectively. For mathematics the χ^2/df increases slightly to 11.05. The $\chi^2/df = 3.22$ for the average of reading and mathematics is about unchanged.

[18] These different measures illustrate various distinctions of Lazarsfeld and Menzel (1972, 227–229) concerning properties of collectives. The measures on cohorts of students in the schools are analytical properties of the schools; the writing tests are analytical properties of the fourth grade of the schools; and the quality ratings are pseudo-global characteristics of the schools.

[19] Jill Tao of SAS clarifies the function of this nesting operator in this applications as follows:

> Hi Bob,
>
> The reason why the results are identical between subject = school and subject = school (treatment) is probably because your schools have unique ids; in other words, schools are uniquely identified by the value of itself. You will see different results between the two specifications when the same set of school ids are used across treatments. For example, for treatment 1, you have schools 1 to 10, for treatment 2, the schools are also labeled 1 to 10, although they are different schools, in which case, you must use school (treatment) to uniquely identify the schools.
>
> The GROUP = option does totally different things. When you say GROUP = treatment, it instructs Proc Mixed to estimate separate sets of variance-covariance for different treatment. So if you have 3 treatments, you would have 3 sets of the covariance parameter estimates, one for each treatment.

[20] If the UN option is used, then only variance and covariances are printed out and these data appear in a different order, thus necessitating a different code for this test.

[21] The estimate statements for the overall change and change from Time 1 to Time 2 follow:

```
Title 'Difference in differences estimator for the third time period
(Time 2)relative to the first (Time 0)';
   * for comparison group;
   estimate '0,2 - 0,0'period -1 0 1 treatment*period -1 0 1;
   *for the target group;
   estimate '1,2 - 1,0'period -1 0 1 treatment*period 0 0 0 -1 0 1;
   *this estimates the difference between the two differences above;
   estimate '(1,2 - 1,0)vs (0,2 - 0,0)'treatment*period 1 0 -1 -1 0 1;
Title 'Difference in differences estimator for the third time period
(Time 2)relative to the second (Time 1)';
   *for the comparison group;
```

```
estimate '0,2 - 0,1'period 0 -1 1 treatment*period 0 -1 1;
*for the target group;
estimate '1,2 - 1,1'period 0 -1 1 treatment*period 0 0 0 0 -1 1;
*this estimates the difference between the two differences above;
estimate '(1,2 - 1,1)vs(0,2 - 0,1)'treatment*period 0 1 -1 0 -1 1;
```

[22] The lsmestimate statements have this form:

```
*DID estimates for Time 2 - Time 0;
lsmestimate treatment*period 'diff in comp group' -1 0 1;
lsmestimate treatment*period 'dif in trt group' 0 0 0 -1 0 1;
lsmestimate treatment*period '(1,1−1,0) versus (0,1 - 0,0)'
   1 0 -1 -1 0 1/or ilink cl;
*DID estimates for Time 2 - Time 1;
lsmestimate treatment*period 'diff in comp group' 0 -1 1;
lsmestimate treatment*period 'dif in trt group' 0 0 0 0 -1 1;
lsmestimate treatment*period '(1,1−1,0) versus (0,1 - 0,0)'
```

[23] DID_{12} can be calculated easily from the proportions of Box 12.6 by taking the difference between the proportions in the treatment group minus the difference between the proportions in the comparison group for those time periods: (0.9683 −0.9668) − (0.8918 −0.7528) = 0.0015 −0.139 = −0.1375. For the comparison group this difference is (−1)(−0.1375) = 0.1375.

[24] For reading tests, the estimated V correlation matrix for the school nested by treatment, based on a CS covariance structure looks like this:

Row	Column 1	Column 2	Column 3
1	1.0000	0.5470	0.5470
2	0.5470	1.0000	0.5470
3	0.5470	0.5470	1.0000

All on-diagonal correlations have the same value and all off-diagonal elements have the same value. The R-side covariance parameter estimates are school (treatment) = 4.47 (0, 10.7268), Z = 1.4, Pr Z = 0.1614; the residual = 3.702 (2.1391, 7.9112), Z = 3.08, Pr Z = 0.001.

[25] For mathematics tests, the estimated V correlation matrix for the school nested by treatment, based on a TOEP(2) covariance structure looks like this:

Row	Column 1	Column 2	Column 3
1	1.0000	0.6066	
2	0.6066	1.0000	0.6066
3		0.6066	1.0000

There are two bands of information; this structure lacks parameters in row 3 and column 1 and in row 1 and column 3. The TOEP(2) R-side covariance parameters are both statistically significant. School (treatment) = 7.21.47 (1.5288, 12.8913), Z = 2.49.4, Pr Z = 0.0129; the residual = 11.8856 (6.8351, 25.6330), Z = 3.05, Pr Z = 0.0012.

[26] On the average of the reading and mathematics tests, the estimated V correlation matrix for the schools nested by treatment, based on a CS covariance structure looks like this:

Row	Column 1	Column 2	Column 3
1	1.0000	0.5415	0.5415
2	0.5415	1.0000	0.5415
3	0.5415	0.5415	1.0000

All on-diagonal correlations have the same value and all off-diagonal elements have the same value. The R-side covariance parameter estimates are school (treatment) = 3.836 (0, 9.1123), Z = 1.42, Pr Z = 0.1542; the residual = 3.2486 (1.8618, 7.0528), Z 7= 3.02, Pr Z = 0.0012.

[27] For the fourth grade writing tests, the estimated V correlation matrix for the school nested by treatment, based on a UN(1) covariance structure looks like this:

Row	Column 1	Column 2	Column 3
1	1.0000		
2		1.0000	
3			1.0000

All on-diagonal correlations have the same value and all off-diagonal elements are null. The R-side covariance parameter estimates are: UN (1, 1) school (treatment) = 2.2318 (0.9381, 10.414), Z = 1.76, Pr Z = 0.0389; UN (2, 2) school (treatment) = 14.2834 (6.9128, 44.9467), Z = 2.2, Pr Z = 0.0138; and UN (3, 3) school (treatment) = 4.5156 (2.1249, 15.2605), Z = 2.1, Pr Z = 0.0177. The generalized $\chi^2 = 26$; $\chi^2/df = 1.00$.

[28] The section on "Corrections for Multiplicity" in the previous chapter explicated the step-down Bonferroni. To illustrate the calculations for the step-up false discovery rate using the four obtained *p*-values of panel 12.1, first organize them in ascending order from left to right as done in the box below, and then apply the equations below (SAS 1997a, 802):

Number of Parameters	R-3 = 1	R-2 = 2	R-1 = 3	R = 4
Symbols	p_1	p_2	p_3	p_4
Raw *p*-values	0.005	0.012	0.038	0.121
Adjusted *ps* False Discovery Rate	$s_1 = 0.020$	$s_2 = 0.024$	$s_3 = 0.051$	$s_4 = 0.121$
Adjusted *ps* Bonferroni	$s_1 = 0.020$	$s_2 = 0.036$	$s_3 = 0.076$	$s_4 = 0.121$

$s_{\mathrm{R}} = p\mathrm{R} = 0.121$;
$s_{(R\text{-}1)} = \min\,(s_{\mathrm{R}}, [R/(R-1)p(R-1)] = 0.121$ or $(4/3)(0.038) = 0.121$ or $0.0505 = 0.051$;
$s_{(R\text{-}2)} = \min\,(s_{(\mathrm{R}\text{-}1)}, [R/(R-2)p(R-2)] = 0.051$ or $(4/2)(0.012) = 0.051$ or 0.024;
$s_{(R\text{-}3)} = \min\,(s_{(\mathrm{R}\text{-}2)}, [R/(R-3)p(R-3)] = 0.024$ or $(4/1)(0.005) = 0.020$.
The false discovery correction retains the significance of three of the three significant raw probabilities; the more stringent Bonferroni retains the significance of only two. SAS (1997a, 798–802) discusses the assumptions of these and other tests.

Part IV
Research Summaries

Chapter 13
Consolidations and Critiques

America's health care is the costliest in the world, yet quality is patchy and millions are uninsured. Incentives for both patients and suppliers need urgent treatment. . . . If the United States couples its efforts to expand coverage with. . . a radical restructuring of the underlying drivers of cost inflation, there is every reason to think its health system can become the best in the world – and not merely the priciest.

—*The Economist* (June 27, 2009, 75, 77)

Two social problems of health care in the USA motivate the research summaries of the subsequent two chapters. On the one hand, many people have restricted access to appropriate care because of their lack of health insurance and spiraling costs. On the other hand, preventive vaccinations are available to children, but some parents restrict access to this care because of mistaken beliefs that vaccinations cause autism and have other adverse consequences. To address this first problem of restricted access to medical care, because of the lack of insurance and funds to cover its costs, many people advocate health insurance reforms. Regardless of what components of the Obama Administration's comprehensive reforms are eventually instituted, the implementations will need to reduce the utilization and costs of health care appropriately. The first of these chapters bears directly on this problem. It applies meta-analytic techniques to summarize findings from evaluative studies that assessed the bounded effectiveness of precertification and onsite review nurses in reducing costs to the insurer by averting inpatient admissions and hospital care that lack evidence of medical necessity. In principle, Medicare and Medicaid plans, which are predominantly fee-for-service plans, could reconsider applying these techniques with the aim of reducing costs.

To address this second problem of parental reluctance to vaccinate their children, because of their beliefs about adverse consequences of childhood vaccinations, the second of these chapters summarizes the results of assessments of the scientific evidence for and against a causal association. It does so by considering how health scientists and biostatisticians reviewed the evidence from epidemiological and other studies about the linkage of childhood vaccinations to the onset of autism and other maladies. The original reviews were conducted by Demicheli et al. (2005) and the

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_13, © Springer Science+Business Media B.V. 2011

Institute of Medicine (1994, 2001a, b, 2004). Based on these reviews, this chapter echoes their findings that the evidential support for such causal connections does not meet scientific standards. However, based on anecdotal evidence from parents of autistic children, some overly concerned and overly protective parents still cling to their erroneous beliefs that childhood vaccinations cause autism and are otherwise harmful, even though health scientists and courts-of-law have debunked any causal linkages on the basis of rigorous evidence.[1] The reticence of such parents to vaccinate their children not only endangers the health of their own children, but it also weakens the vaccination coverage in their area, thereby endangering other people's children, both locally and internationally. Dearth of vaccinations creates avoidable illnesses and deaths that are tragic, thereby reducing human development, especially in the least developed countries (United Nations Development Program 2003, 8–10; 2010, 201).

Although this book can do little to change the attitudes of such parents directly, it can help the reader learn how experts have evaluated evidence for or against supposed causal relationships, and thus help to weaken anti-vaccination and other implausible beliefs indirectly (Mazur 2008). The concluding chapter of this book provides an opportunity for the reader to apply principles bearing on assessments of causality. Focusing on the adequacy of the evidence for assumed causal relationships, it presents the author's critical summary of the findings of those chapters that present multilevel models. The reader can make her or his own judgments and disagree.

Problems of Health Care

The absence of insurance coverage, rising costs, and spotty care characterize the malfunctioning health care system of the USA, which President Obama's reforms aim to correct.

The Uninsured

From 1987 to the recent past (circa 2007) the number of Americans without health insurance has increased from about 31 million, or 13% of the population, to about 46 million, or 15.3% of the population (DeNavas-Walt et al. 2008, Fig. 6, 22). In the USA, economic recessions worsen this problem because many insecurely employed and newly unemployed citizens lose their employer-provided health insurance and are unable to pay for private insurance or medical care. For the uninsured, the costs of health care and medical insurance have become prohibitively expensive. Consequently, the quality of health care varies across social and economic categories, as does support for comprehensive reform.

For several social categories, Table 13.1 compares the ratio of the proportion of a category's uninsured to its proportion of a total population. A ratio greater than 1 indicates that the social category is overrepresented among the uninsured; a ratio of 1 indicates no disparity; and a ratio less than 1 indicates that the social category is

Table 13.1 For pivotal social categories, the ratio of the proportion uninsured to the proportion of a population

Social category	Proportion of the total population (base number is 299,106,000)	Proportion of the total uninsured (base number is 45,657,000)	Ratio of the proportion of uninsured to the proportion of a population
Ethnicity			
Hispanic	0.154	0.323	2.10
Black	0.126	0.162	1.29
Asian	0.044	0.049	1.11
White	0.658	0.450	0.68
Household Income			
<$25,000	0.185	0.297	1.60
$25,000–$49,999	0.230	0.318	1.38
$50,000–$74,999	0.195	0.186	0.95
$75,000 or more	0.390	0.200	0.51
Age in Years			
Under 18	0.249	0.179	0.72
18–24	0.095	0.175	1.84
25–34	0.134	0.226	1.69
35–44	0.141	0.169	1.20
45–64	0.258	0.236	1.09
65 or older	0.123	0.015	0.12
Work Experience, 18–64 year old	(Base number is 187,913,000)	(Base number is 36,822,000)	19.6%
Full-time	0.659	0.570	0.87
Part-time	0.132	0.157	1.19
Did not work	0.209	0.271	1.30

DeNavas-Walt et al. 2008, Table 6, 22.

underrepresented among the uninsured. Ethnic minorities, poor people, and younger adults are overrepresented: Regarding ethnicity, Hispanics are very over represented (2.1), followed by African-Americans (1.29), and Asians (1.11). Although Whites have the highest number of uninsured (20,545,650), relative to their proportion in the population, they are underrepresented (0.68). Regarding household income, the poorer the family, then the greater the overrepresentation among the uninsured, from 1.60 for households with an income less than $25,000 per year, to 0.51 for those with an income of $75,000 or more. Regarding age, children under 18 (0.72) and especially adults 65 years or older (0.12) are underrepresented among the uninsured; the former possibly because they are covered by their parent's insurance and targeted governmental programs; the latter through Medicare. However, children living in poverty are more likely to be uninsured compared with all children, 17.6 to 11% (DeNavas-Walt et al. 2008, Figs. 8, 24).

Among the adult age categories, the uninsurance ratio of 1.84 for younger people 18 to 24 years old is much higher than the ratio of 1.09 for older people 45 to 64 years old. This difference is probably due to different rates of employer-provided health insurance. For those in the prime of their lives, that is, from 18 to 64 years old, 19.6% are uninsured, which is higher than the 15.8% for the total population. The uninsurance ratio for full time workers is 0.87; for part-timers, 1.19; and for those not working, 1.30.

Employer-provided health insurance dominates the insurance market. In 2007 the percentage of the population covered by this employee benefit was 59.4%, an additional 8.9% purchased their own private insurance, 27.8% were covered by government-provided insurance, and 15.3% were not covered by any insurance; because of multiple coverage, the sum of these percentages is 111.4% (DeNavas-Walt et al. 2008, Figs. 7, 21). Thus, the current health insurance system is a mixture of both private and public insurance; under the current reforms it is likely to remain so, regardless of inefficiencies and inflating costs.

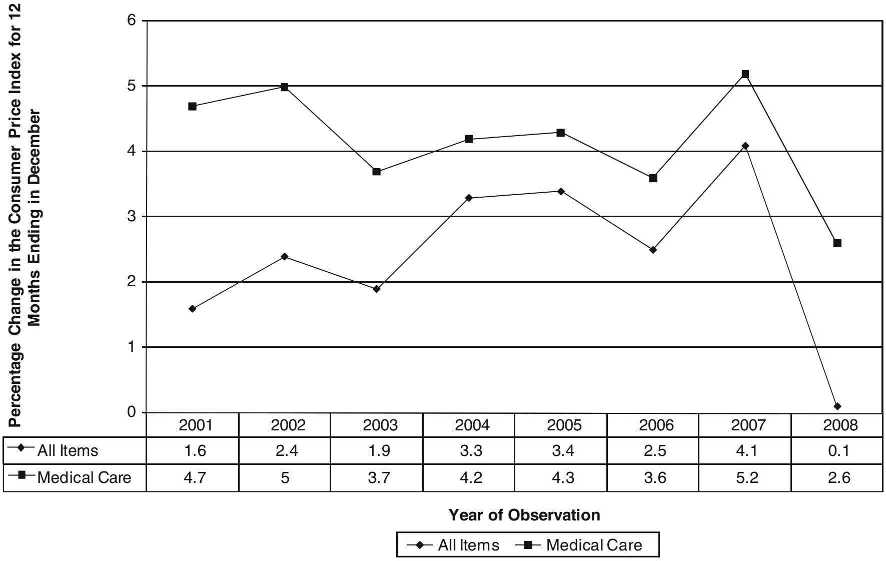
Medical Costs

Inequities increase costs in addition to such drivers of medical cost increases as innovative medical treatments and diagnostic technology, the absence of information technology that can improve the coordination of care, provider incentives that encourage utilization and not quality of care, insurance plans that aim to reduce risk and not improve quality, and patients who do not follow fitness and nutrition guidelines. Because of the purchasing power of managed medical care networks, the cost for the same unit of medical services for the insured population often is less than that for the uninsured. Because many poor people receive free services in emergency rooms, the hospitals may make up the costs they incur by in-effect increasing fees for the services they provide to patients with insurance coverage. These different costs for the same services add to administrative expense that contributes to medical price inflation, which is considerably higher than the inflation gauged by the consumer price index (CPI).

For the period 2001 through 2008, Fig. 13.1 depicts the yearly percentage change for the 12 months ending in December for all items of the CPI compared with that for medical care. At present, health care spending accounts for about 17.6% of the gross domestic product in the USA (Siska et al. 2009), but the actual resources (e.g., nurses per 1,000, physicians per 1,000, acute care beds per 1,000) devoted to health care lag behind other developed nations (Organisation for Economic Co-Operation and Development 2000). Even so, the affluent and securely employed can obtain health care that may be the best in the world, but for the poor and uninsured their health care may be inadequate.

Quality of Care

The USA spends more on health care per capita and as a percentage of its gross domestic product than any other affluent democracy that is a member of the



	2001	2002	2003	2004	2005	2006	2007	2008
All Items	1.6	2.4	1.9	3.3	3.4	2.5	4.1	0.1
Medical Care	4.7	5	3.7	4.2	4.3	3.6	5.2	2.6

Fig. 13.1 Yearly percentage change of consumer price index compared with its medical care cost component, 2001 to 2008

Organisation for Economic Co-operation and Development (OECD), the Group of 20 (G20). Yet, it ranks considerably below countries like Japan, where the life expectancy is 82.1 years compared with the 77.9 years in the USA (Organisation for Economic Co-Operation and Development 2007). Although various indicators document the improvements in health in the USA, its citizens experience higher than appropriate rates of infant mortality, asthma mortality, obesity, and Hepatitis B; the Canadian health care system produces better odds of surviving colorectal cancer and childhood leukemia than the US system (Organisation for Economic Co-Operation and Development 2007; Daschle 2008, 32–38). These and other indicators of spotty care vary from region to region of the USA (Burd-Sharps et al. 2008, 50–79), and stem in part from the lack of appropriate health insurance and restricted access that results in questionable quality of care and poor health; contemporary health care reforms aim to alleviate these problems.

Policy Orientations and Reform Plans

Because people have different experiences with health care delivery, and different beliefs about the appropriateness of governmental interventions in health care, attitudes about comprehensive reform vary. The public and its representatives are divided about the roles of the federal and state governments on the one hand, and the private sector of health insurers and employers on the other. Moreover, the more

Table 13.2 Three Policy Orientations about American Health Systems

Value orientation	Health care is a right	Health care is both a privilege and a right (i.e., Ambivalent)	Health care is a privilege
Preferred System	Public plans, "Medicare for All"	Mixed public and private plans	Only private plans (an employee benefit)
New Public Plan, 8/17/09	Strongly for = 33%	Moderately for or against = 32%	Strongly against = 35%
Government Reform of Health Care, 8/17/09	Strongly necessary = 35%	Somewhat for or against = 30%	More harm than good = 35%
Two-Item Index of Comprehensive Reform, 11/03/92	Favored (+ +) = 29%	Some reform (+ − or − +) = 39%	Opposed (− −) = 32%
Political Support, 8/17/09	The left liberals = 20%	The center moderates = 40%	The right conservatives = 37%
Political Party, 8/17/09	Democrats = 35%	Independents = 34%	Republicans = 35%
Typical Organizational Support	Labor Unions, Progressive Social Movements, MoveOn.com	American Association of Retired People	Industry Lobbyists, FreedomWorks, Tea Party Patriots, ResistNet

The Washington Post-ABC News Poll, conducted 13–17 August 2009. Index constructed from two items in 1992 the Election Night survey conducted by Frederick/ Schneiders; See Smith (1999b, 30) for the index.

fortunate are not eager to accept the tax increases needed to pay for the health care insurance of the less fortunate (Smith 1993, 56; Toner and Elder 2007, Blendon et al. 2009). Encapsulating these differences, a person's basic philosophical orientation toward systemic reform depends upon whether he or she thinks health care is a right or a privilege; or, indicating ambivalence, both a right and a privilege.[2] Based on opinion polls Table 13.2 classifies about a third of the public as believing that health care is a right, another third as believing that health care is a privilege; and the remaining middle-third as ambivalent. These distinctions parallel the basic political cleavages in the USA that can be summarized by political party identification as Democrat, Independent, or Republican, or by political philosophy as liberal, moderate (i.e., perhaps pragmatic), or conservative.[3] On the political left, labor unions and such progressive social movements as "MoveOn.org" support comprehensive reforms and public health insurance; the political center (represented here by the American Association for Retired People) wants an improved public and private system that facilitates the health care of senior citizens; and the radical right associated with the "Tea Party Patriots," "FreedomWorks," and "ResistNet" strongly opposes governmental interventions in health care, especially if it would raise taxes, lead to "socialized" medicine, or provide women with abortions.

Because the political right is strongly against a transformative public plan that might eventually dismantle private insurance, and the political left strongly opposes the shortcomings of private insurance, the Obama reforms strike a pragmatic middle course aiming to improve the present mixed system composed of public and private insurance. As Tom Daschle (2008, 107) optimistically wrote: "The consensus in the middle of the political spectrum, among both Democrats and Republicans, is that we should create a public-private hybrid that preserves our private system within a strengthened public framework." However, the reforms being instituted build incrementally on the existing mixed private and public system without a new public option.

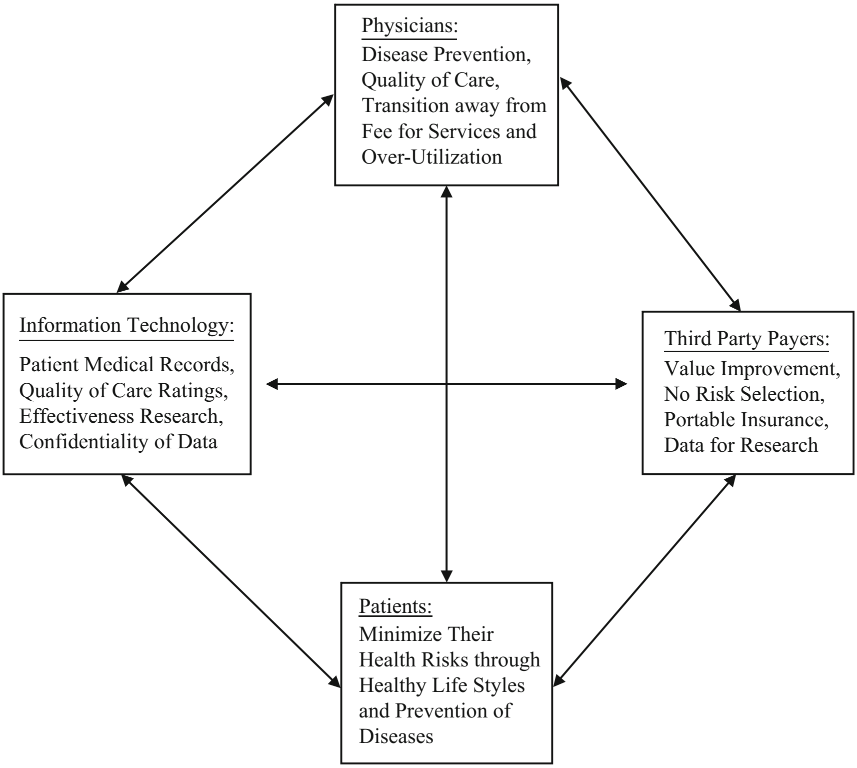
President Obama's Plan

President Obama's pragmatic plan has three aims (9 September 2009): to strengthen the private insurance of people who are presently covered, to provide new coverage for the uninsured, and to cut costs. Obama would strengthen employer-based insurance by ending risk selection on the basis of preexisting conditions, and premium discrimination on the basis of gender and age. He would prevent insurers from dropping coverage when people are sick, limit out-of-pocket expenses for episodes of care, and eliminate extra charges for such preventive care as mammograms, colonoscopies, flu shots, and tests for diabetes and other diseases. For senior citizens he would protect Medicare and eliminate the gap in coverage for prescription drugs. Children's health care, Medicare, and Medicaid are included in this plan, with the costs offset by federal and state governments, as done now. These reforms would stabilize and improve private and public insurance for those who are already covered.

However, when people lose jobs, many lose their health insurance. During the first year of the 2008–2009 recession an additional million people lost their health insurance so that by the end of 2009 about 47 million people were uninsured. The reform plan aims to solve the problem of uninsurance allowing dependent children to be eligible for coverage under their parents' plans and by creating state-based health insurance exchanges where individuals not covered by employer plans could choose from a portfolio that includes indemnity insurance and managed care networks offered by private insurers at competitive prices. There may be a basic plan that all insurance exchanges would offer, plus different levels of enhancements at additional costs. Tax credits for small businesses and for low- and middle-income people will enable them to purchase the insurance. Businesses over a certain size will be required to provide basic insurance for their employees, or to pay into a fund that will offset the costs of the basic plan. All people will be required to be covered by at least the basic plan. In this way healthy young people will be included in the risk pools, perhaps lowering the insurance costs for people with greater health risks. The exchanges will encourage low-cost, private-sector plans that would provide the uninsured with basic coverage. Until the exchanges are fully operational, there also will be low-cost insurance based on a national pool of high-risk people with preexisting conditions and ill health.[4]

Cost Savings

The Obama Administration believes that these reforms coupled with the elimination of wasteful spending, especially in Medicaid and Medicare, can pay for most of these new programs. Additionally, to cap and even to reverse the long-term health care spending curve, most probably the instituted reforms will build on the recommendations of a group of experts affiliated with the Engelberg Center for Health Care Reform at the Brookings Institution. They have focused on the critical problem of the disjunction between the costs of medical care and the results—the present system offers few incentives to improve quality and reduce overall costs. The Engelberg reforms (2009) would develop incentives for quality improvement and healthy lifestyles, new uses for information technology, and insurance that would reduce inappropriate utilization and its costs; the graphic of Fig. 13.2 ties together and depicts a number of their recommendations for cost savings.



Source: This new depiction is based on recommendations from a group of experts affiliated with the Brookings Institution (Engelberg Center for Health Care Reform, 2009).

Fig. 13.2 Four interrelated elements for controlling healthcare utilization, costs, and quality

The vertical two-headed arrow portrays the pivotal reciprocal relationship between physicians and patients. The Engelberg experts recommend that physicians focus more on the prevention of diseases, improve quality by better coordination of care, and shift away from fee-for-services payments toward bundled systems of reimbursement that encourage coordinated care and not overutilization of medical services. The patients, on the other hand, should be given incentives for improving their own health by minimizing their health risks through physical fitness, healthy lifestyles (no smoking, no obesity, no drug usage, and so forth), and should not pressure providers for care that is not medically necessary.

The horizontal two-headed arrow depicts the reciprocal relationship between information technology (IT) and the insurance companies and governmental agencies that are the third-party payers. Innovative IT would develop and maintain confidential medical records that only the group of physicians treating a patient could access. However, the database stripped of patient identifiers could be used to assess the quality and costs of episodes of care, and the effectiveness of alternative treatments. The third-party payers could use this information to guide their relationships with their managed care networks of providers. These payers would be prevented from selecting as enrollees only those applicants with low risks of illness and the absence of preexisting conditions. Their insurance policies would be portable, not permanently linked to the enrollee's employer.

The two-headed arrows along the periphery of the model imply other reciprocal relationships among the four elements. Patients enrolled in an insurance plan and the physicians treating them would emphasize healthy lifestyles and prevention of diseases. In return the insurers would focus on improving quality of care, provide hassle-free reimbursements for services, and offer open and portable insurance. The patients would make their medical records available to IT and the IT professionals would guarantee the confidentiality of the records. However, the patient's physicians could use the patient's medical record to coordinate care.

By stabilizing employer-provided insurance, reducing the number of uninsured Americans, creating incentives to reduce utilization and costs, and improving the coordination of medical care, the health care reforms of the Obama Administration and the Engelberg Center could reduce the costs and enhance the quality of the health care produced by the dysfunctional American system of recent years.

The Subsequent Chapters

Hoping to contribute to the amelioration of problems of health care even slightly, Chapter 14, "Gatekeepers and Sentinels" presents some findings that may support the use of precertification and concurrent review nurses for containing the inpatient expense of Medicaid and Medicare patients, and Chapter 15 summarizes the evidence about linkages between childhood vaccinations and unintended adverse consequences. The concluding chapter provides practice in assessing causal relationships; such practice may have some indirect benefits for resolving the health care issue.

Gatekeepers and Sentinels

Regardless of the reforms eventually implemented, to control for overutilization of care that lacks evidence of medical necessity, utilization and cost management strategies similar to those that Chapter 14 evaluates are still relevant.[5] This chapter presents meta-analytic consolidations of findings from evaluative research about the effects of nurses who monitored the use of medical care. These nurses threatened monetary sanctions against the providers of the care if supporting evidence indicated that the care was not medically necessary. In retrospect, their activities are consistent with the Engelberg Center's (2009, 3) recommendation to: "Reduce payments for care of low value relative to cost; for example, by reducing clearly inappropriate utilization and overpayments, as identified by the Medicare Payment Advisory Commission (MedPac)."

Two main types of programs encompass the specific studies. The first program entails precertification nurses, or "Gatekeepers," who assign an initial length of hospital stay and approve reimbursement. If the proposed hospital admissions lack evidence of medical necessity, then the nurses request additional evidence and, if that is not forthcoming, they can deny reimbursements for the health services they deem not medically necessary. The second program entails onsite nurses, or "Sentinels," who monitor the medical necessity of any hospital admissions that were not precertified, and approve or deny additional bed days. In essence, this chapter tests the effects of a two-step organizational control system that assesses cases prior to admission and also after admission.

Study design

The studies apply the basic difference-in-differences (DID) design developed earlier in Part 3. But here the unit of analysis is an inpatient stay, an episode of care; the outcomes are the length of stay, bed days, various measures of expense, and indicators of quality. A given patient has realized outcomes if assigned to the group being monitored by a nurse, and counterfactual outcomes if assigned to the comparison group; and conversely. Since a given patient cannot be assigned simultaneously to both the target and comparison groups, the fundamental problem of causal inference holds. Consequently, after pooling the data, the evaluations compare the change in time periods before and after intervention in the target group to the change the same time periods in the matched comparison group. The program effect coefficient equals the difference between these differences, which can be estimated by the coefficient on the interaction of the post-intervention time period and the target group.

Meta-analysis methods

For each set of studies and for a specified outcome, this consolidation first calculates the composite fixed effect of the treatment by taking a weighted average of the program effect coefficients, where a coefficient's weight is the reciprocal of its squared standard error divided by the sum of the reciprocals of the squared standard errors across the set of program effect coefficients. It tests the components of the composite effect for homogeneity and establishes a confidence interval. If the null hypothesis of homogeneity is not rejected, then this implies that the composite effect is an appropriate summary of its weighted components; that is, the variance between the estimates from the various studies is inconsequential. If the null hypothesis of homogeneity is rejected, then this implies that there is variance between the estimates from the various studies that needs to be taken into account. After calculating this random effect, a new composite measure is calculated with a wider confidence interval than the fixed-effect measure. This procedure can be carried out with an Excel program that is available at the Springer website.

Fixed and random effects

Chapter 14 attempts to clarify the distinction between fixed and random effects and, analogously and implicitly, between the average causal effect conditional on the covariates in the sample (CATE) and the population average treatment effect (PATE), as follows. Assume that there are a number of replicated evaluative studies focusing on the effects of a particular policy, say precertification of inpatient health care. The results of these studies form the database for a meta-analysis. The fixed-effects model ignores inter-study variation assuming that there is one common program effect for the data at hand. It assumes that there is one single true effect across the studies and that each study provides an estimate of that effect; the weighted average of these effects provides an estimate of the true effect. The fixed-effects model thus answers this question: "For the studies at hand, did the program on average produce and effect?" (Fleiss and Gross 1991, 132). Implicitly, this estimator is $\hat{\delta}_{\mathrm{CATE}}$ because the inference is limited to the sample of studies forming the database.

The random-effects model takes into account variation among the observations assuming that there is a distribution of true effects that characterizes a super-population. Moreover, each study has its own true effect that would result if it were replicated an infinite number of times. The range of inference is not limited only to those observations included in the database; the random-effects model aims to estimate the grand-mean treatment effect of all studies (Cook et al. 1992, 311; Hasselblad et al. 1995, 212–217). It provides an answer to this question: "Based on the studies at hand, will the program on average produce and effect?" (Fleiss and Gross 1991, 132). Implicitly, this estimator is $\hat{\delta}_{\mathrm{PATE}}$ because the inference is not limited to the sample of studies forming the database; the goal is inference from the sample at hand to parameters of the super-population.

Results

Chapter 14 uses the grammatical conventions implied above about the use of verb tenses to distinguish results based on fixed effects from results based on random effects: namely, the past tense indicates the specificity to time, place, and data of findings from fixed-effects analyses; the future tense indicates the generalizability of findings from random-effects analyses. Following this convention, for the universes of studies from which these studies are samples (with inferences limited by time periods and contexts), the analyses of random effects indicate that precertification nurses will reduce admissions, and that onsite concurrent review nurses will reduce bed days and inpatient ancillary expense. Furthermore, when the onsite nurses control the initial and concurrent assignment of hospital days, a random-effects analysis indicates that they will reduce length of stay. A fixed-effects analysis indicates that onsite review nurses reduced hospital admission rates.

At one hospital, surgical confinements that were precertified had higher rates of surgical complications; this may have been due to selection effects or to the effects of the program. The nurses' monitoring had no adverse effects on obstetrical complications or on the pooled rates of surgical and obstetrical complications. Since these programs may reduce inadvertent negative effects introduced by hospital care, these nurses may have indirectly enhanced the patients' well-being.

Implications for policy

President Obama hopes to fund his health care reforms in part by eliminating waste in Medicare and Medicaid, especially in the original fee-for-service Medicare plan that controls utilization through co-payments and limitations of coverage and physician payments through reimbursement schedules. About 78% of Medicare enrollees have this plan; the other 22% are enrolled in managed care plans provided by private insurers. Perhaps surprisingly, these managed care plans have not produced the savings that were promised initially (Kaiser Family Foundation, April 2009). Consequently, any savings extracted from Medicare will mostly likely come from the fee-for-service program. Because the precertification and onsite review nurses operated in the context of private fee-for-service plans, perhaps such nurses might be able to reduce excessive utilization in Medicare. But Medicare's experience with precertification for ten inpatient procedures during the period 1986–1990 indicated that it did not avert much inpatient care because the beneficiaries were older and frailer than the general population, and the evidence for the medical necessity of their care was persuasive. However, there could be a role for onsite concurrent review nurses if their responsibilities were expanded beyond gatekeeping and monitoring to include such services as patient advocacy, case management, and discharge planning. These suggestions await further empirical testing and assessment of the evidence for effectiveness and causality.

Childhood Vaccinations and Autism

The mass media and the Internet can provide timely information about health-related and other topics, but there is a risk that these media also can provide misinformation that has serious negative consequences. The controversy about childhood vaccinations and autism took off when Andrew J. Wakefield, the lead investigator of an exploratory study of questionable quality, participated in a press conference publicizing his speculative findings as if these provisional results were established scientific regularities. Ignoring previous research studies that showed that the combined measles–mumps–rubella (MMR) vaccination was safe, Wakefield opined that he could not support the use of the MMR vaccine until the issue of their safety is resolved.

As a consequence of his overstating the veracity of his findings, and the mass media's amplification of his warning about the MMR vaccine, in the UK and the USA many concerned parents decided not to vaccinate their children. Reinforcing the decisions of these parents, websites disseminated anecdotal information and misinformation about autism, along with advertisements about their products that promised to alleviate the symptoms of autism. Soon after this controversy erupted, vaccination rates tumbled in the UK from 92% to 80% at the peak of alarm—in the UK childhood vaccinations are at the discretion of the parents. Consequently, cases of measles in the UK increased from 56 in 1998 to about 1,348 in 2008, and several children died from this disease (Deer, February 8, 2009, e2).[6]

In the wake of Wakefield's report and the breakdown of immunization rates, a number of epidemiological studies found no persuasive evidence supporting causal effects of vaccinations on autism spectrum disorders (i.e., autism, atypical autism, and Asperger's syndrome). However, a number of studies claimed to uncover causal relationships. Some of the latter studies focused directly on the supposed adverse effects of the MMR vaccinations; other studies, by Mark Geier, M.D. and his son David Geier, primarily focused on thimerosal-containing vaccines (TCV) and their supposed adverse effects due to the very small amounts of mercury used as a preservative in the vaccines. [7] Because the results conflicted, it was appropriate to have experts review the quality of these studies, separating the wheat from the chaff; that is, solid scientific findings from those that are questionable.

Responding to the diversity of these findings, the Cochrane Collaboration in Europe and the Institute of Medicine (IOM) in the USA formed study groups composed of health scientists and biostatisticians. These groups aimed to assess the quality of the evidence for or against the putative linkages between vaccinations and adverse consequences. Chapter 15 explicates the methods these review groups applied to reach their separate conclusions of no causal effects on autism of vaccinations using the MMR vaccines or TCVs. The Cochrane group applied meta-analysis techniques to select their cases and to judge quality; they concluded that the evidence did not support a causal association between the MMR vaccinations and autism. The IOM group applied an informal Bayesian approach, taking a position of neutrality regarding the safety of the vaccines and their components. After establishing in advance their criteria for judging the quality of a study and the set of statements that would frame their conclusions, the committee reviewed the relevant studies identified by an exhaustive search of the literature. For both research questions the IOM concluded: "The evidence favors rejection of a causal relation."

Gauging Causality in Multilevel Models

The procedures of the Cochrane Collaboration and the IOM can guide assessments of the causality zone achieved by the various multilevel models of this book. Consequently, in the concluding chapter I first discuss the different notions of causality; namely, no causality, stable association, potential outcomes, and dependency networks. Then I present criteria for judging the quality of a social scientific study in terms of its validities, briefly: Fit validity concerns the adequacy of the conceptual scheme. Construct validity gauges the appropriateness of the measures of the concepts. Internal validity focuses on how well the study design operationalizes the research questions, and the robustness of the relationships among the measures of the concepts. External validity concerns the generality of the findings. Statistical conclusion validity entails the appropriateness of the statistical models applied to the study's data. I apply these criteria to judge the results of the multilevel modeling of the core chapters, summarizing the argument for causality and identifying the notion of causality that in my judgment the study fulfills. If a study does not possess at least the first three validities, then its evidence will indicate "no causality." I invite the reader to form his or her own judgments and to disagree with my assessments.

Endnotes

[1] Parents of children with autism presented three typical cases to special masters serving on the USA Court of Claims. To receive compensation the cases had to show a slight preponderance of the evidence that thimerosal-containing vaccines caused the child's autism and other adverse events. The court ruled that the petitioners' theories of causation were speculative and unpersuasive, and that scientific evidence did not support causal linkages between vaccinations and autism (Vedantam 2009).

[2] Few people now fall in the null category of rejecting that health care is a right and rejecting that it is a privilege. In the past, some physicians did not approve of any third-party payers and wanted to receive fee-for-services payments directly from the patient, similar to a typical business transaction. In the present, some people prefer the business model and others reject modern medicine altogether, preferring to let the wisdom of the body cure what ails them.

[3] In addition to reporting that opposition to reform was concentrated among higher-income voters and those over 65, Gelman et al. (2010, 11) summarize findings linking partisanship to attitudes about health care reform as follows:

> Thus a survey from March 2010 showed attitudes on health care also to be extremely partisan. In the aggregate, 46% support the proposed health-care reform and 48% opposed it, but Obama voters supported it by 81% to 11% while McCain voters opposed it by 90% to 7%. These numbers exactly mirror Obama approval among the same groups: 83% to 10% approval among Obama voters and 90% to 8% disapproval among McCain voters (Public Policy Polling, 2010), and represent a much higher level of partisan polarization than, for example, opinions about abortion or Iraq during the George W. Bush presidency.

[4] Bernard and Tara (2010) summarizes some of the features of the health care reform plan.

[5] Mays et al. (2004), W4, 429–430) report that health plans are reintroducing such “first wave managed care” innovations as prior authorization and concurrent review of inpatient care, strategies Chapter 14 develops and tests.

[6] Ten of the 13 authors of the controversial 1998 paper withdrew their interpretation that the MMR vaccinations may be causally linked to autism (Murch et al. 2004). Wakefield eventually left his position at the Royal Free and University College London and now offers colonoscopies to children in his clinic in Austin, Texas; some parents of autistic children respect his activities (Deer, 8 February 2009). On 2 February 2010 the editors of *The Lancet* formally retracted Wakefield et al. (1998) from the public record. On 23 May 2010 the media reported that Wakefield was banned from practicing medicine in the UK.

[7] Geier and Geier now advocate treating autistic children with the castration drug Lupron. For further information see their entry in the Wikipedia (downloaded October 3, 2009) and related websites about medical malpractice.

Chapter 14
Gatekeepers and Sentinels

> *More than four decades ago, this nation stood up for the principle that after a lifetime of hard work, our seniors should not be left to struggle with a pile of medical bills in their later years. . . . And that is why not a dollar of the Medicare trust fund will be used to pay for this plan. . . . The only thing this plan would eliminate is the hundreds of billions of dollars in waste and fraud, as well as unwarranted subsidies in Medicare that go to insurance companies. . . . Reducing the waste and inefficiency in Medicare and Medicaid will pay for most of this plan.*
>
> —President Barack Obama (September 9, 2009)

By applying a basic fixed- and random-effects paradigm of meta-analysis, this chapter summarizes findings from numerous evaluations of managed care programs in which nurses preauthorized and monitored inpatient care. Although the backlash against managed care circa 2000 reduced the use of such programs, the rising costs of medical care have led health plans to reconsider their use (Mays et al. 2004, W4: 429–431). So that Medicare can meet the future needs of the retiring baby boom generation, the National Academy of Social Insurance formed panels of experts to develop recommendations about reforming Medicare (Aaron and Reischauer 1995, 1998; Marmor and Oberlander 1998; Wilensky and Newhouse 1999). The panel on fee-for-service Medicare (January 1998) requested the identification of promising private-sector innovations for managing health care costs and quality that Medicare could apply (Fox 1997, 50; Miller and Luft 1997).

This chapter bears on the use of cost-management programs in Medicare's dominant fee-for-services plan. It presents a consolidation of findings from evaluations of private, cost-managed fee-for-service insurance conducted by researchers at a large health care insurance company. This consolidation assesses the effectiveness of a two-step control system designed to manage the utilization and costs of inpatient medical care. One component—precertification by nurses acting as gatekeepers—focuses on admissions to the hospital, prompting the attending physician to eliminate admissions that are not medically necessary. The second component—onsite concurrent review of inpatient care by nurses acting as sentinels—focuses on procedures and hospital days, prompting the attending physician to eliminate care that lacks medical necessity. The chapter concludes by discussing the relevance of the gatekeeper and

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_14, © Springer Science+Business Media B.V. 2011

sentinel effects for controlling the costs of the 78% of Medicare beneficiaries currently enrolled (circa 2010) in fee-for-service insurance (Kaiser Family Foundation 2009; Nyman et al. 1990, 127–137; Bailit et al. 1995; Kane and Friedman 1997; Davis et al. 1990, 223; White 1999).[1]

Program Characteristics

Utilization review (UR) programs that precertify admissions, or that concurrently review inpatient stays, penalize patients by denying reimbursements when their physicians or hospitals provide care that lacks evidence of medical necessity. To insure that actual admissions and lengths of stay are appropriate, program nurses apply accepted standards for the type of illness being treated—these are medically expected durations (Merton 1984, 274–275). The doctor's failure to obtain pre-approval for an admission through telephone conversations with precertification nurses or, in urgent care, concurrent approval, results in increased cost sharing for the patient. Based on the patient's medical record and their own observations, nurses conducting onsite concurrent review may question the necessity of a weekend or emergency admission that was not precertified, additional bed days, and charges for ancillary and other services. By discussions with physicians who review utilization, the attending physician may contest the decision of a nurse reviewer to deny payment for the admission or hospital bed day. However, what results have these programs achieved?

New Contributions

This consolidation assesses the unique effects of precertification and onsite review. Other investigators have assessed the conflated cost savings of these programs compared to no management of utilization; see Feldstein et al. (1988); Wickizer et al. (1989); Wickizer (1990); Wheeler and Wickizer (1990); Wickizer (1991); and Wickizer et al. (1991). Because only one core data set from one insurer was available, and because the programs were bundled together, these investigators could not distinguish the effects of the telephone-based precertification and review program from the effects of the onsite concurrent review program, as this consolidation does.

Preauthorization Review

Regarding precertification, the Congressional Budget Office cited a study by Khandker and Manning (1992, 52) as providing "the most convincing evidence to date of the impact of utilization review under fee-for-service insurance" (Langwell 1992, 11–12).[2] Compared to units with no management of utilization, units with the telephone-based precertification and review program experienced significant reductions in the following inter-related measures: total expense, by \$16.49 per insured employee (EE) per quarter ($t = -2.42$) or 4.4% (gross savings); inpatient expense, by \$19.63 per EE per quarter ($t = -3.27$) or 8.1% (gross savings); length of stay, by about half a day ($p = 0.0001$); and bed days, by 25.3 hospital days per 1,000 EEs per quarter ($t = -5.67$). The increase in outpatient expenditure per EE of \$3.13 per quarter was not significant ($t = 1.27$); neither was the quarterly reduction of 0.37 admissions per 1,000 EEs ($t = -0.63$).

To test whether the latter lack of a reduction in admissions is singular, this consolidation integrates this finding with those from seven other evaluations conducted by their research group. It compares units with precertification to units that do not manage utilization. Whenever possible, it also assesses the effects of precertification on rates of medical complications.

Concurrent Review

Regarding onsite concurrent review, in a controlled experiment in which Medicare beneficiaries and their physicians were unaware subjects, Restuccia (1983, 51, 55–62) found that when the rules of the program allowed nurse-coordinators discretion to directly inform attending physicians about questionable hospital days, savings in bed days and length of stay were greater than in a conventional program in which the rules limited their discretion about whom to contact and when. He referred to the program that allowed direct contact as "cybernetic" (Hage 1974) because it encouraged feedback from the nurses directly to the attending physicians, in contrast to a conventional program in which feedback was indirect, passing through physician reviewers before reaching attending physicians. This chapter compares the effectiveness of these two programs.

By comparing sites that use precertification and onsite concurrent review to sites that only use precertification, this chapter identifies the effects of onsite concurrent review. It advances past research by assessing the extent to which onsite concurrent review influences a comprehensive set of outcomes (admissions per 1,000 EEs, length of stay, bed days per 1,000 insured lives, expense for ancillary services and other inpatient expense, rates of medical complications, and cost savings). It reports the studies' consolidated effects and the effects nested within the cybernetic and conventional onsite programs. (EEs are defined as insured employees not including dependents; there are about 2.5 insured lives per EE.)

Research Methods

A meta-analysis aims to assemble and summarize the full literature on a topic, whereas a consolidation is more limited—it aims to summarize and make available to the research community the published and unpublished results of a specific program of research. Consolidations and meta-analyses thus differ in scope but not in method.

They test hypotheses about the effect of stimulus variables on response variables; they compare and contrast any diverse results; and, on the basis of the summarized findings, they may guide subsequent primary research, meta-analyses, or explanatory theorizing. By integrating the results of various consolidations, research synthesizers can form a comprehensive meta-analysis. (For a classic example of a consolidation in the form of an inventory of propositions see Festinger 1950; for a formal theory based on these findings see Simon 1957, 115–144; for a meta-analysis of the effects of educational reform see Borman et al. 2003.)

Evaluators at a private insurance company conducted quasi-experimental studies of precertification and onsite concurrent review for the period 1986–1990. This chapter consolidates all of their published and unpublished reports on these topics. The research group critiqued the unpublished reports prior to their distribution. In their explication of current problems of meta-analysis, Hasselblad et al. urge researchers not to restrict meta-analysis by including only randomized experimental studies; they state (1995, 229): “Policy makers cannot afford to ignore the vast majority of information available today merely because it does not fit the simple model of classical meta-analysis.”[3] Since unpublished studies may have weaker program effects than published studies, including them in a meta-analysis or in a consolidation may make it more difficult to establish that an effect exists (Wolf 1986, 37–39). But unpublished studies, if properly done, should be included in a research synthesis (Cook et al. 1992, 292–293; D’Agostino and Kwan 1995, AS, 102; Cook et al. 1993), especially in consolidations that aim to make available to the research community findings that are rare and that facilitate the estimation of random effects—results that generalize to the universe of such studies.

Study Characteristics

Precertification

The eight studies of precertification are described in Table 14.1 in chronological order, spanning the period April, 1986 through November, 1990. They assess the medical claims of at least 1,250,000 privately-insured EEs in the comparison groups and 285,000 in the target groups; Medicare beneficiaries are excluded. The studies are comparable, combinable, and independent (Cook et al. 1992, 315–316), as explicated next. To provide a common metric, this chapter annualizes their findings so that the time periods before and after implementation of the target program each have duration of one year. Hereafter, for brevity, it refers to the time period prior to program implementation as the pre-period, and the time period of program implementation as the post-period.

The research designs of the eight studies are similar. The typical study combines a cross-sectional comparison with measures of change across time. The group that will undergo the target program is matched with a contemporaneous comparison group that will not receive the program. During the pre-period neither group is exposed to

Table 14.1 Characteristics of precertification studies

	Study P-1	Study P-2	Study P-3	Study P-4	Study P-5	Study P-6	Study P-7	Study P-8
Location of program	Dispersed	Southern California	National USA	Boston-area	Midwest West Coast	Texas dispersed	National USA	Alaska Northwest
Study duration	2 years	21 months[a]	2 years[a]	18 months[a]	2 years	3 years[a]	3 years[a]	2 years
Pre time period	4/86–3/87	1/87–12/87	1/87–12/87	4/87–12/87	1/87–3/88	1/87–12/88	1/87–12/89	12/88–11/89
Post time period	4/87–3/88	1/88–9/88	4/87 to 12/88	4/88–12/88	1/88–3/89	1/89–12/89	4/87 to 3/90	12/89–11/90
Private insurance	No medicare	No medicare	No medicare	No medicare	No medicare	No medicare	No medicare	No medicare
Unit of analysis	Inpatient stays	Group accounts	Group accounts	Inpatient stays	Inpatient stays	Inpatient stays	Group accounts	Inpatient stays
Number of units								
Target pre	1,875	838	828	419	719	7,512	2,846	1,731
Target post	1,692	861	828	460	883	7,288	2,846	1,593
Comparison pre	15,723	112	4,381	15,105	72,295	60,885	2,459	1,174
Comparison post	15,194	136	4,381	13,510	67,108	51,548	2,459	1,182
Enrollees (EEs)								
Target pre	7,441	3,880	176,364	2,348	4,133	43,990	254,896	8,469
Target post	7,325	3,957		2,365	5,669	42,418		8,704
Comparison pre	87,352	653	468,767	88,526	343,141	321,491	327,047	10,084
Comparison post	87,824	721		88,132	331,671	292,342		10,646
Control variables	Design variables	Design variables	Design vars. & 15 predictors	Design variables	Design variables	Design variables	Design vars. & 15 predictors	Design variables
Report authors	R. B. Smith, J. Clive, and H. M. Allen, Jr.	R. Khandker, W. Manning, and H. M. Allen, Jr.	R. Khandker and W. Manning	R. B. Smith, J. Clive, and H. M. Allen, Jr.	R. B. Smith	R. B. Smith	R. Khandker, W. Manning, and T. Ahmed	R. B. Smith
Report Date	21-Mar-89 (Rev. 12-May-89)	9-May-89	14-Feb-90 Pub. 1992	5-May-89	14-Dec-89	1-Jun-90	12-Dec-90 Pub. 1992	24-Apr-91

[a] Data and rates have been annualized.

the target program. During the post-period, the new program is implemented only in the target group. At the end of the post-period, for each of the two groups, the pre-to-post difference on an outcome is assessed. The difference between these two differences defines the effect of the program. Statistical methods are used to control for demographic and other differences between these groups.

The units of analysis in the studies are different, but this difference has no effect on admissions. In five studies (P-1, P-4, P-5, P-6, and P-8), the evaluators analyzed inpatient confinements. For admissions rates, they only controlled for the effects of the design variables—the pre- versus post-period, the cross-sectional difference between the target and comparison groups, and the program effect indicator. In three other studies (P-2, P-3, and P-7), for each quarterly time period other evaluators aggregated claims to the group-account level and then, using statistical methods to control for case-mix and other variables, they analyzed the aggregate rates. Because of the complexity of its study design, for simplicity, this consolidation uses the effect of P-2 with controls only for the design variables.

To test whether the different units of analysis—aggregated claims versus inpatient confinements—produced different effects on admissions, this chapter compares the consolidated effect of the three studies based on aggregated claims to the consolidated effect of the five studies based on inpatient confinements. In a test of the null hypothesis (H_0 = no difference between the mean reductions in admissions due to the different units of analysis) versus the alternative hypothesis (H_a = the different units of analysis produce results that are different), the χ^2 test statistics are such that H_0 can not be rejected. When this test is based on the difference between the combined fixed-effects estimates for the studies with the different units of analysis, the homogeneity $\chi^2 = 2.38$ ($0.25 > p > 0.10$); when the test is based on the difference between the combined random-effects estimates, the homogeneity $\chi^2 = 3.68 \times$ E-05 ($p > 0.90$). Thus, regarding admission rates, the two different units of analysis have no effect on the results.

The assumption of independent studies is appropriate: Each study has a different time period; this may imply that the implementation of each program represents a different sample from the distribution of true effects. The study sites are scattered throughout the United States; this may imply that the patients and physicians at one site are different from those at other sites. Although P-3 and P-7 share some common data elements (especially in the comparison group), the post-period is much shorter in P-3 than in P-7, P-7 has three authors (not two), and the studies report significantly different effect sizes. In a test of the null hypothesis (H_0 = no difference between the mean reductions in admission rates of the two populations) versus the alternative hypothesis (H_a = the population means are different), the two-sample t test decisively rejects the null ($t = 4.31, p < .001$). The parameters of studies P-3 and P-7 are at least very different, if not totally independent (Moore and McCabe 1989, 538–555). Since the assumption of independence covers all of the other studies of precertification, to avoid possible bias due to non-independence between P-3 and P-7, this consolidation calculates the summary effects first with P-3 included (which are the preferred estimates), and then with that study excluded (see the relevant endnotes).

Onsite concurrent review

The four studies of onsite concurrent review are described in Table 14.2 in chronological order, spanning the period January 1986 through November 1990. The studies cover the claims experience of at least 37,451 privately-insured EEs in the comparison group and 14,832 in the target group; any Medicare beneficiaries are excluded. These studies are comparable, combinable, and independent. Although the study periods vary (17 months, two years, and three years), the annualized results provide common metrics. All four studies apply the same difference-in-differences (DID) study design taking the difference between the difference between the pre to post means in the target group and the difference between the pre to post means in the

Table 14.2 Characteristics of studies of onsite concurrent review

	Study O-1	Study O-2	Study O-3	Study O-4
Location of program	Southwestern Virginia	Central Ohio	Greater Miami, Florida	Anchorage, Alaska
Study duration	17 months[a]	3 years[a]	2 years	2 years
Pre time period	1/87–9/87	1/86–6/87	9/87–8/88	12/88–11/89
Post time period	10/87–5/88	7/87–12/88	9/88–8/89	12/89–11/90
Private insurance	No medicare	No medicare	No medicare	No medicare
Precertification nurse	Only certifies admission	Certifies admission and assigns LOS	Only certifies admission	Only certifies admission
Onsite review nurse	Certifies each day, direct feedback	Certifies additional days, indirect feedback	Certifies each day, direct feedback	Certifies each day, direct feedback
Number of admissions				
Target pre	1,336	2,426	957	892
Target post	700	1,936	853	845
Comparison pre	2,149	9,090	1,754	1,731
Comparison post	921	7,507	1,745	1,593
Enrollees (EEs)				
Target pre	Not available	4,853	5,379	4,600
Target post	Not available	4,581	5,487	4,728
Comparison pre	Not available	19,546	9,436	8,469
Comparison post	Not available	16,933	9,625	8,704
Control variables in regressions	Design variables demographics, emergencies, intensity of service, and case mix.	Design variables selection effects, geographic area, seasonality, demographics, case-mix, and benefit plan.	Design variables demographics, geographic area, seasonality, hospital type, and case-mix.	Design variables admission day, geographic area, seasonality, demographics, case-mix, and illness intenisty.
Complications studied	No	Yes	Yes	No
Report authors	R. B. Smith	R. B. Smith and T. D. Gotowka	R. B. Smith	R. B. Smith
Report date	2-Nov-88	1-Dec-89 Pub. 1991	7-Mar-90	12-Apr-91

[a] Data and rates have been annualized.

comparison group, and then the difference between those differences. The target group was exposed to both precertification and onsite review while the comparison was exposed only precertification. All four studies utilize inpatient hospital confinements as the unit of analysis in the regressions, and form rates per EE using the enrollment census. Because of labor unrest, enrollment data for O-1 are not available; consequently, this site provides only per-confinement measures of utilization and cost. The effects on length of stay, bed days, and expense are derived from regression analyses in which the design variables, patient demographics, case-mix, geographic area, and other variables are controlled. The effects on admissions are derived from analyses that control only for the design variables.

It is appropriate to assume that these studies are independent. Their time periods differ; this may imply that the implementation of each program represents a different sample from the distribution of true effects. The sites are geographically dispersed, with target and comparison groups in southwestern Virginia (O-1), central Ohio (O-2), greater Miami, Florida (O-3), and Anchorage, Alaska (O-4); this implies that the patients, physicians, and onsite nurses are different. Each study has unique employer-based target and comparison groups; this implies that the studies do not share any common data elements.

The three cybernetic programs (O-1, O-3, and O-4) allowed direct feedback between the nurses and the attending physicians. In these programs the precertification nurses only certified the admission and the onsite nurses had discretion to assign all subsequent hospital days, and to negotiate directly with the attending physicians about the appropriateness of care. Since these onsite nurses more directly controlled the duration of the stay, this consolidation assesses the extent to which these cybernetic programs reduced utilization and expense, compared with the conventional program.

The conventional program (O-2) required the precertification nurse to certify the admission and to assign the initial length of stay. This procedure limited the discretion of the onsite nurse to shorten the duration of the hospital stay. When the onsite nurse's protocols indicated that certification should have been denied, the nurse would contact a physician reviewer, who would then contact the attending physician if the denial was warranted. In many cases this process delayed the issuance of a denial; retrospective review was often necessary (Smith and Gotowka 1991, 83).

Medical complications

The providers' rates of surgical and obstetrical complications reflect the complexity of the patients' medical problems and the appropriateness of the care they receive. To determine how the precertification and onsite review programs influenced rates of complications, this consolidation assesses data from a published study (O-2) and an unpublished study (O-3) of onsite concurrent review. Although the data are sparse, they are sufficient to estimate roughly the effects of both programs on rates of complications.

The codes of the International Classification of Diseases, Clinical Modification, 9th revision (ICD-9-CM), may indicate complications. Surgical complications may be indicated by the presence on a claim of codes for infections, intraoperative errors, hemorrhage, wound dehiscence, fistulae, and complications related to the cardiac, peripheral vascular, respiratory, gastrointestinal, and urinary systems. Obstetrical complications may be indicated by the presence on a claim of codes for eclampsia, puerperal infection, obstetrical trauma, hemorrhage, surgical wound dehiscence, retained placenta, neonatel infection, birth trauma, and complications from anesthesia, surgery, or abortion.

Statistics

When the available data allow, this consolidation estimates the programs' fixed and random effects (following DerSimonian and Laird 1986; Fleiss 1981, chapters 9 and 10; Fleiss and Gross 1991; Hasselblad et al. 1995, 203, 212–215; DuMouchel and Waternaux 1992). The fixed-effects model ignores interstudy variation. It assumes that there is one common program effect and that each study provides an estimate of that effect—there is a single true effect of the treatment across the studies (Cook et al. 1992, 311). It provides an answer to the question: "In the studies at hand, *did* the program on average produce an effect?" (Fleiss and Gross 1991, 132).

The random-effects model quantifies the inter-study variation, assuming that there is a distribution of true effects—each study has its own unique true effect that would result if it were replicated an infinite number of times (Cook et al. 1992, 311). Because the random-effects model aims to estimate the grand-mean treatment effect of all studies, the range of inference is not limited to only those studies included in the consolidation (Hasselblad et al. 1995, 212–217). It thus provides an answer to the question: "Based on the studies at hand, *will* the program on average produce an effect?" (Fleiss and Gross 1991, 131–132).

Whether or not a study utilizes the 2 by 2 design (pre-post vs. target-comparison), when a regression coefficient and its standard error (hereafter, SE) define the study's program effect, these statistics are used in the calculations of the fixed and random effects. When these statistics are unavailable, then, based on the 2 by 2 design, the effect size is calculated as the difference between the pre-to-post difference in means of the comparison group on the response variable, and the pre-to-post difference in means of the target group on that response variable. The SE of this difference is estimated by the square root of the sum of the variance estimates of the two differences (Fleiss 1981, 29–30). Given the program effect coefficients and their SEs, a spreadsheet program calculates the estimates of fixed and random effects.

Fixed effects

Following DerSimonian and Laird (1986, 180–182), let d_i denote the estimated program effect from study i, $i = 1,\ldots, K$, and let s_i denote its estimated SE. The fixed-effects model then assumes that

$$\mathrm{d}_i \sim iid\, N(\mu, \sigma_i^2)$$

where μ is the true program effect and σ_i is the true SE of d_i. The model assumes that the studies are similar enough in design and subjects so that the program effect is constant across studies ($\delta_i = \mu$).

The common program effect μ is estimated by a weighted average of the individual effect estimates $d_{i,}$

$$\hat{\mu} = \sum_{i=1}^{K} W_i d_i$$

The weight W_i assigned to d_i is $W_i = \sigma_i^{-2} / \sum_{i=1}^{K} \sigma_i^{-2}$ where σ_i^2 is estimated by s_i^2. An estimate of the SE of $\hat{\mu}$ is provided by $SE(\hat{\mu}) = 1/(\sum_i \sigma_i^{-2})^{1/2}$.

The assumption of homogeneity of treatment effect across studies can be tested by referring the calculated value of:

$$Q = \sum_{i=1}^{K} (d_i - \hat{\mu})^2 / \sigma_i^2$$

to a χ^2_{K-1} distribution. To reject the assumption of homogeneity, the critical value of the test is 0.05; to accept this assumption, the critical value is about 0.5—the null model must fit very well (Hasselblad et al. 1995, 213). If the null (of homogeneous effect) is not accepted, then the random effects are calculated.

Random effects

By using the following two-stage assumption (DerSimonian and Laird 1986, 180–183), the random-effects model allows the treatment effects to vary across studies:

$$d_i = \delta_i + \varepsilon_i, \quad \varepsilon_i \sim iid\, N(0, \sigma_i^2) \tag{1}$$

$$\delta_i = \mu + u_i, \quad u_i \sim iid\, N(0, \tau^2) \tag{2}$$

This model thus relaxes the assumption of a common treatment effect by assuming that the individual treatment effects δ_i follow a normal distribution with mean μ and variance τ^2. The parameter μ is an average treatment effect and τ^2 is a measure of the variability in the individual treatment effects δ_i's across studies. The model implies that $\mathrm{var}(d_i) = \sigma_i^2 + \tau^2$.

The average treatment effect μ is estimated by

$$\hat{\mu} = \sum_{i=1}^{K} W_i d_i$$

where the weights W_i now are

$$W_i = \frac{\left(\tau^2 + \sigma_i^2\right)^{-1}}{\sum_j \left(\tau^2 + \sigma_j^2\right)^{-1}}$$

The SE of $\hat{\mu}$ is

$$SE(\hat{\mu}) = \frac{1}{\left[\sum_j \left(\tau^2 + \sigma_j^2\right)^{-1}\right]^{1/2}}$$

where σ_i^2 is estimated by s_i^2.

An estimate of the variance component τ^2 based on the method of moments (DerSimonian and Laird 1986, 183) is,

$$\hat{\tau}^2 = max\left|0, \frac{Q-(K-1)}{\sum \sigma_i^{-2} - \{\sum \sigma_i^{-4} / \sum \sigma_i^{-2}\}}\right|$$

where Q is the homogeneous component of the total χ^2 as defined earlier (Fleiss 1981, 163). Clearly, if Q is less than or equal to the $(K - 1)$ degrees of freedom (hereafter *df*), then the random-effects model reduces to the fixed-effects model, and the random-effects estimates are the same as the fixed-effects estimates.

Poisson regression

To summarize the change in bed-day rates for three studies of onsite review (O-2, O-3, and O-4), this consolidation extends the fixed- and random-effects paradigm to include Poisson regression analysis. For each site a 2 by 2 design table (pre-post vs. target-comparison) organizes the counts of bed days and lives. TargetPost indicates the absence (0) or presence (1) of the program effect. The Eq. (3) below are evaluated (for the target group in the post-period), when TargetPost is 0 and when it is 1 (Agresti 1996, 80, 86–87).

$$\begin{aligned} \log\text{ (bed days/lives)} &= \alpha + \beta_1\text{Post} + \beta_2\text{Target} + \beta_3\text{TargetPost} \\ \log\text{ (bed days)} - \log\text{ (lives)} &= \alpha + \beta_1\text{Post} + \beta_2\text{Target} + \beta_3\text{TargetPost} \qquad (3) \\ \text{bed days} &= \text{lives } \times \ e^{(\alpha+\beta_1\text{Post}+\beta_2\text{Target})} \times \left(e^{\beta_3}\right)^{\text{TargetPost}} \end{aligned}$$

In the last equation when TargetPost = 0, then $e^{\beta_3} = 1$; that is, $e^0 = 1$. If Post = 1, Target = 1, and TargetPost = 0, then the predicted bed days at a target site when the onsite program is absent = lives × $e^{(\alpha\ +\ \beta1\ +\ \beta2)}$. When the onsite program is present, TargetPost is 1, and e^{β_3} will be a multiplicative factor (usually less than unity). The predicted bed days at a site when the onsite program is present equals the product of the predicted bed days at that site when the program is absent multiplied by this factor. When the difference between these two counts of bed days is divided by the number of insured lives in the target group in the post-period, the change in the bed-day rate per 1,000 lives is obtained. To obtain the upper and lower bounds of this change, the upper and lower bounds of the 1.96 confidence interval (CI) for β_3 are exponentiated and used in these calculations.

The summary fixed- and random-effects estimates, and their upper and lower bounds, are obtained from the usual spreadsheet calculations. Exponentiation of these coefficients produces factors that adjust the counts of bed days. To produce the predicted bed days when onsite review is absent, the actual number of bed days when onsite review is present (the count in the target group in the post-period) is divided by the factor. To obtain the change in the bed-day rate, the difference between the predicted bed-day count when onsite review is absent and the actual bed-day count when onsite review is present divided by the number of lives in the target group in the post-period. This is done for the fixed-effect estimates, the random-effects estimates, and their upper and lower values.

Counternull effect size

When there is a finding of no statistically significant effect, to stress that the same evidence could also support a finding of some effect, Rosenthal and Rubin (1994, 329–330) have proposed the counternull value of the effect size (ES). They define this quantity as that nonnull magnitude of effect size that is supported by exactly the same amount of evidence as is the null value of the effect size. Whether an effect size is statistically significant, for symmetric reference distributions (the normal and *t* distributions), the counternull value of the effect size (hereafter, the counternull) is the difference between twice the obtained effect size minus the null effect size (that is, $\mathrm{ES}_{\mathrm{counternull}} = 2\mathrm{ES}_{\mathrm{obtained}} - \mathrm{ES}_{\mathrm{null})}$. In the subsequent analyses, since the null value is zero, the counternull is twice the value of the obtained effect. For a full explication of this statistic see Rosenthal et al. (2000, 13–16).

Logit and log-linear models

The original research reports dichotomized medical complications as confinements that had some complications versus confinements with no complications. Consequently, this consolidation analyzes these data using statistical methods appropriate for dichotomous responses: logistic regression and log-linear modeling. The former assumes that the programs have asymmetric effects on complications; the latter assumes a symmetric effect, and solves for the best-fitting model. If the precertification and onsite review programs did not increase complications, then this null finding would indicate that these programs did not compromise the appropriateness of the medical care. A significant increase could be an effect of the program or it could indicate selection bias—the programs may have screened out the less severe cases. Because the data are so sparse, the random effects cannot be calculated.

Savings

Total inpatient net submitted expense includes bona fide charges for hospital expense (room-and-board and ancillary expense), physician services, and anesthesia. Since the onsite programs aimed to save the policyholder money, their monetary savings in total inpatient net submitted expense is a relevant evaluative criterion. The estimates of total inpatient savings per confinement are obtained from each study's program effect, the regression coefficient, and its SE. These results are combined by applying the fixed-effect paradigm to obtain the summary per-confinement effect and its SE. (In this particular application, the homogeneity indicated by the small value of the Q χ^2 statistic is such that the random effects reduce to the fixed effects.) For each study and for the summary coefficient, to convert the unit of analysis from per confinement to per EE, the per-confinement effect, its upper bound, and its lower bound are multiplied by the admissions rate (admissions/EEs) for the target group in the post-period. To obtain the net savings per EE, program administration costs are subtracted from the gross savings per EE. These costs are (per EE per year) for Study O-2 = \$27.46, O-3 = \$37.20, O-4 = \$29.60, and \$31.72 overall (which is the average, weighted by the number of EEs in these onsite groups in the post-period).

To calculate the percent savings, this consolidation uses as the denominator the cell value for the comparison group in the pre implementation time period. This quantity represents the maximum possible reduction in expense. In the calculations of the percent savings for the summary statistics, the base value of \$1,616.09 per EE is used; this figure results from the pooling of the data for each individual study. For the individual studies, the predicted cell values based on the regression analyses are used. These are: Study O-2 = \$1,865.60, O-3 = \$1,585, and O-4 = \$1,076.05.

Results

This consolidation follows established conventions for reporting the results of a meta-analysis. For each response variable and for the savings, it plots each study's effect and its 95% CI. It also plots the summary fixed- and random-effects estimates and their 95% CI (Hasselblad et al. 1995, 204–05). Below, after an effect size is reported, the CI is enclosed in parentheses. Since the authors of the studies conceptualized admissions rates, length of stay, and expense as normally distributed continuous variables (Manning 1998; Manning and Mullahy 1999); this consolidation does also.

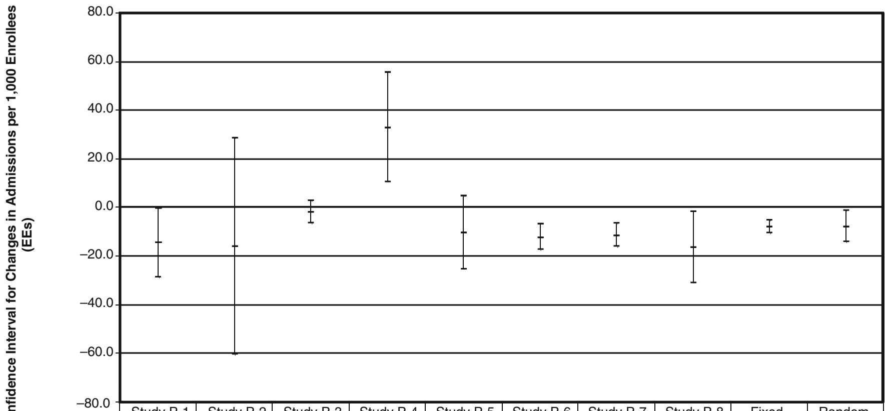
The fixed-effects model assumes constancy of treatment across the studies. The random-effects model assumes that each study has its own unique true effect, which is a realization of a normally distributed random variable.

Effects of Precertification

Admissions

Figure 14.1, which portrays the effects of all eight studies, indicates that P-4 is an obvious outlier—it has a gain in admissions whereas all of the other studies have reductions. This outlying site is a computer hardware company near Boston, which, at the time of the study, was experiencing extensive organizational change. These reorganizations may have caused its high utilization rates. The pre- and post-periods were only nine months long, which, given start-up problems, may have been too short for the program to have an effect. Also, the comparison group was national in scope and may have had lower rates than a closely matched group in the Boston-area (Wennberg et al. 1987). Its extreme variation may signal a study of low quality, a study whose effects may be discounted.

When this anomalous effect is removed, the fixed-effect reduction is −8.2 (−10.9, −5.5) admissions per 1,000 EEs (SE = 1.37), but it is composed of non-homogeneous effects. P-3 now contributes 8.13 to the Q χ^2 of 12.93 ($df = 6, p < .05$), and the random effects must be calculated. That reduction is −9.3 (−14.2, −4.5)



	Study P-1	Study P-2	Study P-3	Study P-4	Study P-5	Study P-6	Study P-7	Study P-8	Fixed	Random
Upper Bound	0.3	28.9	3.1	55.9	5.2	−6.3	−6.1	−1.3	−4.9	−0.8
- Effect	−14.0	−15.6	−1.5	33.4	−9.9	−11.7	−11.0	−16.0	−7.6	−7.3
Lower Bound	−28.3	−60.1	−6.1	10.9	−24.9	−17.1	−15.8	−30.6	−10.3	−13.8

Upper Bound - Effect Lower Bound

Fig. 14.1 Fixed- and random-effects estimates for admissions, eight studies of precertification

admissions per 1,000 EEs (SE = 2.46).[4] The counternull reductions are, respectively, −16.4 and −18.6 admissions per 1,000 EEs.

The meta-analytic literature, however, suggests that this consolidation should include the anomalous study. Because researchers prior to the consolidation did not independently assess the quality of all eight studies, and attach to each a consensual, objective quality score, the outlier cannot be discounted now on the basis of its low quality (Fleiss and Gross 1991, 137). Moreover, the fixed- and random-effects paradigm allows the heterogeneity of treatment effects across observations to be incorporated in the analysis (DerSimonian and Laird 1986, 187). The distribution of true effects may include the effects of sites in which there may be start-up problems, inadequate comparison groups, and unfavorable program effects. Since this paradigm weights treatment effects by the inverse of their squared SE, an effect size that has a high SE is weighted less than the same effect size that has a low SE. The removal of a study because the treatment effect is unfavorable would bias the summary estimates. Fleiss and Gross (1991, 130) are adamant: “It is invalid to delete from the set of studies to be meta-analyzed those whose results are in the ‘wrong direction,’ for the opportunity for bias in identifying the ‘deviant’ studies is too great.” Alternatively, the summary can be updated by incorporating new information from other studies. For these reasons, the preferred estimates are based on all eight studies, even though they indicate a smaller, but favorable effect of precertification on admissions.

For the fixed-effects model, Fig. 14.1 reports that the eight groups with precertification, experienced a statistically significant reduction of −7.6 (−10.3, −4.9) admissions per 1,000 EEs (SE = 1.36). The counternull is −15.2 admissions per 1,000 EEs. When the homogeneity of the effects is tested, the statistics confirm that P-4 is an outlier—it contributes 12.75 to the Q χ^2 of 25.86. As expected, the assumption of homogeneity is rejected (DF = 7, $p < .001$) and the random effects must be calculated. That reduction is −7.3 (−13.8, −0.8) admissions per 1,000 EEs (SE = 3.3). The counternull of −14.6 admissions per 1,000 EEs roughly equals the fixed-effect counternull. Thus, precertification will reduce admissions, even when the consolidation includes the outlier.[5]

Medical complications

Using data from O-3 for the period January 1985 to September 1988, precertified confinements in Florida hospitals can be compared to similar confinements that were not precertified. Precertified surgical confinements had a higher risk of complications than those that were not precertified; Panel (A) of Table 14.3 reports that the odds ratio equals 1.87 ($p = 0.042$). This ratio expresses the odds (0.01236) of a surgical complication in metropolitan Miami hospitals when a hospital confinement was precertified to the odds (0.00661) of a surgical complication in those hospitals when a confinement was not precertified. After precertification was introduced, confinements that had surgical complications increased from 6.57 to about 12.21 per

Table 14.3 Counts of medical complications by precertification status and location of hospital in Florida, January 1985 to September 1988

Panel (A), surgical complications	Hospitals in metropolitan Miami		Hospitals elsewhere in Florida	
	86,244–88,244	85,001–86,243	86,244–88,244	85,001–86,243
Time period	Precertification	No precertification	Precertification	No precertification
Surgical complications				
Some	13	6	19	10
None	1,019	967	1,731	1,604
Total confinements	1,032	973	1,750	1,614

Effects of precertification: $\beta = 0.626$, SE = 0.307, Exp (β) = 1.87, $p = 0.042$

Miami metro hospitals: $\beta = 0.099$, SE = 0.297, Exp (β) = 1.104, $p = 0.739$

Constant: $\beta = -5.118$, SE = 0.0277, $p < .0001$

Best log-linear model: Precert × Surgical Complications, Miami hospitals; $\chi^2 = 0.323$, DF = 3, $p = 0.956$

Panel (B), obstetrical complications	Hospitals in metropolitan Miami		Hospitals elsewhere in Florida	
	86,244–88,244	85,001–86,243	86,244–88,244	85,001–86,243
Time period	Precertification	No precertification	Precertification	No precertification
Obstetrical complications				
Some	77	69	102	86
None	349	312	420	311
Total confinements	426	381	522	397

Effects of precertification: $\beta = -0.073$, SE = 0.122, Exp (β) = 0.930, $p = 0.552$

Miami metro hospitals: $\beta = -0.155$, SE = 0.123, Exp (β) = 0.856, $p = 0.207$

Constant: $\beta = -1.317$, SE = 0.107, $p < .0001$

Best log-linear model: Obstetrical complications, Miami hospitals, Precert; $\chi^2 = 4.96$, DF = 4, $p = 0.29$

Panel (C), surgical and/or obstetrical	Hospitals in metropolitan Miami		Hospitals elsewhere in Florida	
	86,244–88,244	85,001–86,243	86,244–88,244	85,001–86,243
Time period	Precertification	No precertification	Precertification	No precertification
Total complications				
Some	95	84	130	102
None	1,038	966	1,694	1,577
Total confinements	1,133	1,050	1,824	1,679

Effects of precertification: $\beta = 0.119$, SE = 0.103, Exp (β) = 1.126, $p = 0.247$

Miami metro hospitals: $\beta = 0.231$, SE = 0.104, Exp (β) = 1.260, $p = 0.026$

Constant: $\beta = -2.71$, SE = 0.088, $p < .0001$

Best log-linear model: Miami Hospitals × Total Complications, Precert; $\chi^2 = 1.69$, DF = 3, $p = 0.64$

Best log-linear model across Panel (A) and Panel (B): Surgeries × Precert × Complications, Surgeries × Miami Hospitals; $\chi^2 = 4.99$, DF = 6, $p = 0.55$

Partial Associations: Surgeries × Precert × Complications; $\chi^2 = 4.74$, $p = 0.029$. Precert × Complications; $\chi^2 = 0.093$, $p = 0.761$

Data are from Study 3 of Onsite Concurrent Review.

1,000 surgical confinements. The best-fitting log-linear model ($p = 0.96$) corroborates this effect. It is composed of the two-variable interaction, precertification × surgical complications, and the marginal effect of metropolitan Miami hospitals.

Because precertification may screen out the less complicated cases, this increased rate of surgical complications does not necessarily imply that precertification produced a decline in the appropriateness of care. Panel (B) of Table 14.3 indicates that precertification had no effect on obstetrical complications in these hospitals, and Panel (C) indicates that precertification had no effect on the pooled rates of obstetrical and surgical complications. However, the best fitting log-linear model across Panel (A) and Panel (B) does include the three-variable interaction, surgical confinements × precertification × complications ($p = .029$), but not the two-variable interaction, precertification × complications ($p = .775$).

To summarize, the random-effects analysis indicates that precertification will reduce admissions. A fixed-effects analysis suggests that precertification was associated with increased rates of surgical complications, but not with increased rates of obstetrical complications, and not with increased rates of both surgical and obstetrical complications.

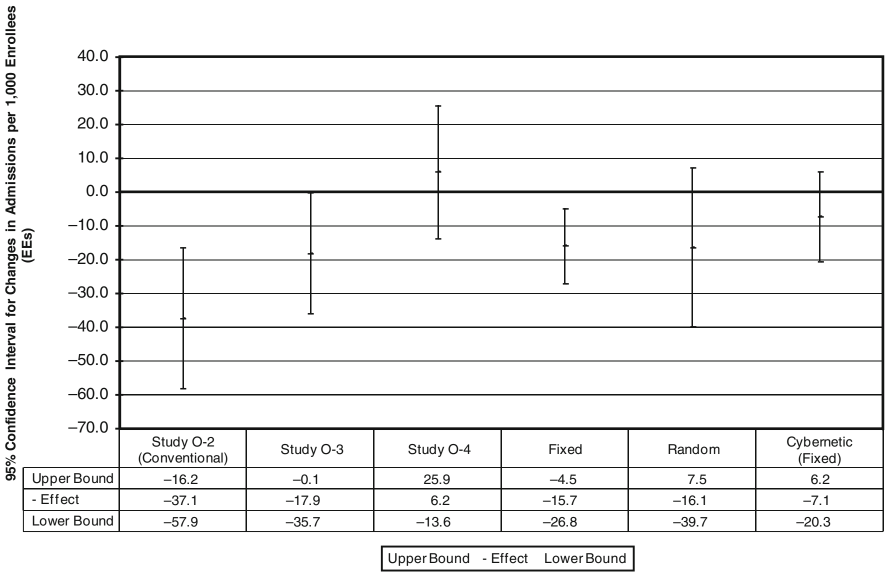
Effects of Onsite Concurrent Review

The subsequent analyses estimate the overall consolidated effects of the onsite programs and the stratified effects of the conventional and cybernetic programs.

Admissions

The fixed-effect model, but not the random-effects model, suggests that the onsite nurses significantly reduced the admissions rates; see Fig. 14.2. When the results from three sites (O2, O3, O4) are combined, the fixed-effect reduction is −15.7 (−26.8, −4.5) admissions per 1,000 EEs (SE = 5.7). The counternull is −31.4 admissions per 1,000 EEs. The effects are not homogeneous—the Q χ^2 is 8.81, whereas the $\chi^2_{.05}$ is 5.99 ($df = 2$); the random-effects formulation is more appropriate. That reduction is −16.1 (−39.7, +7.5) admissions per 1,000—not statistically significant (SE = 12). However, the counternull of −32.2 admissions per 1,000 EEs is as likely to be true as the null value of zero; it is about the same size as the fixed-effects counternull.

Two sites (O-3 and O-4) had the cybernetic program in place; one site (O-2), the conventional program. To determine whether the lack of homogeneity found above is due to these different programs, the summary effect of the two cybernetic programs is calculated and compared to the effect of the conventional program. Although O-3 had a reduction in admissions and O-4 an increase, because their homogeneous $\chi^2 = 3.15$ is less than the $\chi^2_{.05}$ critical value of 3.84 (1 *df*), these two cybernetic sites can be combined. Their summary reduction in admissions per 1,000 EEs is −7.1 (−20.3, + 6.2) whereas the conventional program's significant reduction is −37.1 (−57.9, −16.2), SE = 10.6. Its counternull of −74.2 is much



	Study O-2 (Conventional)	Study O-3	Study O-4	Fixed	Random	Cybernetic (Fixed)
Upper Bound	–16.2	–0.1	25.9	–4.5	7.5	6.2
- Effect	–37.1	–17.9	6.2	–15.7	–16.1	–7.1
Lower Bound	–57.9	–35.7	–13.6	–26.8	–39.7	–20.3

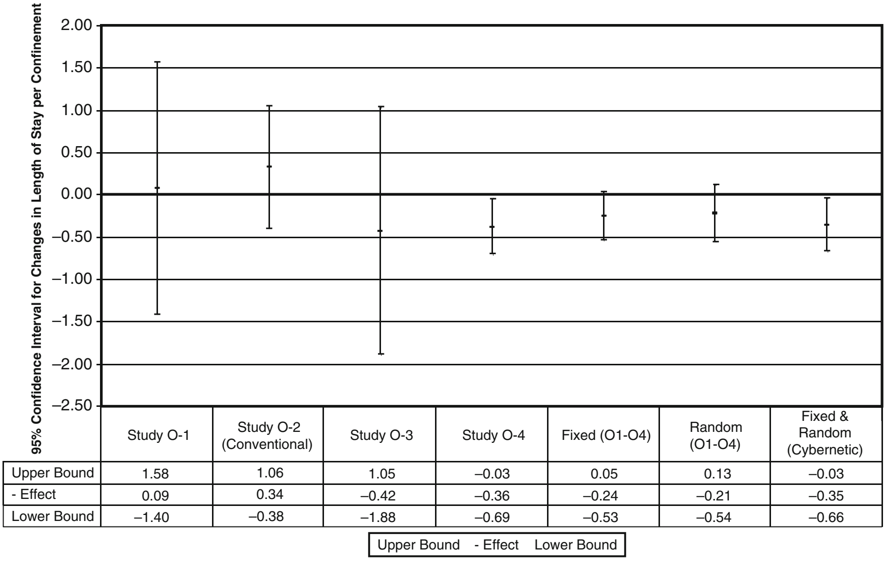
Fig. 14.2 Fixed- and random-effects estimates for admissions, three studies of onsite concurrent review

larger than the cybernetic programs' counternull of −14.2. When the homogeneity of the effects of the two types of programs is tested, the $\chi^2 = 5.68$ is greater than the $\chi^2_{.025}$ critical value of 5.02 (1 *df*); the effects are clearly different.

The conventional site had a generous benefit plan that covered many retired employees and their beneficiaries. To enable their employed children to remain at work, many of these elderly people desired custodial care in the local hospitals rather than care in their homes. Consequently, the target group of the conventional site had a much higher pre-period admissions rate than the two cybernetic sites: 333.4 admissions per 1,000 EEs, compared to 177.9 at the Florida site and 193.9 at the Alaska site. With so many admissions to monitor, it was easier for that onsite nurse to find admissions needing evidence of medical necessity (Smith and Gotowka 1991, 85).

Length of stay

Figure 14.3 presents the fixed- and random-effects estimates for length of stay per confinement. The four target sites experienced a statistically insignificant fixed-effect reduction of −0.24 (−0.53, +0.05)—about one hospital day per four admissions (SE = 0.15). The CI ranges from an increase of one day every 20 admissions (+0.05) to a decrease of one day every two admissions (−0.53). The counternull of −0.48 suggests that a reduction of about one day every 2.1 admissions is as likely as no reduction at all. When the homogeneity of the four



	Study O-1	Study O-2 (Conventional)	Study O-3	Study O-4	Fixed (O1-O4)	Random (O1-O4)	Fixed & Random (Cybernetic)
Upper Bound	1.58	1.06	1.05	−0.03	0.05	0.13	−0.03
- Effect	0.09	0.34	−0.42	−0.36	−0.24	−0.21	−0.35
Lower Bound	−1.40	−0.38	−1.88	−0.69	−0.53	−0.54	−0.66

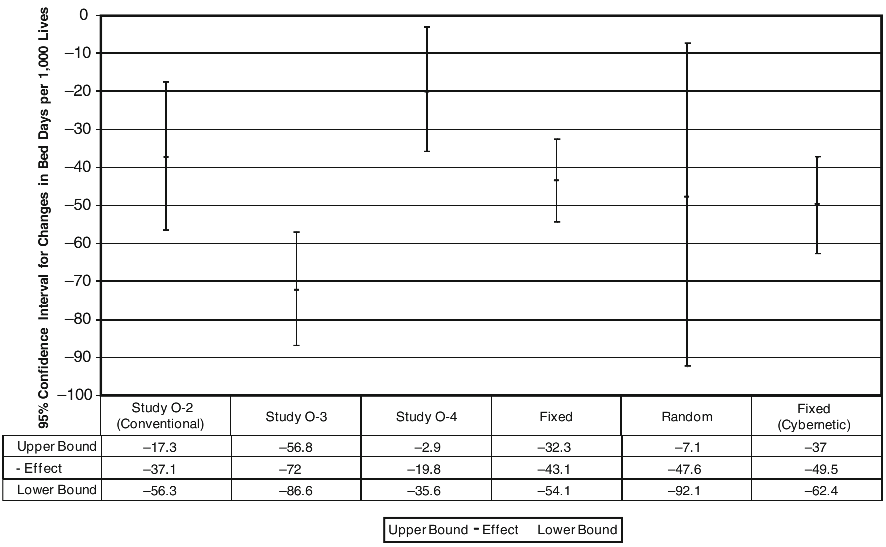
Fig. 14.3 Fixed- and random-effects estimates for length of stay, four studies of onsite concurrent review

sites is tested, the Q χ^2 of 3.24 is not large enough to cause the rejection of the null hypothesis; it is much smaller than the $\chi^2_{.05}$ of 7.81 (DF = 3). However, the obtained χ^2 is greater than 2.37 (the $p = .5$ value) and O-2 is an outlier; the random-effects model is appropriate. Its estimate is not statistically significant: the reduction is −0.21 (−0.54, +0.13), one hospital day every five admissions (SE = 0.17). The CI ranges between a decrease of one day every two admissions and an increase of one day every eight admissions. The counternull of −0.42 implies that a reduction of two days every 4.75 admissions is as likely as no reduction at all.

Because the cybernetic programs allowed the nurses to assign the inpatient days, their consolidated effect indicates a statistically significant reduction in length of stay of −0.35 (−0.66, −0.03), a reduction of one hospital day every three admissions (SE = 0.16). Because these three sites are very homogeneous, their Q χ^2 of 0.35 is less than the *df* and also less than the $Q_{.750}$ value of 0.575 (*df* = 2), the random effects reduce to the fixed effects. The conventional program produced a gain in length of stay of +0.34 (−0.38, 1.06). However, the difference in length of stay between these two programs only approaches statistical significance: The Q χ^2 of 2.90 (*df* = 1) falls between $\chi^2_{.10} = 2.71$ and $\chi^2_{.05} = 3.84$.

Bed days

Bed-day rates per 1,000 lives (EEs plus insured dependents) express the combined effect of length of stay and admissions. Since onsite nurses may reduce both of



	Study O-2 (Conventional)	Study O-3	Study O-4	Fixed	Random	Fixed (Cybemetic)
Upper Bound	–17.3	–56.8	–2.9	–32.3	–7.1	–37
- Effect	–37.1	–72	–19.8	–43.1	–47.6	–49.5
Lower Bound	–56.3	–86.6	–35.6	–54.1	–92.1	–62.4

Fig. 14.4 Fixed-and random-effects estimates for reductions in bed days per 1,000 lives, Poisson regression analysis of three studies of onsite concurrent review

these measures, they also may reduce bed days. The conventional site (O-2) and two cybernetic sites (O-3 and O-4) report bed-day rates. For each of these sites, and for their consolidated effects, Fig. 14.4 plots the Poisson regression estimates for the reductions in these rates, and their lower and upper bounds. The reductions of −37.1, −72, and −19.8 bed days per 1,000 lives are based on the following statistically significant program effects (for Target × Post): O-2 = −0.0593 (SE = 0.0164); O-3 = −0.201 (SE = 0.0234); and O-4 = −0.0702 (SE = 0.0307).[6]

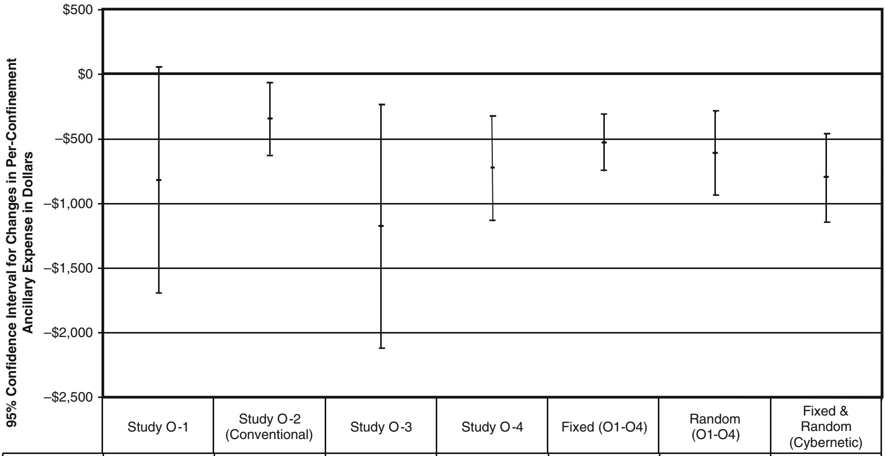
The fixed-effects and random-effects analyses indicate reductions in bed days. The consolidated fixed-effect estimate for the three sites is −0.1002 (SE = 0.0123), which implies a reduction of −43.1 (−54.1, −32.3) bed days per 1,000 lives—the onsite program reduced bed-day rates. Since O-3 is an outlier, contributing 18.6 to the Q χ^2 of 25.7, the effects lack homogeneity ($p < .0005$) and the random-effects model is appropriate. Its estimate of −0.1102 (SE = 0.0475) implies a reduction of −47.6 (−92.1, −7.1) bed days per 1,000 lives—the onsite programs will reduce bed-day rates.[7]

Do the two types of programs have effects of different size? The conventional site (O-2) indicates a reduction of −37.1 (−56.3, −17.3) bed days per 1,000 lives. Taken together, the cybernetic sites (O-3 and O-4) indicate a reduction of −49.5 (−62.4, −37) bed days per 1,000 lives. The program effect for O-2 is −0.0593 (SE = 0.0164). The consolidated fixed-effect estimate for O-3 and O-4 is -0.1529 (SE = 0.0186). The Q χ^2 of 14.25 ($df = 1$) decisively rejects the null hypothesis of no difference in bed-day rates between the two types of programs ($p < .001$)—the cybernetic sites have the larger reductions.

Expense

Total net submitted inpatient expense per admission is defined as the sum of the bona fide charges for physician services, anesthesia, room and board, and inpatient ancillary services (tests, diagnostic procedures, surgical gowns, and so forth). Hospital expense is the sum of ancillary expense and room-and-board expense. Onsite nurses, especially in cybernetic programs, will reduce the per-confinement expense for ancillary, hospital, and total net submitted expense. Their effects on per-confinement room-and-board expense (−$61, SE = $64) and physician expense (random effect = −$24, SE = $40) are not significant. When ancillary expense is subtracted from hospital expense, or from total expense, the reductions in these costs are not significant; the reduction in ancillary expense is pivotal.

For the four onsite programs and their consolidated fixed and random effects, Fig. 14.5 plots the per-confinement reductions in expense for ancillary services. The fixed-effect reduction is −$523 (−$740,−$306) per confinement (SE = $111). Although the Q χ^2 statistic of 4.8 is much less than $\chi^2_{.05} = 7.81$ (3 *df*), which is the critical value for rejection of the hypothesis of homogeneity, it is greater than $\chi^2_{50} = 2.37$, which is the value needed for the acceptance of the assumption of homogeneity. Consequently, the random-effects model is appropriate; it increases the consolidated per-confinement effect to −$605 (−$929, −$281), SE = $165. The random-effects counternull is −$1,210, the fixed-effects counternull is −$1,046.



	Study O-1	Study O-2 (Conventional)	Study O-3	Study O-4	Fixed (O1-O4)	Random (O1-O4)	Fixed & Random (Cybernetic)
Upper Bound	$58	−$60	−$229	−$316	−$306	−$281	−$453
- Effect	−$815	−$340	−$1,171	−$721	−$523	−$605	−$795
Lower Bound	−$1,688	−$621	−$2,113	−$1,127	−$740	−$929	−$1,138

Fig. 14.5 Fixed- and random-effects estimates for reductions in per-confinement expense for ancillary services, four studies of onsite concurrent review

The three cybernetic programs yield stronger per-confinement reductions in ancillary expense than the conventional program. Their consolidated reduction is −$795 (−$1,138, −$453), SE = $175. The three effects are very homogeneous: the Q χ^2 of 0.74 is less than the *df*, and the random effects reduce to the fixed effects. The reduction for the conventional program is much less: −$340 (−$621, −$60), SE = $143; and the difference between the programs is statistically significant ($0.05 > p > 0.025$).

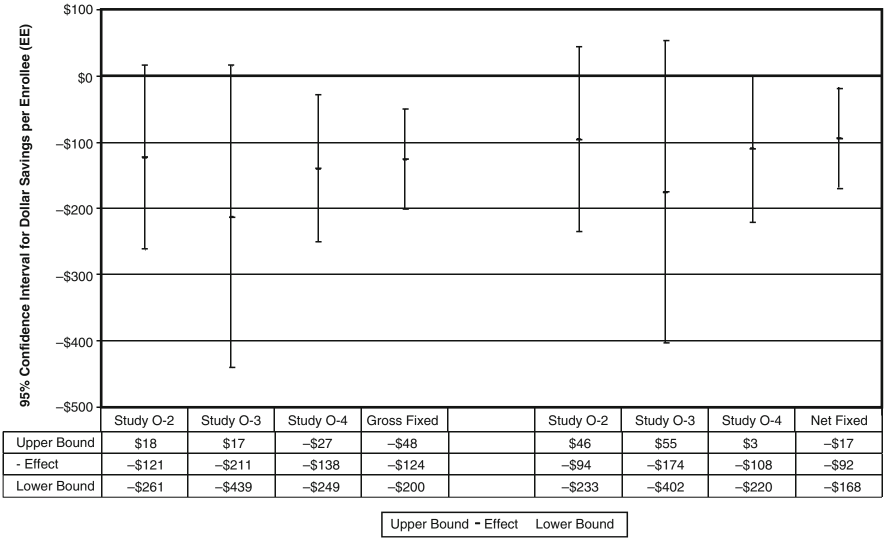
Onsite nurses will significantly reduce per-confinement hospital expense, because they will reduce per-confinement ancillary expense, a component. The random-effects reduction in hospital expense is −$605 (−$939, −$272), SE = $170. The consolidated reduction for the cybernetic programs is even larger: −$873 (−$1,329, −$417), SE = $233. These sites are very homogeneous ($\chi^2 < df$); the random effects reduce to the fixed effects. The conventional program's reduction in hospital expense is not statistically significant; it is −$350 (−$770, +$71), SE = $215. Even so, the difference between these two types of programs only approaches statistical significance, the Q $\chi^2 = 2.74$ ($0.10 > p > 0.05$, $df = 1$).

For total net submitted expense, the consolidated per-confinement reduction is −$618 (−$980, −$257), SE = $184. The four sites have very homogeneous effects (Q $\chi^2 = 1.76 < df = 3$); the random effects reduce to the fixed effects. The three cybernetic programs have a larger consolidated reduction: −$832 (−$1,360,−$304), SE = $269; their effects also are very homogeneous ($\chi^2 < df$) and the random effects reduce to the fixed effects. The conventional program's reduction is not statistically significant; it is −$431 (−$926, +$65), SE = $253. Even so, the difference between the two types of programs is not statistically significant, the Q $\chi^2 = 1.18$ ($df = 1, p \sim .25$).

Savings

For the conventional site (O−2), two cybernetic sites (O-3 and O-4), and their consolidated fixed effects, Fig. 14.6 presents annual reductions (and their bounds) in total inpatient net submitted expense per EE (gross and net of program costs). Since the Q χ^2 of 1.76 is less than the *df*, the random effects reduce to the fixed effects. The bounds of the consolidated gross reduction of −$124 (−$200, −$50) per EE do not include zero. When the program administration costs are taken into account, the bounds of the net reduction of −$92 (−$168, −$17) per EE also do not include zero. The percent gross inpatient savings is 7.7% (3%, 8.7%); the net savings is 5.7% (1%, 10.4%).

The two cybernetic programs taken together produce slightly higher net savings than the conventional program. Their consolidated net reduction is −$111 (−$206,−$16) per EE compared to −$94 (−$233, +$46) per EE for the conventional site. The net percent savings are, respectively, 9% (1.3%, 16.7%) and 5% (−2.5%, 12.5%).



	Study O-2	Study O-3	Study O-4	Gross Fixed		Study O-2	Study O-3	Study O-4	Net Fixed
Upper Bound	$18	$17	–$27	–$48		$46	$55	$3	–$17
- Effect	–$121	–$211	–$138	–$124		–$94	–$174	–$108	–$92
Lower Bound	–$261	–$439	–$249	–$200		–$233	–$402	–$220	–$168

Fig. 14.6 Gross and net annual savings in total net submitted expense per enrollee, three studies of onsite concurrent review

Medical complications

The onsite nurses did not significantly affect rates of surgical or obstetrical complications, or their pooled rate, the effects indicate very small reductions, see Table 14.4.

Discussion

Summary of Findings

Table 14.5 summarizes many of the key findings about these precertification and onsite concurrent review programs. The consolidation of the effects of precertification on admissions compares this program to units that do not manage utilization. For eight studies, the random effects indicate a statistically significant reduction of −7.3 (−13.8, −0.8) admissions per 1,000 EEs. For the universe of studies from which these are samples, precertification will reduce admissions.

Because rates of medical complications express the complexity of medical problems and the skills of the providers, they are at best imperfect measures of the appropriateness of care. Even so, in Florida, precertified surgical confinements had a slightly higher risk of surgical complications than surgical confinements that were not precertified, the odds ratio was 1.87 ($p = 0.04$). The precertification process may

Table 14.4 Counts of medical complications by onsite review status, enrollment status, and location of hospitals, January 1986 to September 1989

Panel (A), surgical	Ohio Site 2, active EEs		Ohio Site 2, inactive EEs		Florida Site 3	
confinements	Onsite Hospitals	Other Hospitals	Onsite Hospitals	Other Hospitals	Onsite Hospitals	Other Hospitals
Pre time period	86,001–87,181	86,001–87,181	86,001–87,181	86,001–87,181	86,244–88,244	86,244–88,244
Only precertification						
Some complications	10	30	3	41	13	19
No complications	956	1,943	173	2,032	1,019	1,731
Total confinements	966	1,973	176	2,073	1,032	1,750
Post time period	87,182–88,366	87,182–88,366	87,182–88,366	87,182–88,366	88,245–89,243	88,245–89,243
Precert + onsite review						
Some complications	4	30	1	40	7	4
No complications	727	1,496	125	1,765	453	832
Total confinements	731	1,526	126	1,805	460	836

Effects of onsite review in post time period: β = −0.403, SE = 0.385, Exp (β) = 0.668, $p = 0.295$

Best log-linear model with sites distinguished: Onsite Hospitals × Sites × Complications, Post Time × Sites; $\chi^2 = 9.58$, DF = 9, $p = .39$

Best log-linear model pooling sites: Onsite Hospitals × Post Time, Onsite Hospitals × Surgical Complications; $\chi^2 = 1.37$, DF = 2, $p = .50$

Panel (B), obstetrical	Ohio Site 2, active EEs		Ohio Site 2, inactive EEs		Florida Site 3	
confinements	Onsite Hospitals	Other Hospitals	Onsite Hospitals	Other Hospitals	Onsite Hospitals	Other Hospitals
Pre time period	86,001–87,181	86,001–87,181	86,001–87,181	86,001–87,181	86,244–88,244	86,244–88,244
Only precertification						
Some complications	15	61	1	7	77	102
No complications	82	231	1	26	349	420
Total confinements	97	292	2	33	426	522
Post time period	87,182–88,366	87,182–88,366	87,182–88,366	87,182–88,366	88,245–89,243	88,245–89,243
Precert + onsite review						
Some complications	12	37	2	7	28	49
No complications	40	157	31	14	161	209
Total confinements	52	194	33	21	189	258

Effects of onsite review in post time period: β = −0.154, SE = 0.251, Exp (β) = 0.858, $p = .54$

Best log-linear model with sites distinguished: Sites × Onsite Hospitals × Post Time, Complications; $\chi^2 = 12.70$, DF = 11, $p = .31$

Best log-linear model without sites: Onsite Hospitals, Post Time, Complications; $\chi^2 = 4.32$, DF = 4, $p = .37$

Panel (C), surgical and/or	Ohio Site 2, active EEs		Ohio Site 2, inactive EEs		Florida Site 3	
obstetrical	Onsite Hospitals	Other Hospitals	Onsite Hospitals	Other Hospitals	Onsite Hospitals	Other Hospitals
Pre time period	86,001–87,181	86,001–87,181	86,001–87,181	86,001–87,181	86,244–88,244	86,244–88,244
Only precertification						
Some complications	28	100	4	48	95	130
No complications	955	1,943	172	2,032	1,038	1,694
Total confinements	983	2,043	176	2,080	1,133	1,824
Post time period	87,182–88,366	87,182–88,366	87,182–88,366	87,182–88,366	88,245–89,243	88,245–89,243
Precert + onsite review						
Some complications	20	79	3	48	38	55
No complications	719	1,500	126	1,761	458	817
Total confinements	739	1,579	129	1,809	496	872

Effects of onsite review in post time period: $\beta = -0.157$, SE $= 0.188$, Exp $(\beta) = 0.855$, $p = .403$

Best log-linear model with sites distinguished: Sites × Onsite Hospitals × Complications, Sites × Post Time; $\chi^2 = 5.36$, DF $= 9$, $p = .80$

Best log-linear model pooling sites: Onsite Hospitals × Post Time, Complications; $\chi^2 = 5.37$, DF $= 3$, $p = .15$

Table 14.5 Summary of key findings

	Fixed-effects estimates		Random-effects estimates	
Precertification				
Admissions per 1,000 EEs	−7.6	(−10.3, −4.9)	−7.3	(−13.8, −0.8)
Odds ratios for complications				
Surgical	1.87	($p = 0.04$)	Not applicable (NA)	
Obstetrical	0.93	($p = 0.55$)	NA	
Surgical or obstetrical	1.13	($p = 0.25$)	NA	
Onsite concurrent review				
Admissions per 1,000 EEs	−15.7	(−26.9, −4.5)	−16.1	(−39.7, 7.5)
Cybernetic program	−7.1	(−20.3, +0 6.2)	NA	
Conventional program[a]	−37.1	(−57.9, −16.2)	NA	
Length of stay per confinement	−0.24	(−0.53, +0.05)	−0.21	(−0.54, +0.13)
Cybernetic program[b]	−0.35	(−0.66, −0.03)	Same as fixed-effects estimates.	
Conventional program	0.34	(−0.38, +1.06)	NA	
Bed days per 1,000 insured lives	−43.1	(−.54.1, −32.3)	−47.6	(−92.1, −7.1)
Cybernetic program[a]	−49.5	(−62.4, −37)	NA	
Conventional program	−37.1	(−56.3, −17.3)	NA	
Ancillary expense per confinement	−$523	(−$740, −$306)	−$605	(−$929, −$281)
Cybernetic program[a]	−$795	(−$1,138, −$453)	Same as fixed-effects estimates.	
Conventional program	−$340	(−$621, −$60)	NA	
Annual net savings per EE	−$92	(−$168, −$17)	Same as fixed-effects estimates.	
Percent	5.7%	(1%, 10.4%)		
Cybernetic program	−$111	(−$206, −$16)	NA	
Percent	9%	(1.3%, 16.7%)		
Conventional program[c]	−$94	(−$233, +$46)	NA	
Percent	5.0%	(−2.5%, 12.5%)		
Odds ratios for complications				
Surgical	0.67	$p = 0.30$	NA	
Obstetrical	0.86	$p = 0.31$	NA	
Surgical or obstetrical	0.86	$p = 0.40$	NA	

Significance of Difference Between Cybernetic and Conventional Programs. Not applicable(NA)
[a]$p < 0.05$, [b]$0.10 > p > 0.05$, [c] Difference is not significant.

have averted some surgeries that lacked evidence of medical necessity, thus increasing the pool of patients with complicated cases, and thus raising the odds of surgical complications. Of course, this effect could be directly due to the program or, given its weak probability level, it could be due to chance. Precertification did not affect the rate of obstetrical complications ($p = 0.55$), or the pooled rate of surgical and obstetrical complications ($p = 0.25$). When these three p-values are corrected for multiplicity using the step-down Bonferroni procedure, all of their values are not statistically significant: respectively, $p = 0.12$, $p = 0.55$, and $p = 0.50$; implying no causal effect of precertification on these measures of complications.

The consolidation of the effects of onsite concurrent review compares sites that use both precertification and onsite review to sites that use only precertification. It also compares the cybernetic and conventional programs. Cybernetic programs allow the onsite nurses to negotiate directly with the attending physicians about the appropriateness of the intensity of care, and to use their discretion about the assignment of additional bed days. Contrariwise, a conventional program requires the onsite nurse to contact an offsite physician reviewer, who may contact the attending physician to discuss the patient's care. Since the overall effects of the onsite program combine the often larger effects of the cybernetic programs with the usually smaller effects of the conventional program, the overall effects are less favorable than those of the cybernetic program. The conventional program not only is more complicated to administer than the cybernetic program, its reliance on indirect feedback produces an inordinate number of retrospective denials, which intensifies the providers' hostility toward the program. Observations and interviews at the conventional site and discussions with the program administrators clearly indicate that the design of the cybernetic program is much better than the design of the conventional program. Because a cybernetic program encourages the nurse to exercise discretion, to directly discuss the case with the attending physician, and to collaborate with the providers of care, it moves the provision of care closer to appropriate standards (Restuccia 1983, 60–62).

Overall, the onsite programs will not significantly reduce admissions. For the fixed-effects model, the reduction is −15.7 (−26.9, −4.5) admissions per 1,000 EEs; but for the random-effects model the reduction is −16.1 (−39.7, 7.5). The counternull value of −32.2 is as likely to be true as the null value of zero. The conventional program's reduction of −37.1 (−57.9, −16.3) admissions per 1,000 EEs is significantly greater than the cybernetic program's reduction of −7.1 (−20.3, +6.2). Most probably, this large difference is an artifact of the very high rate of admissions at this particular conventional site. The target group had a pre-period rate of 333.4 admissions per 1,000 EEs, compared to about 186 in the cybernetic groups. With so many admissions to monitor, the onsite nurse could easily find some that lacked evidence of medical necessity. Thus, to determine if future onsite review programs—either conventional or cybernetic—will lead to reductions in admissions, more evidence is needed.

Onsite nurses will reduce the length of stay when they have discretion to assign all hospital days. For the cybernetic program the random-effects estimate is −0.35 (−0.66, −0.03) days per confinement. Because it is difficult to take away days that have already been given, in the conventional program, which required the precertification nurses to assign the initial length of stay, the onsite nurses did not significantly reduce the number of days—the fixed-effects estimate is +0.34 (−0.38, 1.06). The difference between the two programs approaches statistical significance ($.10 > p > .05$).

The onsite nurses will reduce bed days. The random-effects estimate from the Poisson regressions across the three studies indicate an overall reduction of −47.6 (−92.1, −7.1) bed days per 1,000 insured lives. The cybernetic program's reduction, −49.5 (−62.4, −3.5), is larger ($p < 0.05$) than the conventional program's reduction, −37.1 (−56.3, –17.1).

Onsite nurses will reduce expense for ancillary services, the overall random-effects reduction is −$605 (−$929, −$281) per confinement. The per-confinement reduction for the three cybernetic programs, −\$795 (−\$1,138, −\$453), is significantly larger ($p < 0.05$) than that for the conventional program, −\$340 (−\$561,−\$60). Because of the large overall reduction in ancillary expense, the random-effects estimate also indicates that the onsite nurses will reduce per-confinement hospital expense, −\$605 (−\$939, −\$272), and total net submitted expense, −\$618 (−\$980, −\$257)—the cybernetic program will produce even larger reductions. The onsite nurses did not significantly reduce expense for physician services, room and board, and anesthesia.

Onsite review nurses did not significantly change rates of complications. Since the precertification and onsite programs may reduce negative iatrogenic effects, they may actually improve the wellbeing of patients.

Interpretation of Findings: Gatekeepers and Sentinels

The reductions in admissions due to precertification and in bed days due to onsite concurrent review exemplify the effect of a gatekeeper. Precertification merely requires a physician to assemble the evidence that validates the medical necessity of a patient's admission to the hospital. The precertification nurse assesses the evidence the physician provides by comparing it to accepted standards for the type of illness. By applying protocols similar to those used by the precertification nurse, the onsite nurse also tries to avert any utilization that lacks evidence of medical necessity. To contest a nurse's decision to deny an admission or hospital day, the attending physician can negotiate directly with the nurse, strengthen the evidence, appeal the nurse's decision to a utilization review physician, or provide the care at the risk of financial penalties for the patient.

The reduction in expense for inpatient ancillary services due to the onsite nurse exemplifies the effect of a sentinel. The onsite nurse did not explicitly monitor ancillary expense, but a nurse's presence in the hospital seems to have inhibited the physician and hospital staff from billing charges for ancillary services. Davis et al. (1990, 52) attribute a similar mechanism to Medicare's prospective payment system (PPS).

To sort out whether the reduction in inpatient ancillary expense is due to reduced length of stay or to a sentinel effect, Study O-3 assesses changes in the size of the program effect when length of stay is introduced in the regression equations. If this additional control for length of stay reduces the size of the program effect, then the more appropriate interpretation of the impact of the program on ancillary expense would be the change in length of stay and not a sentinel effect (Davis et al. 1990: 18). Study O-3 reports that the inclusion of length of stay in the regression equation predicting ancillary expense had little effect—it changed the per-confinement regression coefficient from −\$1,171 (−\$2,113, \$−229) to −\$1,023, which falls within the CI. This small difference is consistent with the interpretation that it was a sentinel effect of the nurse that inhibited the physician and hospital staff from charging for ancillary services. However, the mechanism that brought about this sentinel effect is unknown—it may have been due to a stable, normative change in practice patterns; or it could be a transient effect, dependent on the continued presence of the onsite nurses; or it could merely indicate a shifting of ancillary services to outpatient settings.

Generalization of Findings

Although the significant random-effects estimates reported above suggest that these findings can be generalized to the universe of such studies, because the studies cover only the period from 1986 to 1990 and describe the experience of only one private insurer, it is a rather limited universe. Consequently, the findings and the specific estimates of net savings to the corporate policyholder attributable to precertification (3.3% total; 6.4% inpatient) and to onsite concurrent review of inpatient care (5.7% inpatient) should not be over-generalized to other time periods, demographic populations, and insurance programs—private or public.

The interpretive constructs of gatekeeper effect and sentinel effect are relevant to Medicare, which has used both to control its payments to providers. In 1986, to supplement its Prospective Payment System (PPS), which used diagnostic related groups (DRGs) to price inpatient stays, Medicare instructed its peer-review organizations (PROs) to act as gatekeepers by precertifying inpatient care. During the period 1986–1990 Medicare required precertification for a variable list of 10 inpatient procedures that included hernias, gallbladders, hysterectomies, cataracts, and other maladies. Since the Medicare beneficiaries were older and frailer than the members of the privately insured groups in this consolidation, the evidence supporting the medical necessity of their care usually was persuasive—the precertification nurses almost always certified the care. Moreover, because of changes in style of medical practice and the need to reduce costs, physicians began to perform many of these procedures on an outpatient basis. Consequently, during the period 1990–1993 Medicare phased out inpatient precertification and began to focus on improving the quality of care by identifying and encouraging best practices. Circa 1999 the attending physician decided when a Medicare patient required hospitalization. Relying on the sentinel effect of the threat of retrospective reviews, PROs focused more on reducing payment errors—the miscoding of DRGs—and inappropriate admissions. (Interview with Janet Moroney and Christopher G. Richards of MassPro, November 17, 1999.)

In his review of possible new cost-containment strategies for Medicare, Peter Fox (1997, 50) states, “Prior authorization of selected high-cost services has potential for achieving savings.” However, he neither enumerates which procedures Medicare should precertify nor presents any evidence bearing on the program’s potential costs and benefits. Clearly, given Medicare’s past history with precertification, before Medicare broadly reintroduces this program, it should test it in pilot studies and carefully evaluate its consequences.

Medicare uses its resource-based relative value scale (RBRVS) to pay physicians on a cost per unit of service basis; it thus has little control over the volume of services physicians may charge (David Smith 1992, 181–218; Wennberg and Cooper 1999, chapter 6). Assuming that health plans are reluctant to reintroduce the gatekeeping of admissions by precertification nurses (Mays et al. 2004, W4: 430), then this absence of monitoring may encourage undue patient- and physician-created demand for inpatient care. It may also encourage DRG-creep, a provider's classification of a patient in a DRG with higher payments by up-coding the level of service. If a Medicare nurse was onsite—a sentinel—then this person could judge the appropriateness of the admission and length of stay; the level of care (especially to minorities who often receive less care than others); the number of physician visits to the patient; and the number of services. These nurses could also: prevent DRG-creep; prevent hospitals from shaving a patient's length of stay to increase their DRG profit; act as a patient advocate; perform case management; and help the patient with discharge planning and home health care.

Although Restuccia's 1982 study implies that fee-for-service Medicare had previously used onsite nurse coordinators, it definitely did not use them (at least in Massachusetts) during the period 1986–2000, and most probably to the present (ca. 2011). Because onsite review nurses can combine both the gatekeeping and sentinel functions, along with provision of some care if needed, patient advocacy, case management, and discharge planning; they are a better option for Medicare than precertification nurses. Because anecdotal evidence suggests that physicians will oppose the precertification of inpatient care and onsite nurses who monitor it, this recommendation awaits further empirical tests and evaluation of the evidence for causality.

Acknowledgements The author thanks Dr. Greta Ljung for explicating aspects of the fixed- and random-effects statistical models and for her collaboration and editorial critique of early versions of this chapter. He also thanks the following people for sharing their knowledge with him: Duane Hayes of the SAS Institute clarified aspects of Poisson regression analysis; Dr. Thomas E. Gotowka critiqued the original reports; Dr. Rezaul Khandker, Dr. Sharon Fox, and Dr. Joanna Handlin Smith critiqued drafts of the article on which the present chapter is based; Ms. Janet Moroney and Mr. Christopher G. Richards of the Massachusetts Peer Review Organization (MassPRO) clarified how PROs control the costs and quality of Medicare. Earlier versions of this chapter were presented at the 1996 Joint Statistical Meetings in Chicago, August 4–8, 1996 and at the Conference in Honor of Clifford Clogg, Pennsylvania State University, September 26–28, 1996. Main portions of this chapter were published in the *Evaluation Review* (vol. 25, no. 3, 228–330; © 2001 Sage Publications). The author thanks the publisher for allowing him to reuse this material in this book. The views expressed herein are the author's own and do not necessarily reflect those of any colleagues or organizations.

Endnotes

[1] A symposium on the Future of Medicare discussed strategies for reform but did not consider the possible use of private-sector innovations to control costs (McClellan 2000; Cutler 2000; Fuchs 2000; Reinhardt 2000; Saving 2000). However, Mays et al. (2004, W4:429) indicate that some

health plans have reintroduced prior authorization requirements after having eliminated these requirements.

[2] A recent study by Gary King and colleagues (King et al. 2009) has advanced this research paradigm by introducing random assignment. The treatment encouraged previously uninsured Mexicans to enroll in a health insurance program and in upgraded medical facilities.

[3] Deaton (2009, 25–41) critiques the applicability of randomized control trials (RCTs) to studies of development and wellbeing because of their narrow scope and assumptions that often are not met. He prefers a model that links contexts, mechanisms, and outcomes.

[4] When P-3 is removed from the analysis (because of its possible dependency with P-7) and P-4 is also removed (because it is an outlier), the fixed-effects summary is −11.6 (−15, −8.3) admissions per 1,000 EEs. When the homogeneity of the effects is tested, the six effects are found to be homogeneous, the Q χ^2 of 0.604 is less than the *df*; the random effects are the same as the fixed effects. The counternull is −23.3.

[5] When only P-3 is removed from the analysis (because of its possible dependency with P-7), the fixed-effects summary is −10.7 (−14, −7.4) admissions per 1,000 EEs. When the homogeneity of the effects is tested, the statistics clearly indicate that P-4 is an outlier—it contributes 14.74 to the Q χ^2 of 15.66. The assumption of homogeneity is rejected (DF = 6, $p < 0.025$) and the random effects must be calculated. That reduction is −8.4 (−15.3,−1.4) admissions per 1,000 EEs; the counternull is −16.8.

[6] For each specific site, bed-day rates are calculated using this equation from (3) above:

$$\text{bed days} = \text{lives} \times e^{(\alpha + \beta 1\text{Post} + \beta 2\text{Target})} \times \left(e^{\beta 3}\right)^{\text{TargetPost}}$$

For O-3, for example: lives = 13,718, $\alpha = -0.532$, $\beta_1 = -0.319$, $\beta_2 = -0.0769$, and $\beta_3 = -0.201$. When TargetPost is 0 (onsite review is absent), then $e^{\beta 3} = 1$; the factor has no effect. Then the estimated bed days = lives × $e^{(\alpha\ +\ \beta 1\text{Post}\ +\ \beta 2\text{Target})} = 5{,}423.9$ (= 13,718 × 0.395383). When TargetPost is 1 (onsite review is present), the factor $e^{\beta 3} = 0.8179$ will affect the estimate of bed days. When the onsite program is present at site O-3, the estimated bed days are 5,423.9 × 0.8179 = 4,436.2. The difference 5,423.9 − 4,436.2 indicates a reduction of −987.7 bed days due to onsite review. When this reduction in bed days is divided by the number of lives, the reduction in the bed-day rate for O-3 is −987.7/13,718 lives, or −72 bed days per 1,000 lives, as reported in Fig. 14.4. The upper and lower bounds for the bed-day rates are calculated using as the factors the exponentiated 1.96 CI bounds of the program effect coefficients.

[7] The change in the bed-day rates are calculated as follows—this example uses the random-effects estimates. The exponentiated value of $e^{-0.1102}$ is 0.8957. Since the total bed days in the target group in the post-period is 16,207, the predicted bed days when the onsite program is absent would be 16,207/0.8957 = 18,094, for a difference of −1,887 bed days attributable to the program. When this difference is divided by 39,648, the number of lives in the target group in the post-period, the reduction of −47.6 bed days per 1,000 lives is obtained. When the exponentiated values of the upper and lower CI bounds for the random-effect estimate are used in the calculations, the reductions range from −7.1 to −92.1 bed days per 1,000 lives.

Chapter 15
Childhood Vaccinations and Autism

Time to look beyond MMR in autism research. . . . Is the measles, mumps, rubella (MMR) vaccine safe? Yes, acceptably so, is the only conclusion possible to reach in the face of the totality of the epidemiological evidence. There are no substantiated data to suggest that the MMR vaccine causes autism, enterocolitis, or the syndrome first described by Andrew Wakefield and his colleagues in The Lancet in Wakefield et al. 1998.

—Editorial, *The Lancet*, (23 February 2002)

In particular, the claims in the original paper that children were "consecutively referred" and that investigations were "approved" by the local ethics committees have been proven to be false. Therefore, we fully retract this paper from the published record.

—Editors, *The Lancet*, (2 February 2010)

The furor over possible adverse effects of childhood vaccinations provides an informative case study that can help uncover approaches for assessing assumed causal relationships in empirical studies. Briefly put, a causal relationship between vaccinations and autism-spectrum disorders (i.e., autism, atypical autism, and Asperger's syndrome) is not supported by the evidence from well-designed and executed clinical studies. But, due to unwarranted publicity-touting flawed studies that claim that vaccinations cause autism and other adverse events, large numbers of parents in the USA and the UK now believe that the risks presumably associated with vaccinations outweigh their disease-prevention benefits, and are choosing not to vaccinate their children, even to prevent the consequences of polio and other diseases. As the percentage of unvaccinated children in an area increases, the risk of outbreaks of diseases increases, creating potentially severe public health problems. Moreover, as the vaccination coverage decreases in developed countries, it becomes more difficult to justify vaccinations in countries with lower levels of human development, thus worsening the health of children globally (United Nations Development Program 2003, 254–257).

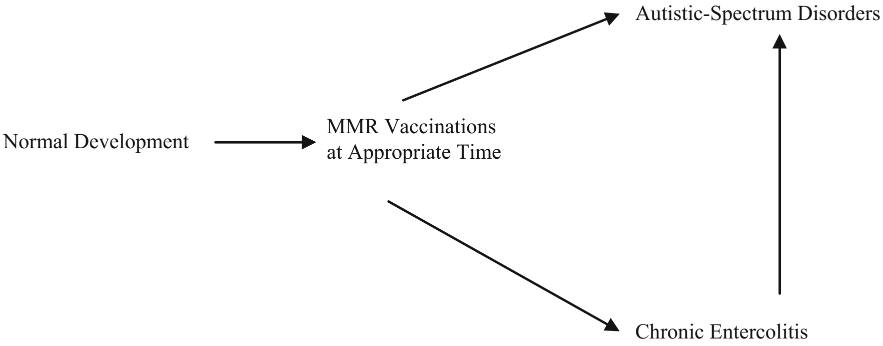
This chapter traces how a flawed but publicized study (Wakefield et al. 1998) intensified this controversy about the supposed adverse consequences of vaccinations—*The Lancet* has now retracted this study (Editors, February 2, 2010); examines

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_15, © Springer Science+Business Media B.V. 2011

the beliefs of parents whose children are autistic; and shows how health scientists in Europe and the USA critically reviewed studies leading them to decide that the empirical evidence does not support a causal linkage between childhood vaccinations and autism-spectrum disorders (ASD). Their procedures provide methodological frameworks for assessments of causality; the concluding chapter of this book applies aspects of these frameworks to determine the level of causality of the findings from the multilevel modeling of the previous chapters.

The MMR-Autism Controversy in Britain

Research studies on sensitive topics can have unanticipated consequences.[1] A peer-reviewed article in the *The Lancet* (1998, 637–641), by Andrew J. Wakefield and 12 other British physicians and health workers, suggested that the single-shot childhood vaccination for measles, mumps, and rubella (MMR) predetermined behavioral symptoms of autism and chronic colitis, and the latter provided a biological mechanism that predetermined the autism. The 12 children presumably exhibited a history of normal development followed by loss of acquired language and other skills, diarrhea, and abdominal pain after they had received the MMR vaccination. The investigators hypothesized that normal development preceded the MMR vaccination, which then led to the autistic-spectrum disorders, and to biological maladies, which then predetermined the behavioral disorders; these health researchers implicitly suggested this pattern of associations:



The investigators concluded their article saying (1998, 641): “We have identified a chronic enterocolitis in children that may be related to neuropsychiatric dysfunction. In most cases, onset of symptoms was after measles, mumps, and rubella immunisation. Further investigations are needed to examine this syndrome and its possible relation to this vaccine.” Their article only reported early findings and it clearly stated that a causal linkage between MMR vaccinations and the syndrome of intestinal dysfunction in children with autistic-spectrum disorders was not proven (1998, 641).

Contrarily, Wakefield, the lead author, participated in a press conference soon after the article's publication, stating that he could not support the continued use of these three vaccines given in combination, until the issue of their safety is resolved (Deer, February 8, 2009, e2; McCartney 2009, 10).[2]

Because of the mass media's reporting of Wakefield's views, his press conference precipitated the "MMR-autism controversy" (Ramsey 2001). Consequently, many concerned parents in Europe, the UK, the USA, and the developing world now question the safety of childhood vaccinations and choose not to vaccinate their children, even though numerous authoritative research studies report that childhood vaccinations and autism are unrelated (e.g., Institute of Medicine 1994, 2001a, 2001b, 2004; Demicheli et al. 2005). Moreover, Brian Deer, an investigative reporter, claimed in *The Sunday Times* (8 February 2009, e1–e7) that the 1998 study lacks credibility because, contrary to what the authors state in their article, the medical records that he recently examined indicate that some of the children had symptoms of autism prior to their vaccinations and others were not diagnosed with autism-spectrum disorders. Deer (8 February 2009, e6–e7) also points out that Wakefield received money for helping prepare legal cases about the adverse consequences of vaccines and that this relationship may have shaped the nature of the study.[3]

The controversy stemming from the 1998 article and press conference triggered serious public health problems (Deer, February 8, 2009, e2):

> In Britain, immunisation rates collapsed from 92% before the Lancet paper was published, to 80% at the peak of Britain's alarm. Measles has returned as officially "endemic." With less than 95% of the population vaccinated, Britain has lost its herd immunity against the disease. In 1998 there were 56 cases reported; last year there were 1,348, according to figures released last week that showed a 36% increase on 2007. Two British children have died from measles, and others put on ventilators, while many parents of autistic children torture themselves for having let a son or daughter receive the injection.[4]

Addressing this public health problem, *The Lancet* has published articles by epidemiologists that find no causal association between the MMR vaccinations and the subsequent onset of autistic-spectrum disorders (e.g., Taylor et al. 1999; Farrington et al. 2001; Smeeth et al. 2004); and medical researchers have tested the assumed biological mechanisms and found no convincing relationships (as related by De-Stefano and Chen 1999, 1987–1988; Ashaf 2001, 1341; Editorial, *The Lancet* 2002 637; Editors, *The Lancet* 2 February 2010). Subsequently, ten of the 13 coauthors of the Wakefield article disavowed its interpretation that the MMR vaccinations may be causally linked to autism-spectrum disorders (Murch et al. 2004).

Because of personal attacks, charges, and countercharges, by mutual agreement Wakefield resigned his research position at the Royal Free and University College London (Ramsey 2001), but he continues to study relationships between childhood vaccinations, biological mechanisms, and autism in the USA.[5] Parents of children with autistic-spectrum disorders, who believe that childhood vaccinations caused their child's autism, still value his research. Wakefield received the first annual Andrew J. Wakefield Award for Courage in Medicine (www.autismone.org, viewed 5/19/09).

The Vaccination–Autism Controversy in the USA

The British controversy spilled over to the USA providing support for the many parents in this country who believe that childhood vaccinations cause autism and other adverse events. Such views are reinforced by mass media reports, some physicians, parents of autistic children, and web sites. For example, AutismCoach is a web site (www.autismcoach.com) that disseminates information (and misinformation) about autism, and sells books, educational software for autistic children, and dietary supplements that presumably prevent or reduce the manifestations of autism.[6] It conducted a poll of its customers about their attitudes and beliefs about the causes of their children's autism-spectrum disorders. By 14 October 2004, 53 adults (parents and grandparents) responded. The response categories are Yes = 70% (Immunization triggered my child's autism); Partially = 5.5% (Immunization was partially responsible); Don't Know = 12%; and No = 12.5% (Immunization didn't trigger my child's autism). Many parents of autistic children interpret the concurrent association between the timing of the vaccinations and the onset of behavioral manifestations of autism as a causal chain, with the autism following very soon after the vaccinations:

Age of Child → Vaccinations → Autism-Spectrum Disorders (Model 1)

Illustrating parental beliefs about the appropriateness of this causal chain, panel (15.1a) of Table 15.1 presents some of the qualitative responses to this poll. The parents describe their children as initially normal, the children are given shots following the recommended age-based schedule of vaccinations, and very soon after receiving the shots the children's development deteriorates. One parent subscribes to a conspiracy theory that the linkage between vaccinations and autism is being covered up due to the successful lobbying of the pharmaceutical companies and the economic consequences of disclosure. Another parent is a chemical engineer who studies metal poisoning. He believes that the vaccinations and the "mercury poisoning" due to thimerosal, a preservative previously used in many vaccines, induced his child's autism. These parents may exhibit "motivated reasoning" in which a person's prior beliefs and knowledge affect his or her evaluations of evidence—people believe what they want to believe to the extent that reason allows (Kunda 1990).

Panel (15.1b) of the table presents statements of two parents from different families who acknowledge that their children initially had developmental problems that improved prior to their vaccinations and then worsened after the vaccinations; they see no relationship (dashed line) between the initial problems and the post-vaccination disorders:

$$\begin{tikzcd}
\text{Age of Child} \arrow[r] & \text{Vaccinations} \arrow[r] & \text{Autism-Spectrum Disorders (Model 2)} \\
\text{Initial Problems} \arrow[urr, dash, dashed] & &
\end{tikzcd}$$

Both parents indicate that their child was adopted: one child had drug exposure and lack of oxygen as an early problem; the other child had slight development delays

Table 15.1 Qualitative statements suggesting vaccinations cause Autism disorders

(15.1a) Statements that vaccinations preceded onset of Autism disorders:	
Yes	"I believe an immunization triggered my child's autism. He was a perfect child before the immunization."
Yes	"I believe immunizations caused my child's autism."
Yes	"I fully believe that vaccinations (and mercury poisoning) played a role in my son's PDD-NOS [Pervasive Developmental Disorder-Not Otherwise Specified = atypical autism]. I am a chemical engineer and have studied metal poisoning for the past 3 years."
Yes	"I am positive that the vaccination was the environmental factor that brought out my daughter's autism. I am also sure that there is a cover up due to the catastrophic consequences this would bring not only to the pharmaceutical companies (the largest lobbying body in our nation) but to our entire economy."
(15.1b) Statements Acknowledging Prior Problems but Vaccinations Caused Autism	
Yes	"We are absolutely sure that our child's vaccinations were responsible in large part for his retardation and autism. He was adopted at 5 weeks with drug exposure and lack of oxygen being the main concerns. He was developing normally until his 6 mo DPT shots and then began to fall behind. He never stopped screaming. At each set of DPT shots, he lost more skills, speech, imitation, cognition, etc. We did hair analysis at 18 mos-2 yr and his levels of aluminum were very high. Today he is a severely handicapped 10 yr old boy, who is non-verbal, in diapers, with severe behaviors and OCD, and on many meds to try and control his behaviors."
Yes!!	"We adopted a 2 year old boy who had slight delays due to his past environment. Once he was in our home he started developing normally with good speech/communication, improved motor skills etc. He was expected to be fine within a short time! He continued to develop normally, and after he had been living with us for 4 months we took him to his check up and he received the MMR shot. During the shot he screamed MAMA and arched his back and became stiff as a board!!!! We immediately noticed at home that we did not bring home the same child . . . we NEVER heard him speak another word, his muscle tone is extremely tight/high he had to learn all over again how to walk and his motor skills have never got back to normal, he started grunting, hand flapping, and stimming [sic] off lights just to name a few things. None of which he had been doing before the shots!! Put us down for a BIG Yes!!!!!!!!!!"

Source: Immunization Poll of AutismCoach Customers, Updated 14 October 2004. Downloaded from www.AutismCoach.Com, circa 10 September 2009.

prior to his adoption at age 2. Even so, both parents discount a direct linkage between these prior conditions and the autism-spectrum disorders; the vaccinations are the culprit. In both cases the children's disabilities are severe. For these problems the adoptive parents could hold responsible themselves, the birth parents, the child, or the adoption agencies, or they could hold responsible the vaccinations; the latter may be the easier choice for them. But, regardless of the existence of initial problems or not, a number of recent case–control studies reviewed by Demicheli et al. (2005, 10–11) have shown that the childhood MMR vaccinations do not directly lead to autism-spectrum disorders. Relevant studies include Taylor et al. (1999); Farrington et al. (2001); Smeeth et al. (2004); and DeStefano et al. (2004).

Some parents in the poll do not say that the vaccinations caused their child's autism, thereby implicitly agreeing with the empirical findings of no relationship between vaccinations and autism; see Table 15.2. Rather, they believe that the disorders are related to the child's preexisting conditions and not to the vaccinations:

$$\begin{tikzcd}
\text{Age of Child} \arrow[r] & \text{Vaccinations} \arrow[r, dash, dashed] & \text{Autism Spectrum Disorders (Model 3)} \\
\text{Initial Problems} \arrow[urr] & &
\end{tikzcd}$$

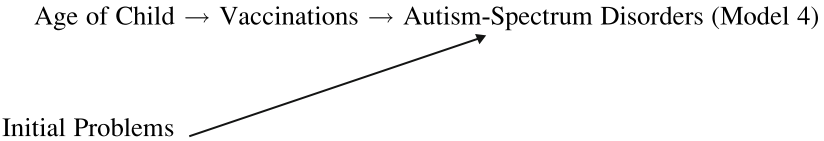
The first panel of Table 15.2 displays the reasoning of these parents and grandparents: some characterize their children as having autism at birth; while others see no correlation between the vaccinations and their children's autism-spectrum disorders. Rigorous evaluations of evidence from clinical studies show that vaccinations and autism-spectrum disorders are unrelated, and support this model of causality (Institute of Medicine 1994, 2001a, 2001b, 2004; Demicheli et al. 2005).

Table 15.2 Qualitative statements about prior problems and Autism disorders

(15.2a) Statements that vaccinations did not cause Autism disorders:	
No	"My son was odd from the start."
No	"My grandson's autism was present from birth. I knew he was not normal when I first held him. He was 1 week old, totally unable to make eye contact."
No	"I don't believe immunizations caused my daughter's autism. She is now 6, and seemed to have slight developmental issues from a very early age, although her progress compared to normal peers worsened the older she got. She never had any obvious setbacks that correspond with vaccines."
No	"We have two sons, the oldest has Asperger's, the younger one does not. While some children may be affected by the vaccines, our son did not begin to show any clear signs of Asperger's until right before kindergarten. Because he has mild Asperger's, it may be the social bit did not really matter until then, as preschool requirements are not very high, and he has never been disruptive. However, we did not notice any correlation to the vaccine in the way some parents do."
(15.2b) Statements about Predisposing and Vaccine Interactions as Causes of Autism:	
Partially	"We believe a genetic predisposition combined w/many assaults on our child's immune system beginning w/antibiotics via the birth canal and Hep B at birth followed by many other vaccines including MMR caused our child's autism. Genetics alone would not have caused our son to reach the severity he was at when he was 2 y.o. My husband has MANY autistic traits and would probably be classified as autistic if evaluated. As a child, I had many sensory issues that I have learned to overcome but still carry w/me to a certain extent. These 2 factors would not have caused my son to be in a vegetative state. That had to have been caused by the external/environmental factors mentioned above."
Partially	"When my son was born, they used a vacuum extractor to help remove him from the birth canal. According to this lady, this could have been detrimental to him, as it causes the head to expand when it should be contracting, bursts blood vessels, and causes an extremely important bone near the eyes to peak downward instead of upward as it is supposed to. This floored me, as NO ONE has ever told me that a vacuum extractor has any side effects . . . our son is constantly needing pressure on his head. . . . I think vaccines did have a role in causing our son to become autistic with perhaps the vacuum extractor being the original problem for causing things to go wrong."

Source: Immunization Poll of AutismCoach Customers, Updated 14 October 2004. Downloaded from www.AutismCoach.Com, circa 10 September 2009.

Panel 15.2b indicates that some parents think that clearly identified initial problems interact with the vaccines to produce the autism:


Age of Child → Vaccinations → Autism-Spectrum Disorders (Model 4)
Initial Problems

One parent believes that a genetic predisposition interacted with the vaccines to cause the child's autism; the other parent believes that trauma at birth induced by a vacuum extraction of the fetus interacted with the vaccines to cause the child's autism. However, if vaccinations do not directly cause the autism-spectrum disorders, as research studies indicate, then there cannot be a statistical interaction effect between the vaccinations and the initial problems; the direct effect of the vaccinations on the outcomes must be included in the statistical model and it is zero; this simplifies the model implying that the initial problems are associated with the autism.

Because the same child both cannot receive and also not receive the vaccination at the same time, this fundamental problem of causal inference limits the evaluation of causality. Circumventing this limitation, some parents compare their children's reactions. One parent in this survey related that both of her children were vaccinated, but only one child developed autism (i.e., vaccination does not always cause autism). Contrarily, another parent related that one of her children had been vaccinated and developed autism, but her other children, who were not vaccinated, did not develop autism (i.e., vaccination causes autism). In both cases the genetic background was roughly controlled, but the relationship between vaccination and autism was different. Without a verified biological mechanism linking the vaccinations to the outcome, the causal effects for an individual child are near impossible to determine, but average causal effects can be estimated and have been, with the result that Model 3 mostly likely is the preferred causal model; initial problems and not vaccinations engender the autism-spectrum disorders.

Methods for Reviewing Evidence

The European and American reviews of the evidence about the supposed linkage between childhood vaccinations and autism reached this same conclusion: the empirical evidence does not support a causal linkage between childhood vaccinations and autism-spectrum disorders. European health scientists conducted the review for the Cochrane Collaboration (Demicheli et al. 2005) by applying a meta-analysis paradigm somewhat similar to that of Chapter 14. American health scientists conducted the reviews for the Institute of Medicine (Institute of Medicine 1994, 2001a, 2001b, 2004) by applying an informal Bayesian approach in which they initially took a neutral position about causality and then modified their beliefs on the basis of the evidence; this approach informs that of the concluding chapter of this book.

The Cochrane Collaboration's Meta-Analytic Approach

The international health scientists of the Cochrane Collaboration aim to help people make well-informed decisions about health-care alternatives by preparing, maintaining, and promoting the accessibility of systematic reviews of the effects of health-care interventions. Systematic reviews attempt to answer a specific research question by collating all the empirical evidence on a topic that fits prespecified eligibility criteria. This paradigm gives a high priority to the results of randomized trials and to controlled epidemiological studies that study change in treatment and control groups in time periods before and after an intervention. Table 15.3 presents an abridged version of the main topics of a typical Cochran Collaboration report; these topics are those of a scientific paper, but there also is a plain language summary for the lay reader. The typical abstract of a review highlights these topics, which the report elaborates: Background, Objectives, Search Strategy, Selection Criteria, Data Collection and Analysis, Main Results, Author's Conclusions, and Plain Language Summary.

The objectives of the review set the criteria for the selection of studies and guide the actual selection process. For example, the objectives of the review by Demicheli et al. (2005, 3–5) limited the selection of studies to those that reported the effectiveness and unintended effects of the combined MMR vaccinations. The reviewers included all comparative prospective or retrospective studies of healthy individuals up to 15 years of age. The selection criteria specified any report in which the vaccination used the combined MMR compared with singular vaccinations, do-nothing, or placebo; and with such outcomes as clinical cases of measles, mumps, or rubella; systematic adverse events ranging from fever to Crohn's disease, ulcerative colitis, autism, and aseptic meningitis; and such local adverse events as soreness and redness at the site of the inoculation.

The investigators' initial broad selection criteria for effectiveness and safety identified about 5,000 studies for screening from such databases as the Cochrane Central Register of Controlled Trials, MEDLINE, EMBASE, Biological Abstracts, Cochrane Database of Systematic Reviews (CDSR), the National Health Service (NHS) Database of Abstracts of Reviews of Effects (DARE), the Science Citation Index, and so forth. From these 5,000 reports, teams of two persons retrieved 139 candidate studies that possibly fulfilled all of the selection criteria. Of these, the investigators eliminated 108: some because the data sets were used several times resulting in redundant publications; and the vast majority because they did not meet all of the inclusion criteria; their final review is based on 31 studies that varied in their potential biases and generalizability. The designs of these 31 studies include five randomized controlled trials (RCTs), one controlled clinical trial (CCT), 14 cohort studies, five case–control studies, three time-series trials, one case-crossover trial, one ecological trial, and one self-controlled case-series trial.

The investigators based their conclusions on their valuation of the results from studies with a low risk of bias within the diverse types of study designs,

Table 15.3 Organization of a Cochrane collaboration review (Abridged)

Abstract: Background, Objectives, Search strategy, Selection criteria, Data collection and analysis, Main results, Author's conclusions.

Plain Language Summary: Plain language title, Summary text

The review:

- Background
- Objectives
- Methods
  - Criteria for considering studies for this review
    - Types of studies
    - Types of participants
    - Types of interventions
    - Types of outcome measures
  - Search methods for identification of studies
  - Data collection and analysis
    - Study selection
    - Quality assessment
    - Data extraction
  - Statistical considerations
- Results
  - Description of studies
  - Risk of bias of included studies
  - Effects of interventions
- Discussion
- Authors' Conclusions
  - Implications for practice
  - Implications for research
- Acknowledgments
- References
  - References to studies
  - Other references
- Tables and figures
  - Characteristics of studies
    - Characteristics of included studies (includes "Risk of bias" tables)
    - Characteristics of excluded studies
  - "Summary of findings" tables
  - Additional tables
  - Figures

Supplementary information:

- Data and analyses, Appendices
- Feedback: Title, Summary, Reply, Contributors

About the article:

- Contributions of authors, Declarations of interest, Sources of Support

Source: Box 2.2b of *Cochrane Handbook for Systematic Reviews of Interventions,* edited by Julian Higgins and Sally Green, version 5.0.1, updated September 2008.

interventions, and measures of outcomes. Their plain language summary states these conclusions (Demicheli et al. 2005, 2): "MMR protects children against infections of the upper airways but very rarely can cause a benign form of bleeding under the skin and milder forms of measles, mumps and rubella. No credible evidence of an involvement of MMR with either autism or Crohn's disease was found." This latter conclusion directly contradicts the article by Wakefield et al. (1998). Because the investigators' selection criteria did not focus on the effects of thimerosal, they did not assess the controversial studies of Geier and Geier (e.g., 2003a, 2004), which they characterize (2005, 41) as having an "Uncertain MMR focus, mixed with thimerosal."

However, the review committee of the Institute of Medicine (2004, Table 10, 86–110 and discussion 82–85, 110–126) found the studies of Geier and Geier (2003c, 2004) and Wakefield et al. (1998) to be seriously flawed, and that the evidence from the other studies favors the rejection of a causal relation between MMR vaccinations and autism. To reach this conclusion, which is the same as that of their earlier report (Institute of Medicine 2001a), and that of Cochrane Collaboration, the committee examined new studies that reported no association and found their designs, methods, and results to be more persuasive than the not-credible evidence from the three methodologically flawed studies that reported a causal relationship. The epidemiological studies showing no causal effect of MMR on autism included nine controlled observational studies, three ecological studies, and two studies based on the passive reporting system of Finland. Thus, both the European and American reviewers found that the evidence does not support a causal association between vaccinations using the MMR vaccine and autism-spectrum disorders.

The Institute of Medicine's Bayesian Approach

The Institute of Medicine (IOM) convened an immunization safety review committee initially composed of 13 expert health scientists, none of whom had a conflict of interest regarding childhood vaccinations, and asked them to review and evaluate the evidence for or against a causal association between the MMR vaccine and autism, and between thimerosal-containing vaccines (TCVs) and autism. Following an informal Bayesian approach, the panel initially began with a position of neutrality regarding the safety of the vaccines, making no presumption that a specific vaccine or vaccine component does or does not cause autism or other adverse events. They examined the results from empirical studies of varying designs and quality and, on the basis of their assessments; they classified the evidence according to its level of persuasiveness using the descriptive statements of Table 15.4. These statements range from "No evidence bearing on a causal relation" to "The evidence establishes a causal relationship." For each statement, this scheme specifies the criteria that studies must meet in order to achieve a designated level of evidence. For determining causality the committee prefers controlled epidemiological studies and randomized trials.

Table 15.4 The Institute of medicine's statements about levels of evidence

Statements about results	Levels of evidence
"No Evidence bearing on a causal relation."	"No case reports or epidemiological studies identified."
"The evidence is inadequate to accept or reject a causal relation."	"One or more case reports or epidemiological studies were located, but the evidence for the causal relation neither outweighs nor is outweighed by the evidence against a causal relation."
"The evidence favors rejection of a causal relation."	"Only evidence from epidemiological studies can be used as a basis for possible rejection of a causal relation. Requires a rigorously performed epidemiological study (or meta-analysis) of adequate size that did not detect a significant association between the vaccine and the adverse event."
"The evidence favors acceptance of a causal relation."	"The balance of evidence from one or more case reports or epidemiological studies provides evidence for a causal relation that outweighs the evidence against such a relation."
"The evidence establishes a causal relation."	"Epidemiological studies and/or case reports provide unequivocal evidence for a causal relation."

Institute of Medicine (2001a, 19). Also see Institute of Medicine (2004, 25).

Selection of studies

Following a search-and-retrieval procedure similar to that of the Cochrane Collaboration, the IOM's librarians searched online data bases circa 1992–1993 including those of the National Library of Medicine, MEDLINE, EMBASE, BIOSIS, the Vaccine Adverse Event Reporting System (VAERS), and so forth. Of the 8,000 initial citations about 1,600 were relevant to the committee's work and composed an interim bibliography that formed the basis of the 1994 report (Institute of Medicine 1994, 318–322). The subsequent reviews focused on new clinical and epidemiological studies that probed the hypothesized linkages between MMR vaccinations and autism (Institute of Medicine 2001a, 2004) and between vaccinations with thimerosal-containing vaccines and autism (Institute of Medicine 2001b, 2004). The committee valued published, peer-reviewed studies more highly than unpublished studies, and controlled studies more highly than descriptive studies, but they aimed to include all relevant studies in their reviews, even those that reused slightly modified databases.

Causality and study designs

The IOM's pivotal conception of "Can It?" or synonymously "potential" causality asks (Institute of Medicine 1994, 20–23; Kramer and Lane 1992): "Can the vaccine cause the adverse event at least in certain people under certain circumstances?" Providing answers to this research question, the reviewers evaluated evidence from descriptive studies, controlled epidemiological studies, and randomized trials. Regardless of the study design, the Institute of Medicine (1994, 28–31) stipulated that a plausible underlying biological mechanism should provide a rationale for any association between vaccinations and adverse events, at least theoretically. If no association is found, then a linking biological mechanism cannot be invoked to prove causality.[7]

Because descriptive studies lack a control group, such studies as case reports, case series, and uncontrolled observational studies usually provide evidence for causality that is less persuasive than that from controlled epidemiological studies. But if at least one case unequivocally shows that an adverse event was caused by a vaccine, then "Can It?" causality is satisfied even without controlled epidemiological studies. [8]

Because practical or ethical reasons prevent randomization, epidemiologists seldom are able to randomly assign children to treatment groups exposed or not exposed to a vaccine. Consequently, they provide answers to their research questions by conducting well-controlled observational studies, and by calculating the relative risk as the ratio of the incidence rate of the adverse event among those exposed to the vaccinations to the incidence rate among those not exposed to the vaccinations.[9] An incidence rate is defined as the number of new cases of a disease occurring in a specified time period divided by the total number in the population at risk during that time period. The incidence rate for an adverse event due to an MMR vaccination is the number of new adverse events during the time period divided by the number of MMR vaccinations during that time period (Institute of Medicine 2004, 59). The attributable risk is the difference between such incidences. In case-control studies no direct calculation of relative risk or risk difference can be made. Instead, the exposure odds ratio is calculated as the odds of exposure among the cases divided by the odds of exposure among the controls; it is a very good estimate of the true relative risk when adverse events are rare (Institute of Medicine 1994, 29).

Epidemiological study designs include controlled cohort studies that have exposed and unexposed subjects to the vaccinations; cohort studies comparing adverse events at different time points after a vaccination; case-control studies that compare the rates of prior exposure to the suspected vaccine between children with the adverse event (the cases) and without the adverse event (the controls); ecologic studies that compare the rates of adverse events in areas that have different policies for administering the vaccinations; and meta-analyses. Of course, randomized trials in principle provide the most persuasive evidence, but the adverse events in question are rare and may not appear except in experiments of large scale, which may be impractical or unethical.

Criteria for critiquing studies

The committee of health experts reviewing the studies primarily applied criteria similar to those of Hill (1965) and Susser (1973), which Chapter 3 discussed earlier under the heading "Causality as Robust Dependence." The IOM report (1994, 21–22) lists these six italicized considerations for assessing the "Can it?" causality of the relationship between vaccination (v) and adverse events (a); $v \rightarrow a$.

1. *Strength of association*: The greater the relative risk due to the vaccinations, the less likely the $v \rightarrow a$ relationship is spurious or the result of various biases.
2. *Analytic bias* is minimal: The time order is correct, a does not cause v. Information bias has not affected the result, the variables are measured appropriately and the effects of unblinding, recall bias, and unequal surveillance of vaccinated and unexposed subjects are minimal. The relationship $v \rightarrow a$ holds when a number of relevant test factors T (including selection bias) are controlled, $av.T \neq$ no relationship ($r \neq 0$ or relative risk >2).
3. The *dose–response* relationship indicates that higher doses lead to more adverse events.
4. The $av.T$ relationship is *statistically significant* and not due to chance.
5. *Consistency*: the $av.T \neq$ no relationship replicates; it is found in a number of studies using different populations and locales.
6. *Biologic plausibility and coherence*: The $av.T \neq$ no relationship is consistent with available knowledge based on human and animal experiments, and a well-researched biological mechanism could explain the $v \rightarrow a$ relationship theoretically.

Assessments of thimerosal-containing vaccines (TCVs)

Synthesizing the earlier appraisal (Institute of Medicine 1994, 2001b) with their own; the review committee's report (Institute of Medicine 2004, 7) states that: "the committee concludes that the evidence favors rejection of a causal relationship between thimerosal containing vaccines and autism." This modifies the earlier report's conclusion (Institute of Medicine 2001b, 66) that the evidence was "inadequate to accept or reject a causal relationship between exposure to thimerosal from childhood vaccines and the neurodevelopmental disorders of autism, ADHD [attention deficit hyperactivity disorder], and speech and language delay." This earlier more equivocal conclusion resulted from the scarcity of relevant studies and the wide-ranging response variables composing neurodevelopmental disorders (NDD), rather than just autism.

To determine whether the earlier equivocal conclusion still held, the committee evaluated new epidemiological evidence provided by three controlled observational studies and two uncontrolled observational studies, all of which were rigorously peer-reviewed. Table 15.5 presents an abridged summary of the evidence from these epidemiological studies. It illustrates the format of the committee's comprehensive Table 9 (Institute of Medicine 2004, 66–83) that summarized the evidence from all of the studies they reviewed, bearing on TCVs and autism. The column headings of these tables specify the study, the design, the country and population of children, the vaccine exposure, the outcomes of the study, the results, the reviewers'

Table 15.5 Abridged evidence table, epidemiological studies find thimerosal-containing vaccines (TCVs) unrelated to Autism

Citation	Design	Country	Vaccine exposure	Outcomes	Results	Reviewers' comments	Contribution to causality argument
Controlled observational studies							
Hviid et al. (2003)	Cohort (Time Series)	Denmark, 467,450 children	Thimerosal, No thimerosal	Autism-spectrum disorders	No significant effects, trend Increases	Strong internal validity	TCV and autism not related
Verstraeten et al. (2003)	Retrospective Cohort	USA, 110,833 children	Cumulative thimerosal exposure at 1, 3 and 7 months	Autism and developmental problems	No cumulative dose–response effect	Autism diagnosis is valid	TCV and autism not related
Miller (2004)	Cohort	Great Britain	Cumulative doses by 3, 4 and 6 months	Autism ICD-9 codes	No cumulative dose–response effect	Increased exposure does not increase autism risk	TCV and autism not related
Uncontrolled observational studies							
Madsen et al. (2002)	Ecological	Denmark, children diagnosed with autism	Thimerosal, No thimerosal	Autism-spectrum disorders	Increasing secular trend after TCVs eliminated	Increase may be due to inclusion of outpatients	TCV and autism unrelated, but study design is limited
Stehr-Green et al. (2003)	Ecological	Denmark and Sweden, children diagnosed with autism	Thimerosal, No thimerosal	Autism-spectrum disorders	Increasing secular trend after TCVs eliminated	Increase may be due to changing diagnostic criteria	TCV and autism unrelated, but study design is limited

Institute of Medicine (2004, Table 9, 66–83; and discussion, 42–65). This abridged table only briefly summarizes the complete information.

comments on the study, the committee's judgment about the study's contribution to the causality argument. All five of these epidemiological studies found that vaccinations with thimerosal-containing vaccines are unrelated to autism.

Contrariwise, in two ecological studies of birth cohorts and three comparative studies based on passive reporting systems of data, Geier and Geier claim that exposure to thimerosal-containing vaccines is causally associated with autism-spectrum disorders. Table 15.6 presents an abridgment of the review committee's analysis of their ecological studies. They conclude that the methodological flaws of these studies make their results uninterpretable and not persuasive with respect to causality: The investigators have questionable data and measures at the aggregate level, and from these data and measures they attempt to make causal inferences for individual children. The aggregate data make questionable their estimates of the vaccination dosages that the children actually received and its linkage to the investigators' flawed measure of autism.

During the time period of the study the reporting criteria about autism changed: Prior to 1990 the children in the 1981–1984 and 1984–1985 birth cohorts with autism were reported in terms of a different disability category. In 1990 autism was added to the reporting categories thereby artificially increasing the number of children with autism in the 1990–1996 and 1990–1994 birth cohorts. Consequently, the later birth cohorts exhibit higher levels of autism. Regardless of such other problems as vague analytic methods and inappropriately defined measures, this difference in categorizing children with autism invalidates both of the investigators' ecological studies. The IOM report (2004, 55–58) analyzes these flawed studies.

Given that these ecological studies are intrinsically flawed because the changed diagnostic categories increased the number of autism cases in the later birth cohorts, the studies based on data from the passive reporting systems are also intrinsically flawed at least because the cross-sectional treatment groups presumably with higher (87.5 μg) and lower (37.5 μg) amounts of mercury are not valid, thus making any relationships between higher doses and autism and other adverse events fallacious. Table 15.7 presents an abridgment of the review committee's analysis of evidence from the three comparative studies of Geier and Geier; all of these studies are based on inappropriate data and methods. Regarding the first study, the reviewers state (Institute of Medicine 2004, 58): "They do not explain how they categorized individuals into the two exposure groups, nor is it clear how they could." Regarding the second study, the reviewers state (IOM 2004, 59): "Reports of adverse events were again categorized as either following an average exposure of 37.5 μg of mercury or an average exposure of 87.5 μg of mercury, but again no information was provided on how these averages and the exposure groups were derived." Apparently, the third study also is marred by questionable dosage groups.

All three studies suffer from the use of the Vaccine Adverse Event Reporting System (VAERS). Although this database may be useful for developing hypotheses for later testing, because of its voluntary, unmonitored, and unrepresentative submissions, it cannot be used appropriately to estimate incidence, relative risk, and prevalence, and to make conclusions about causality. The review committee states (IOM Institute of Medicine 2004): "VAERS has inherent limitations that include

Table 15.6 Abridged evidence table, Geier and Geier studies find thimerosal-containing vaccines (TCVs) related to autism, ecological data

Citation	Design	Population	Vaccine exposure	Outcomes	Results	Reviewers' Comments	Contribution to causality argument
Ecological studies							
Geier and Geier (2004)	Indeterminate. Classified as ecological because relies on aggregated data	Birth cohorts 1981–1984 and 1990–1996	Number of vaccines distributed from Biologic Surveillance Summaries (BSS)	Autism as reported by Dept. of Education	Prevalence of autism plotted against the estimated average mercury dose. Slope, R^2, and odds ratios are ambiguous	Study has serious methodological flaws (IOM 2004, 55–58, 80–81): prevalence flawed, grouping questionable, level of analysis confused, dosages ambiguous, analytics unclear	Results are uninterpretable and therefore noncontributory with respect to causality
Geier and Geier (2003a)	Indeterminate. Classified as ecological because it relies on aggregated data to make inferences at the individual level	Birth cohorts 1984–1985 and 1990–1994 1984 baseline year	Estimated average amount of mercury each child presumably received from Biologic Surveillance Summaries (BSS)	Autism, speech disorders, orthopedic impairments, visual impairments, deaf-blindness as reported in the 2001 Dept. of Education report	Autism odds ratio = 2.5, it increased by 0.014 μg of mercury Speech disorders OR = 1.4, it increased by 0.12 μg of mercury	Study has serious methodological flaws (IOM Institute of Medicine 2004, 55–58, 78–79): prevalence flawed, grouping questionable, level of analysis confused, dosages ambiguous, analytics unclear	Results are uninterpretable and therefore noncontributory with respect to causality

Institute of Medicine (2004, Table 9, 78–81; and discussion, 55–58). This abridged table only briefly summarizes the complete information.

Table 15.7 Abridged evidence table, Geier and Geier find thimerosal-containing vaccines (TCVs) related to Autism, passive reporting data

Citation	Design	Population	Vaccine exposure	Outcomes	Results	Reviewers' comments	Contribution to causality argument
Studies based on passive reporting of data							
Geier and Geier (2003d)	Passive Reporting System, CDC's Biologic Surveillance Summaries (BSS)	USA, VAERS reports of autism and speech disorders	TCVs DTaP, and DTwP, 1992–2000 vs. no thimerosal DTaP, 1997–2000	Autism and speech disorders for ambiguous groupings of average exposures of 37.5 µg mercury vs. 87.5 µg	Increased relative risks of autism and speech disorder with increased dosage of mercury from TCVs	Study has serious methodological flaws (IOM Institute of Medicine 2004, 58–59): the "incidence rates" are incorrect in the numerator and denominator	Results are uninterpretable and therefore noncontributory with respect to causality
Geier and Geier (2003a)	Passive Reporting System, CDC's Biologic Surveillance Summaries (BSS)	USA, VAERS, Cases: autism, personality disorder, retardation. Controls: seizures, fevers, pain edema, vomiting	TCVs DTaP &DTwP, 1997–2000 vs. no thimerosal DTaP, 1997–2000	Autism and speech disorders for ambiguous groupings of average exposures of 37.5 µg mercury vs. 87.5 µg	OR autism = 2.6 and it increased by 0.029 per µg of mercury	Study has serious methodological flaws (Institute of Medicine 2004, 59–60): the comparison groups and R^2s are questionable	Results are uninterpretable and therefore noncontributory with respect to causality
Geier and Geier (2003c), also see (2003b)	Passive Reporting System, CDC's Biologic Surveillance Summaries (BSS)	USA DTaP TCV vaccinees, 1992–2000 vs. DTaP vaccinees, 1997–2000	Thimerosal (1992–2000), no thimerosal (1997–2000) Autism "Incidence" erroneous	Cases: autism, retardation, speech disorders. Controls: deaths, seizures, etc.	Flawed measures of relative and attributable risks and association	Study has serious methodological flaws (IOM 2004, 60–62)	Results are uninterpretable and therefore noncontributory with respect to causality

Institute of Medicine (2004, Table 9, 66–83; and discussion, 55–65). This abridged table only briefly summarizes the complete information.

variability in reporting standards, reporting bias (e.g., due to other factors such as media attention), unconfirmed diagnoses, lack of information on people who were immunized but did not report an adverse event, lack of an unbiased comparison group, and variable and potentially significant underreporting of adverse events." Consequently, the outcome measures of these three studies are also flawed, as are the investigators' statistical measures of effects. The IOM report (2004, 58–62) explicates the flaws in these studies.[10]

Because the methodologically sound epidemiological studies found that TCVs are unrelated to autism, and all of the studies purportedly showing a linkage are seriously flawed (Institute of Medicine 2004, 7): "Thus, based on this body of evidence the committee concludes that the evidence favors rejection of a causal relationship between thimerosal-containing vaccines and autism."

The meta-analytic review of the evidence conducted by the Cochrane Collaboration and the informal Bayesian review of the evidence conducted by the Institute of Medicine have established that childhood vaccinations with the MMR vaccine are unrelated to autism and autism-spectrum disorders, and that vaccinations with thimerosal-containing vaccines are also unrelated to autism. The public's lack of sophistication in assessing causality coupled with the financial interests of unconscionable people propagandizing a spurious linkage between these childhood vaccinations and autism has led to a dangerous decline in immunization coverage, especially in the UK where vaccinations are at the discretion of parents. Hopefully, these adverse trends are reversing because of the evaluations of evidence conducted by these reviewers and the dissemination of their results through the mass media.

To provide further practice in assessing assumed causal relationships, the following concluding chapter draws on the methods for reviewing evidence of these health scientists. For the chapters presenting multilevel modeling, it locates the causal aspects of the evidence as indicating no causality, robust dependence, potential outcomes, or dependency networks. Earlier, Chapters 3 and 4 defined and illustrated these notions of causality. Studies in the social science seldom meet the most stringent causal criteria but such studies can aspire to reach a relevant zone of causality. The three zones of causal notions and their component subtypes are designed to encourage social problems researchers to think about the causal aspects of their findings and to consider ways of improving their studies and strengthening causal inferences.

Endnotes

[1] At first glance this controversy is similar to the much earlier controversy stemming from Ignaz Semmelweis's (1861) finding that agents (i.e., germs) transmitted by physicians were causing the deaths of mothers soon after their giving birth. The medical establishment in Vienna neither appreciated his findings nor his recommendation that the physicians wash their hands in a chlorine solution, which they followed only sporadically. He resigned his position in Vienna, moved to a less central position, and later on his wife committed him to an insane asylum, where he died. Wakefield challenged the medical establishment, was forced to resign his position, and became

marginal. However, Semmelweis was correct and his recommendation saved lives; Wakefield is incorrect and his recommendation has led to the outbreak of preventable diseases and the deaths of children.

[2] Randomized controlled trials have established that trivalent MMR vaccinations are not more likely to have adverse consequences than the singular vaccinations. Demicheli et al. (2005, 6–8) reviewed six such trials; five preceded the publication of the article by Wakefield et al. (1998). They also reviewed 14 cohort studies that reached the same conclusion; that is, no difference between the trivalent and the single-disease vaccination; nine of these cohort studies where published prior to Wakefield et al. (1998). That article should have taken into account the findings of these earlier studies. No credible empirical research then supported, or now supports, Wakefield's alarming statement. Consequently, ten of the original 13 authors have stated that their data were insufficient to establish a causal link between MMR vaccine and autism (Murch et al. 2004) and *The Lancet* has retracted the article (Editors, 2 February 2010).

[3] Brian Deer states (8 February 2009, e6–e7):

> What parents did not know was that, 2 years before [the study began], Wakefield had been hired by Jabs's lawyer, Richard Barr, a high-street solicitor in King's Lynn, Norfolk. Barr had obtained legal aid to probe MMR for any evidence that could be used against the manufacturers. He is adamant that at all times he acted professionally, and diligently represented his clients. A string of Sunday Times reports have exposed how Wakefield earned £435,643 through his work with Barr, plus funding to support his research. . . .
> The objective," they wrote, "is to seek evidence which will be acceptable in a court of law of the causative connection between either the mumps, measles and rubella vaccine or the measles/rubella vaccine and certain conditions which have been reported with considerable frequency by families who are seeking compensation.

[4] Plotting data from the World Health Organization (WHO) and UNICEF, McCartney (2009, 10) documents the decline and recovery of measles immunization rates in the UK and globally. For the UK the approximate percentages (read from her graphs) are 1998 = 87%, 2001 = 84%, 2004 = 81%, and 2008 = 86%. She attributes this decline to Wakefield's "infamous" 1998 article and his press conference where he claimed the MMR vaccine could cause bowel disorders and, possibly, autism. She notes that all but three of the 13 authors retracted their article's interpretation of the hypothesized linkage between MMR vaccinations and autism (Murch et al., 2004). Global immunization also declined and recovered during this period reaching a higher percentage than in the UK (also read from her graph): 1998 = 80%, 2001 = 72%, 2004 = 85%, and 2008 = 87%. She notes that an outbreak of measles occurred in Dublin in 2000 when the coverage rate was only 79%, much less than the 95% coverage needed to prevent such an outbreak. Consequently, 13 children needed intensive care and another three died because of the measles. A former chairman of the British Medical Association opined that parents in the UK should be compelled to vaccinate their children rather than leaving vaccinations to their discretion.

[5] Brian Deer (8 February 2009, e6) reports that: "Wakefield has left Britain to live in Austin, Texas, where he runs a clinic offering colonoscopies to American children. He tours the country, giving lectures and speeches against the vaccine, and attracting a loyal following of young mothers."

[6] Evidence of how AutismCoach propagates misinformation is provided by its advocacy of the flawed research of David A. Geier, B.A. and Mark R. Geier, M.D (viewed on the AutismCoach web site, 17 September 2009). These researchers purportedly show that the switch to vaccines without thimerosal (a mercury-based preservative) has reduced the occurrence of autism; that is, the mercury previously in the vaccines had caused autism. The AutismCoach web site does not mention that the Institute of Medicine characterizes the Geier studies as uninformative regarding causality, and that the body of evidence favors rejection of a causal relationship between thimerosal-containing vaccines and autism (Institute of Medicine 2004, 6–7, and discussion of findings in this chapter).

[7] The review committee specified three categories of evidence about biological mechanisms: plausible theoretical evidence, experimental evidence on humans or animals, and evidence that the mechanism results in known human diseases (Institute of Medicine 2004, 29).

[8] The IOM defines two other aspects of causality. *Did it*? causality asks if the "evidence strongly suggests that the vaccine *did cause* the adverse event in one or more cases, then it is logical to conclude that it can cause the event" (Institute of Medicine 2004, 23). *Will it*? causality "refers to how frequently a vaccine causes a specific adverse event and can relate to either individuals or populations. … For either individuals or populations, the answer to *Will it*? is best estimated by the magnitude of the *risk difference* (*attributable risk*): the incidence of the adverse event among vaccine recipients minus the incidence of the adverse event among otherwise similar nonrecipients" (Institute of Medicine 2004, 27).

[9] Susser (1973, 87) defines relative risk as follows: Relative risk is the ratio of the incidence of cases among those exposed to the risk or causal factor to the incidence among those not exposed. The relative risk of lung cancer for smokers is the ratio of the rate of lung cancer in smokers to the rate in nonsmokers. Stated another way, this risk is ratio of the incidence observed in smokers to the [counterfactual] incidence to be expected among them had they experienced the rates of nonsmokers.

[10] The Wikipedia entry for the Geiers suggests that they have a financial interest in propagating the spurious linkage between vaccinations and autism:

> The Geiers have developed a protocol for treating autism that uses the castration drug Lupron. Mark Geier has called Lupron "the miracle drug" and the Geiers have marketed the protocol across the U.S. … According to expert pediatric endocrinologists, the Lupron protocol for autism is supported only by junk science. … When treating an autistic child, the Geiers order several dozen lab tests, costing $12,000: if at least one testosterone-related result is abnormal, the Geiers consider Lupron treatments, using 10 times the daily dose ordinarily used to treat precocious puberty. The therapy costs approximately $5,000 per month. The Geiers recommend starting treatment on children as young as possible, and say that some need treatment through adulthood (Downloaded 3 October 2009 from http://en.wikipedia.org/wiki/Mark_Geier). Also see on the internet "Confessions of a Quackbuster Mark Geier Untrustworthy: Autism, Thimerosal, Vaccinations" (Downloaded 3 October 2009).

Chapter 16
Gauging Causality in Multilevel Models

If contributions made by statisticians to the understanding of causation are to be taken over with advantage in any specific field of inquiry, then what is crucial is that the right relationship should exist between statistical and subject matter concerns.

—John H. Goldthorpe (1998, 28)

Studying causation will often require a blend of mathematical analysis, substantive knowledge, and critical reasoning. Competing theories will need to be proposed and tested. Case studies and qualitative description can serve as a necessary complement to quantitative data and mathematical modeling. Some of the most compelling and influential research manages to "paint a picture" by combining descriptive data analysis, statistical modeling, in-depth narratives, and background knowledge.

—Herbert I. Weisberg (2010, 306)

This book illustrates strategies for the development and testing of multilevel models bearing on social problems, all of which deal directly or indirectly on aspects of human development, measured by social and economic indicators. It confronts social problems by ideally following these five steps: analyze the roots of the social problem both theoretically and empirically; formulate a study design that captures the nuances of the problem; gather empirical data bearing on the social problem that enable the design to be operationalized by forming identifiable and repeatable measures; model the multilevel data using appropriate multilevel statistical methods to uncover potential causes and any bias to their effects; and use the results to sharpen theory and to formulate evidence-based policy recommendations for implementation and testing. Applying this process, the core chapters present multilevel models focusing on political extremism, global human development, violence against minorities, the substantive complexity of work, reform of urban schools, and problems of health care. The reader will be better able to conduct state-of-the-art studies on these and other topics by gaining an understanding of these chapters and by using the available data sets and analytic programs to replicate and advance the analyses.

The chapters develop causal inferences about the effects of key stimulus variables on response variables. Most simply, a causal effect results from changing one

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9_16, © Springer Science+Business Media B.V. 2011

variable by external manipulation while keeping others constant; that is, $x \rightarrow y$, all other things being equal (Heckman and Vytlacil 2001, 16). Even so, notions of causality entail rather complex multidimensional concepts. Cox and Wermuth (2001, 65–70) define three levels of causal notions: level-zero, level-one, and level-two.[1] Closely paralleling their distinctions, Chapters 3 and 4 discussed these three notions of causality zones: *Stable association* (i.e., level-zero causality) includes classic causality, where the survey analyst controls for the effects of test factors one at a time; and robust dependence, where the statistically inclined analyst controls for the effects of several test factors simultaneously. *Potential outcomes* (i.e., level-one causality) includes causality as an effect of an intervention and stochastic causal models for attributes. *Dependency networks* (i.e., level-two causality) includes graphical models, association graphs for loglinear models, generative process models, and structural economic models in policy research.

Although these concepts of causality are crisply defined in the literature, the examples in this book do not crisply conform to these criteria, there are areas of ambiguity about the causal aspects of the findings. For these examples and for the social sciences in general, because of the ambiguity due to their imprecision, it is appropriate to think in terms of zones of causal notions rather than precise levels of causality. Moreover, a valid model can produce findings indicative of *no causality*, usually because of the failure to achieve statistical significance; while an invalid model will produce findings that are *not informative as to causality.*

To provide practice in assessing causality, in this concluding chapter I review each multilevel model for any invalidities that limit or negate causal inferences, and then classify the study according to the zone of causality that it may have achieved. In doing so, I apply an informal Bayesian approach, taking an initial position of neutrality with respect to causality, and then on the basis of my explication perhaps changing it. In this chapter I often use the first person because I applied the procedures of these chapters and formed my opinions about the level of causality of the results; I invite the reader to form his or her own judgments about causality, and to disagree with mine.

Zones of Causal Notions

The categories of causality are grouped under the headings “not informative as to causality,” “no causality,” “stable association,” “potential outcomes,” and “dependency networks.”

Two Meanings of Non-Causality

If a study is not conducted with competence and does not exhibit sufficient validities, then its findings are suspect and cannot contribute to the causality argument (Institute of Medicine 1994, 32–33); it is classified as *not informative as to causality*. However, a valid study can reject a causal association producing a finding of *no causality*; for example, it may find (as did Chapter 15) no statistically or substantively significant causal association between childhood vaccinations and autism-spectrum disorders. To be viewed as competently conducted, which is a prerequisite for making appropriate causal (or not-causal) inferences, quantitative studies in the social sciences should exhibit the following five validities that form a hierarchy (Smith 2008b, 136–139; Campbell and Stanley 1963):

Fit validity refers to the appropriateness of the concepts of the study and their relationships to the empirical social world. It asks: Does the conceptual framework include relevant concepts and not exclude those that should be included? Insights from qualitative studies and situation-specific knowledge can enhance the richness of the conceptual framework, thereby improving fit validity.

Construct validity focuses on the relationship between the concepts and their measures. It assesses how well the study operationalizes the theoretical dimensions of the concepts or interventions into measures. For stimulus variables it asks: Is the assumed cause properly defined and operationalized? For control and response variables it asks: Do the operational devices measure without bias what they are supposed to measure? Do these measures have fit validity as well?

Internal validity focuses on the design of the study, assessing the extent to which the research design as implemented is free of contamination from other implementations and from the biased selection of participants. It asks: Is the measured effect free of contamination from other measures or events? Are the quantified relationships between the stimulus and response variables free of selection bias? If there is a comparison group, is it appropriate? Has the research fulfilled the stable-unit-treatment-value assumption (SUTVA)?

External validity indicates the extent to which the research findings can be generalized; it assesses the quality of the samples, overall and for important subgroups. It asks: In what settings, time periods, studies, or populations can the findings of the research be expected to hold?

Statistical conclusion validity refers to the appropriateness of the statistical models that quantify the relationships and their statistical significance. It asks: Are the calculated effect sizes appropriate given the nature of the measures whose relationship they quantify? Do the statistical methods limit the generality of the inferences; that is, to what statistical universe can the results be appropriately generalized? Are the statistical tests fully reported? Does the lack of statistical significance of a relationship lead to a finding of no causality?

Given these criteria it suffices here to rate a study subjectively on each of the five validities as acceptable, limited, or unacceptable, keeping in mind the various questions that determine the validities. Because these validities form a hierarchy, fit validity, construct validity, and internal validity are essential; if a study lacks these validities it is inadequate. If a study exhibits the first three validities plus statistical conclusion validity but lacks external validity it is a valid case study whose results are limited to the sample used in the study. Subsequent studies—replications, coordinated surveys, or meta-analyses—could provide the needed external validity for its results. If a study exhibits the first three validities plus external validity but lacks statistical conclusion validity, as exemplified by a well-designed classic survey or a public opinion poll, then the study only lacks statistical estimates of effect sizes and tests of significance, which an interested researcher could apply to the study's tables.

Stable Association

The survey analyses that exemplify *classic causality* test the time-ordered $x \rightarrow y$ relationship for spurious association and delineate developmental sequences: $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow y$. The classic studies do not apply quantitative statistical methods but are well-designed, reflect the investigator's understanding of the relevant substantive processes and mechanisms, and present carefully analyzed data. The authors of *Communism, Conformity, and Civil Liberties* (Stouffer [1955] 1992) and *The Academic Mind* (Lazarsfeld and Thielens 1958) tested for spurious relationships by controlling for a range of background variables. Stouffer showed that the threat of communism induced a willingness of the public to restrict civil liberties in the name of security. Lazarsfeld and Thielens's study of political extremism on college campuses showed that academic freedom incidents and permissiveness (i.e., progressive attitudes) induced the teachers' apprehension; Chapter 2 reviewed this study. In *Christian Beliefs and Anti-Semitism* Glock and Stark traced the developmental sequence that linked religious dogmatism → religious hostility toward Jewish people → anti-Semitism (1966, 134–135).

Robust dependence advances classic causality by controlling for test factors simultaneously, quantifying effects, testing the quantified effects for spuriousness, and delineating intervening mechanisms. Examples discussed earlier in Chapter 3 include Suppes's notion that x is a genuine cause of y if appropriate controls do not make that relationship disappear; Granger causality in time series as illustrated by Brenner's studies about the effects of economic variables on psychological well-being; meta-analysis of findings from many studies about the $x \rightarrow y$ relationship; and Austin Bradford Hill's considerations for assessing causality in epidemiology.[2]

Potential Outcomes

Rosenbaum and Rubin ([1983] 2006, 170) clearly define the potential outcomes model of causality as it bears on the development of propensity scores for intervention studies when random assignment to the treatment and to the control is absent:

> Inferences about the effects of treatments involve speculations about the effect one treatment would have had on a unit which, in fact, received some other treatment. We consider the case of two treatments, numbered 1 and 0. In principle, the i^{th} of the N units under study has both a response r_{1i} that would have resulted if it had received treatment 1, and a response r_{0i} that would have resulted if it had received treatment 0. In this formulation, causal effects are comparisons of r_{1i} and $r_{0\mathrm{i}}$, for example, r_{1i} - r_{0i} or r_{1i} / r_{0i}. Since each unit

> receives only one treatment, either r$_{1i}$ or r_{0i} is observed, but not both, so comparisons of r$_{1i}$ and r_{0i} imply some degree of speculation. In a sense, estimating the causal effects of treatments is a missing data problem, since either r_{1i} or r_{0i} is missing. … The assumption that there is a unique value r_{ti} corresponding to unit i and treatment t has been called the stable-unit-treatment-value assumption (Rubin 1980a), and will be made here. … The N units in the study are viewed as a simple random sample from some population, and the quantity to be estimated is the average treatment effect, defined as $E(r_1)$ - $E(r_0)$, where $E(.)$ denotes expectation in the population.

Because an individual unit cannot receive both treatments simultaneously, this potential outcomes model is also referred to as the counterfactual model (Morgan and Winship 2007). As developed in Chapter 3, the Coleman causal model for attributes shares some assumptions and features of this potential outcomes approach.

To minimize the differences between the treatment and control groups in observational studies, Rosenbaum and Rubin and their many colleagues create propensity scores that can balance the empirical observations in each of the groups, thereby minimizing bias.[3] With the treatment group coded 1 and the control treatment group coded 0, and using an appropriate generalized linear model, they regress this indicator variable on a vector of observed confounding pretreatment measurements or covariates for the i^{th} unit that are prior to treatment assignment and related to both treatment and response.[4] The propensity score for a unit is its propensity toward the treatment 1 given the observed covariates in the vector x. Given the propensity score for each unit, the treatment and control groups can be balanced using matched sampling, sub-classification (i.e., stratified matching), and covariance adjustment. Then, the treatment effect is the difference between the average outcomes of the balanced treatment and control groups in the study. Chapter 12 provides an example of the construction and use of propensity scores.

Dependency Networks

As Chapter 4 exposited in detail, dependency networks are consistent with level-two causality as defined by Cox and Wermuth (2004, 288), which aims to ascertain if:

> the evidence is convincing that there is no alternative explanation, and especially when the developmental [i.e., generative] process is well understood. We use the cautious approach not to discourage the search for causality, but rather to rule out the possibility that real associations can be deemed causal merely by naming them so.

An example in Chapter 4 develops and tests a generative mechanism that links the home university of exchange students (a level-2 variable) to their consideration of use of knowledge to solve practical problems (a level-1 variable) via an intervening mechanism (i.e., prior research experience → confidence in research skills → self-rated ability to create innovative designs). This structural equation modeling combines subject-matter theory with empirical research, tests the time-ordering of the variables, develops intervening chains of relationships, uncovers the mechanisms that generate the successive response variables, quantifies the direct and indirect effects of the variables, and eliminates alternative explanations. The empirical analyses of voting in Chapter 4 also provide examples of dependency networks. They analyze the level-1 properties of voters as these variables form a path-analytic network of relationships producing a voting choice.

Having presented criteria for classifying social research studies according to their validities and zone of causal notions, let us initially take a position of neutrality regarding the results of the multilevel modeling of the chapters. Then, let us judge the validities of a study, determine the causality zones achieved by the findings from its multilevel modeling, and suggest how these studies could be strengthened, first for the contextual studies, then for evaluative studies, and finally for the meta-analytic consolidations.

Causality Zones of the Contextual Studies

Chapter 2, Contextual Analysis and Multilevel Models

Lazarsfeld and Thielens reported their pivotal contextual analysis in their Fig. 10-9 (1958, 259) of *The Academic Mind*. As explicated earlier in Chapter 2, this figure related the teacher's amount of apprehension to his measure of permissiveness and to the count of academic freedom incidents at his own academic institution. These three variables formed a social mechanism in which the incidents combined with the permissiveness to produce the apprehension. For the reasons mentioned in Chapter 2, this study exhibited strong fit validity, construct validity, internal validity, and external validity, but its a-statistical methodology weakened its statistical conclusion validity.

In a separate analysis I applied contemporary statistical methods to quantify the data of their Fig. 10-9 (Smith 2010a). I treated apprehension as a binomial response and used SAS's Proc Glimmix to estimate the effects of permissiveness and incidents on apprehension. The overall cross-level interaction was not statistically significant, but this masked very significant differences between the conservative and more permissive teachers, and different patterns of growth in apprehension: the apprehension of conservative teachers increased linearly with increases in institutional incidents whereas this linear pattern did not hold for the other teachers; their higher initial level of permissiveness when incidents were zero somewhat limited their increase in apprehension as incidents increased.

Lazarsfeld and Thielens synthesized their contextual analyses by creating an empirical theory linking context, mechanism, and outcome that is composed of the following relationships that they depicted schematically (see their Fig. 7-13, 188): The quality of the academic institution on the *x*-axis was the pivotal institutional variable. The institutions of higher quality were more likely to have permissive rather than conservative professors—the proportion permissive teachers increased rather linearly with institutional quality. Because academic freedom incidents were more frequent at the higher quality institutions, these institutions experienced more pressures on the faculty to conform politically. In spite of these increased pressures, the faculty's apprehension changed in an arc rather than linearly across the institutional quality metric because the higher quality institutions had the more protective administrations, thus limiting the growth of apprehension at these institutions.

Even with my quantification of the contextual relationships among apprehension, permissiveness, and incidents, causality remains in the *classic zone of stable association* because other appropriate level-2 and level-1 variables were not being controlled simultaneously. Because of the investigators' reliance on counter-sorters it was not feasible for them to control simultaneously for all of their key covariates, or to estimate the parameters of the complex model of the relationships that they hypothesized in their summarizing schematic. An interested reader could build on their report by obtaining the archived data from the Roper Center for Public Opinion Research and estimating multilevel models that include all of the relevant level-2 and level-1 variables. Of course, a new study of academic institutions and their faculty as they respond to the stresses of political correctness, economic uncertainty, antiwar dissent, and affirmative action would be most relevant, but very expensive to implement with the same quality as that of this earlier study.

Chapter 7, Global Human Development

This chapter groups the countries of the world into one of 18 regions; the countries and their properties are level-1 variables and the regions are the level-2 variable—the model is thus multilevel.

The validities

The level-1 properties of the countries exhibit fit validity because they are grounded in thoughtful theoretical conceptualizations. These include Huntington's (1996) cultural zones, which here serve as a proxy for the fixed effects of region; indicators of Sen's (1999) instrumental factors, which are indicated by a country's amounts of contemporary slavery, democracy, indebtedness, internal conflict, and corruption; and the United Nations' human development index. These measures exhibit acceptable construct validity. Missing data on slavery and on the very poorest countries eliminated countries from the sample, thereby threatening the internal and external validity of this study. Even so, the remaining countries have geographic spread, the sample's limitations are discussed, the sample size is reasonable, and the use of resources is appropriate (Sudman 1976, 26). The study reports the effects for the 138 countries for which the data are complete; this sample is middling in quality but credible. If ambiguities about the classification of countries coded zero on contemporary slavery had been resolved (zero = no data on slavery or zero = no slavery), then the number of countries classified by their amount of slavery would have been 179. The use of the parametric statistics of Proc Mixed to model the effects on the country's rank on the human development index threatens statistical conclusion validity, but the replication of the main analysis on scores for the components of the human development index and for their summary index alleviates this concern. Overall, I judge the validities of this study to be sufficient for the modeling strategy, which included descriptive and causal analyses of these data.

Descriptive analysis

The descriptive analysis aimed to uncover which fixed covariate exhibited the largest noncausal effect (i.e., correlational effect) on the HDI. With the level-1 countries grouped by the level-2 regions, and with regions designated as random, the HDI was regressed on the cultural zone and the five instrumental factors. The region as random enabled SAS to estimate each region's random effect measured by the amount of variability in the HDI the region exhibits above or below the mean value of zero. SAS also estimated the variance components: the level-2 variance between the regions and the level-1 variance of countries within the regions; SAS refers to these as covariance parameters. Initially, both statistics were statistically significant. To determine which of the factors had the greatest predictive impact on the HDI, each factor was deleted from the full model one at a time and the effect of the deletion on measures of goodness of fit was noted. The factor that produced the largest change in goodness of fit (i.e., increased badness of fit) was thought to have the most important predictive size of effect. When the cultural zone of a country was deleted from the model, compared with the other factors, it produced the largest increase in badness of fit and in this sense it was the most important predictor, but its causal status was not yet determined. Because the study met the validity requirements, and because all of the covariates exhibited stable effects that were quantified and replicated using the underlying scores, I find that *robust dependence* (i.e., stable association) characterized this phase of the analysis.

Causal analysis

The causal analysis of these data differs from the descriptive analysis in that a hypothetical causal factor is tested by forcing it to nest the regions. This nesting of the random regions by a test factor that is a covariate in the model regroups the countries into homogeneous region × test-factor subgroups—the countries in a region are thus matched on the test factor (Cochran 1968) and the mean-square-error (which is proportional to the level-2 variance) is calculated across these sub-tables. Most simply, the test factor is a control and its effects on the size of the level-2 variance and on the size of the random effect estimates for each of the region × test-factor subgroups are assessed. If the control for the test factor eliminates the level-2 variance, and if the subgroups that are homogeneous with respect to the region and the test-factor do not exhibit statistically significant differences above or below the mean of zero, then that test factor would be a cause of the initially observed difference between the regions. Controlling for the level-1 fixed effects, a crucial contrast compares the level-2 variance component when the countries in the regions are only grouped by the region compared with the resulting level-2 variance when the countries are grouped jointly by the region and the test factor.

This test factor is a counterfactual because the regions are not initially grouped by the test factor and the countries within the regions are not homogeneous with respect to their HDI score. The test factor provides an answer to this "what if" question: "What if the countries grouped by their region were made homogeneous with respect to their amount on the test factor, how would this change the level-2 unexplained variance in human development?" The nesting of the regions by a test factor, which is a manipulation of the assumed causal factor, implements this hypothetical condition: If the test factor does not eliminate the level-2 variance then it is not a causal factor; if it does eliminate this variance then it is a causal factor.

The analysis of the countries' rank on the HDI suggests that fully democratic political systems and the absence of the new slavery caused the regional differences in the HDI rank, because the nesting of the random regions by either of these covariates reduced (i.e., accounted for) the between-region variance so that it was insignificant. The replication of this analysis using the countries' underlying human development scores suggest that these explanatory factors work in conjunction with the absence of severe internal conflict, a rough indicator of the rule of law. Emancipatory employment, full political democracy, and human development can work together creating a mutually beneficent circle of feedback relationships. The chapter's conclusion presents qualitative data on the South and the North prior to America's civil war that corroborates these statistical interrelationships. The South was characterized by slavery and the use of force to achieve compliance, limited democracy, and lower human development. The North was characterized more by emancipatory employment and the rule of law, democracy, and higher human development.

Causality argument

The results of the causal modeling of the variance between regions, when the random regions are nested by test factors that are covariates in the model, reaches the zone of *potential outcomes* causality, as follows. Let the reduction in the variance in the human development index between the regions be symbolized by δ; let the variance between the regions when the regions are not nested by a relevant covariate be symbolized by σ_r^2; and let the variance between the regions when they are nested by a relevant test factor t that is a covariate in the model be symbolized by$\sigma_{r(t)}^2$. Then, conceptually, $\delta = \sigma_r^2 - \sigma_{r(t)}^2$. The covariates in the multilevel models that produce these variance estimates are identical; the key difference between the specifications of these models is the random effect between the regions. In the un-nested model this random effect $a_i \sim iid\, N(0, \sigma_r^2)$; whereas in the nested model the analogous random effect is nested by the test factor, $a_{i(t)} \sim iid\, N(0, \sigma^2_{r(t)})$. There are two potential outcomes: (1) If the covariate used as a test factor that nests the random regions has little effect, then δ will be near zero because σ^2_r will approximately equal $\sigma^2_{r(t)}$. Then that covariate will not be a causal factor explaining the variance between the regions. (2) However, if the nesting covariate is causally linked to the between-region variability in human development, then $\sigma^2_{r(t)}$ will be near zero, σ^2_r will be much greater than $\sigma^2_{r(t)}$, and the δ will be larger, approaching δ $= \sigma^2_r$. Thus, the reduction in the between-region variance δ, which signals a causal effect, depends upon which of these potential outcomes is realized.

In sum, the descriptive analysis of the predictors of human development reached the zone of *robust dependence*. The causal analysis highlighting the explanatory effects of the absence of the new slavery, the absence of internal conflict, and the presence of full democracy reached the zone of *potential outcomes causality*. However, the findings of this chapter would be strengthened if they were replicated in a study that has better measures, a full data set, longitudinal observations, and explicit generative mechanisms.

Chapter 8, A Globalized Conflict

To develop an explanation of the contemporary violence against Jewish people, this chapter analyzes repeated aggregated measures on ten Western European countries. Qualitative and exploratory analyses aimed to alleviate any weaknesses of the findings stemming from the use of these aggregated measures. Even with such measures, the study exhibits validities sufficient for meaningful statistical modeling at the level of the country; the evidence for these validities is assessed next:

Fit validity

My detailed reading of the theoretical and substantive literature, a prior pilot study, and a qualitative analysis of factors related to the counts of anti-Jewish violence shaped the conceptual framework for the multilevel modeling. The sensitizing theory suggests that: (1) A country's count of anti-Jewish violence depends on its social structural propensity toward violence that varies with the sizes of its Jewish and Muslim populations; the larger the populations the larger the counts. (2) During the period of the second intifada, the Arab–Israeli conflict acted as a stimulus triggering the perpetrators' violent actions. (3) The inaction of ordinary Europeans allowed such violence to take place. (4) The inhibiters of anti-perpetrator actions are two hypothetical micro-psychological mechanisms: diffusion of responsibility and cognitive ambivalence. (5) Antisemitic propaganda worsened the violence.

Construct validity

The response variable is a country's yearly count of anti-Jewish violent events as reported by the Stephen Roth Institute. Although the Institute applies their procedures for counting events objectively and consistently across the various countries, their event counts are often much lower than the counts of events reported to the police. Because the reporting practices to the police vary from country to country and thus are inconsistent, the Roth Institute's consistent but lower counts better suit this study.

The measures for ten European countries are taken from surveys sponsored by a Jewish agency and implemented by a competent survey research organization. The items in the questionnaires adequately cover the topics relevant to the phenomena being studied and provide aggregate measures of the theoretical constructs for each country. These measures include the belief that the Palestinian Authority truly wants peace, a measure of pro-Palestinian attitudes that is an indirect measure of the triggering events; valuing both Jewish and Muslim lives, a measure of cognitive ambivalence; valuing Jewish lives, an indirect measure of the diffusion of responsibility of bystanders; an index of antisemitism; and other measures thought to be relevant (e.g., Israel's stance on peace, Holocaust remembrance, and so forth).

Internal validity

The sponsor of the surveys would not provide me with access to the individual-level survey data for secondary analysis. To answer my research questions I had to use available aggregated data from the survey reports; this limits the ability of this study to document microlevel causal relationships appropriately. Although the counts of violence were taken yearly, the trend surveys skipped some years. To create a flat file of data for the replication of five years, I assumed that the data were in equilibrium: I assumed that a missing year of data was the same as the previous year of data, and when possible, checked this assumption empirically. This assumption about the data weakens the internal validity of the results.

External validity

The ten countries whose populations were surveyed are all in Western Europe and are democracies. The results therefore should not be generalized to Eastern European countries, Russia, the USA, South America, and so forth; other processes could produce the anti-Jewish violence in these areas of the world. However, such violence tends to increase when the conflict between Israel and the Palestinians escalates.[5]

Statistical conclusion validity

A multilevel modeling strategy is an appropriate choice because the multiple longitudinal observations at two and then five time points are on each of the ten countries, and the countries are grouped by their Muslim and Jewish population category—a context variable. Fixed structural characteristics include the country's population category, the year of the observation on the country, and an indicator variable for Belgium, a country with outlying violence counts given its population category—these fixed covariates create a the country's predisposition toward anti-Jewish violence. Such additional attitudinal covariates as supporting the Palestinian Authority, valuing Jewish lives, and valuing Muslim and Jewish lives are thought to combine with the structural characteristics to generate the counts of violence of a country.

Analytic models

Given that the counts of violence are consistent with a Poisson sampling distribution, I estimated a series of Poisson multilevel models by introducing into the model each explanatory factor one at a time. I observed very carefully how an additional factor affected the size of the random effect parameter for the countries grouped by their population category and the parameter gauging over-dispersion of the modeled data. First, I estimated the intercepts-only model and, as expected, the between-country variability and the deviance of the model was large. Then, to test whether or not the year of the observation affected the results, I added that variable to the model; it reduced the deviance and improved other goodness-of-fit measures. In the next model I added a cross-sectional difference; namely, a country's Jewish and Muslim population category. If this variable had no effect on the between-country variability, then the differences in population sizes would not be a factor. But the introduction of this variable generally and very noticeably improved the measures of model fit—the numbers of Jews and Muslims in a country matters. The residuals of this model indicated that Belgium had higher violence counts than its population category warranted. To reduce this variability, an indicator for Belgium versus not Belgium was added—this binary (0, 1) indicator variable completed the structural model. Prior to the introduction of stimulus variables into this model, the counts of violence increased with the yearly timing of the events in the second intifada, especially in countries with large Muslim and Jewish populations, and in Belgium, compared with the other countries in the sample.

The proportion of a country agreeing with a questionnaire item tapping attitudes about the events in the Middle East—namely, that the Palestinian Authority truly desires peace—when entered in the equation to form the mobilization model, eliminated the variability between the countries. When indicators of anti-Israel and anti-semitic sentiments were entered into the mobilization model, their effects were not statistically significant and less important than the Palestinian Authority item. Moreover, a range of alternative test factors that included social and economic indicators and variables related to the Holocaust had little effect. Pro-Palestinian sentiments working in conjunction with the structural predisposition of a country—the relatively large sizes of its Jewish and Muslim populations—seem to have mobilized the perpetrators to commit anti-Jewish violent acts.

Completing the explanation, the analysis clarified what factors may have allowed the violence to take place. The literature suggests that bystanders to violent events may become indifferent to the interpersonal violence through a mechanism of diffusion of responsibility. Also, ordinary Europeans may experience cross-pressures and ambivalence—cognitive dissonance—that inhibits their actions to curtail the violence. Of the latter two factors, cognitive dissonance, indicated by valuing both Jewish and Muslim lives, had a stronger force than bystander unresponsiveness, indicated indirectly by valuing Jewish lives. I described the effects of the explanatory factors on the reduction of the residual random effects of the countries by plotting these patterns on a graph that depicts the efficacy of the explanatory factors relative to the baseline model.

Parameter studies

Testing these interpretations, a parameter study held constant the structural characteristics of Belgium and gave to Belgium the distribution of attitudes of Germany, which were less favorable to the Palestinians. As expected, Belgium's hypothetical counts of violence were reduced. A second parameter study of attitudes of Belgians suggested that, at low levels of the stimulus variables, bystander indifference is important, but as the attitudes change toward higher simulated values, mobilization, and cognitive ambivalence become more important determinants.

Replications

To counter the possibility that the two yearly observations are too few, the main analysis was replicated on five yearly time periods (some data were interpolated). It found that mobilization indicated by agreeing that the Palestinian Authority desires peace, a proxy for the Mid-east conflict, and cognitive ambivalence indicated by valuing both Muslim and Jewish lives, are basic causes. Less important are bystander indifference, anti-semitic attitudes, anti-Israel sentiments, and a range of other social and economic indicators.

Causality argument

Apparently, on the basis of the above evidence the study transcended the limitations of aggregated data and did achieve at least the zone of *robust dependence* for the relationships among the aggregate measures. Of course, the causal inferences would be strengthened by modeling complete yearly data at the level of individual Europeans, and by more direct measures of longitudinal change in attitudes and violence counts over shorter time periods.

Chapter 9, Will Claims Workers Dislike a Fraud Detector?

This study began as an applied study of the effects of a computerized fraud detector on the morale of claims workers involved with mitigating automobile insurance fraud. Through the reconceptualization of its variables and multilevel modeling it became a more basic research study on the determinants of anti-computerization sentiments. Its validities and results are summarized next.

Fit validity

The purpose of the study, first-hand observations of employees in claims offices, and subject-matter knowledge informed the construction of a list of topics and a survey questionnaire that covered these topics. The pretest of the questionnaire in several claims offices guided the revisions. The final questionnaire had fit validity, appropriately covering these topics: attitudes about computerization of work, knowledge about networks of perpetrators of fraud, receptivity to innovations, indicators of the employee's position in the office hierarchy, and social background variables.

Construct validity

The survey questionnaire provided valid measures of these three key variables: an "anti-index" that gauged negative attitudes about the computerization of their work (the response variable), and two orthogonal fixed covariates derived from a factor analysis: rank in the office and receptivity to innovations. These variables correlated with validating items from the questionnaire and each had high reliability.

Internal validity

The original study design included six claims office and their employees. The design had two dimensions: quasi-experimental and null treatments in closely matched offices versus office size characterized as small, medium, or large. The manipulated treatment was the installation and use of CFRD—a computerized fraud detector. Offices with the null treatment did not have this new computer system. However, this 2 by 3 design was broken by a second computer system (referred to here as Millennium 2000, M2K) that was being installed in several of the offices. Thus, some offices inadvertently received two treatments and some null-treatment offices where in fact receiving the M2K treatment. The presence of M2K in some offices violated the stable-unit-treatment-value assumption (SUTVA) destroying the internal validity of the original design. The study could be salvaged because there were now four types of offices with homogeneous treatments: two offices had both new systems installed, one office only had CFRD, another only had M2K, and two offices had no new systems. Both newly installed computer systems were manipulated stimulus variables. The internal validity of the study now was acceptable because this grouping of offices did not violate SUTVA, the geographic spread of the offices prevented interference due to the different treatments, and selection bias was minimal because there were no significant differences on a range of employee background variables from one office to another.

External validity

This study conceptualizes the six offices as a random sample from the universe of the insurance company's claims offices. Moreover, these six offices were sampled in a given year from the universe of possible years from which they could have been sampled. These assumptions enabled the six offices to be conceptualized as random. Because of the clustering of the observations within the offices and the "random" selection of the six offices, multilevel modeling was a reasonable methodological choice.

Statistical conclusion validity

Using Proc Mixed I estimated the effects on the anti-index, a rather symmetric variable that was assumed to be normally distributed. I first estimated the initial baseline model and the level-1 model comprising the fixed effects of an employee's rank in the office and receptivity to innovations. I then estimated the level-2 causal model of the effects of the various installations of the computer systems, controlling for the level-1 covariates. The baseline model identified two statistically significant variance components: the variance between the offices and the variance among employees within the offices. When I introduced the fixed employee characteristics into the model, both of these variances remained statistically significant. Then, to ascertain whether CFRD only, M2K only, one system (CFRD or M2K), or the joint installation of CFRD and M2K (both systems) caused the employee discontent on the anti-index (relative to the levels in offices with no new computer systems), I coded the types of treatments as indicator variables with the null treatment as the base. Workers in offices with both new computer systems exhibited significantly more discontent than workers exposed to CFRD only, M2K only, their combination as only one new computer system, or the null treatment. There were no significant differences between the effects of CFRD and M2K on worker discontent; both induced slightly more discontent than the null.

Causality argument

The random offices were nested by their type of installation, an explicit global property of the office—a global property of a level-2 unit is not based on aggregations of level-1 characteristics. This nesting of the office by its type of installation is consistent with the view that the computer systems were manipulations, changing the state of the office from the null condition, to one of the treatment conditions. This view was formalized by these causal conceptions of Pearl ([2000] 2009, second edition):

Let z_1 be receptivity to innovation, let z_2 be the rank of the job, and let Office Type distinguish the various implementations. Pearl's $do(x)$ operator indicates that the fixed employee-level variables are not causal: The level-1 causal equation is $Y_{ij} = \gamma_{00} + \gamma_{10}\ do\ (z_{1ij} = z_{1ij}) + \gamma_{20}\ do(z_{2ij} = z_{2ij}) + u_{0j} + r_i$. That is, hold the covariates constant at their initial values. The introduction of the level-2 Office Type indicates a causal impact of Both Systems but not for the singular introductions of new computer systems. Most simply, $y = \mathrm{a} + \mathrm{b}\ do(OfficeType) + \mathrm{c}\ do\ (z_1 = z_1) + \mathrm{d}\ do(z_2 = z_2) + e$. The OfficeType intervention may be M2K Only, CFRD Only, One System, or Both Systems. By holding the two covariates constant the equality $z_i = \mathrm{z}_i$ denotes their static, associational nature (z_1 is receptivity to innovation and z_2 is rank of job). The causal effect of Office Type = Both Systems is $b = \mathrm{E}(\mathrm{y} \mid do(Both\ Systems = 1, z_1, z_2)) - \mathrm{E}(\mathrm{y} \mid do(Both\ Systems = 0, z_1, z_2))$; the estimate of b is statistically significant. However, when $do(CFRD)$ or $do(M2K)$ or $do\ (One\ system)$ are the manipulations, their assumed causal effects are not statistically significant.

The causal equation above is consistent with the potential outcomes perspective as follows. Prior to the joint implementation of both CFRD and M2K, the claims office has potential outcomes under the Both Systems (= 1) treatment, and potential outcomes under the null treatment (= 0), not Both Systems. After assignment to one of these treatments, the office has realized outcomes under that treatment and counterfactual outcomes under the treatment it did not receive. The causal effect is the difference in the average anti-index score between these two types of treatments. The elements of the causal equation express this difference: holding constant the values of the covariates z_1 and $z_{2,}$ the treatment effect b on the anti-index score is the difference between the expectation for the office when it receives the treatment Both Systems = 1 and the expectation when it receives the null treatment, Both Systems = 0.

Chapter 9 interpreted these findings theoretically as stemming from the threat these interventions had for the substantive complexity of the work of the higher status employees (i.e., deskilling of work), and the hassle the employees experienced adjusting to the use of two new computer systems. Thus, on theoretical and methodological bases these analyses of the effects of these interventions on the employees' discontent concerning computerization achieved the zone of *potential outcomes causality*. The longitudinal evaluations of comprehensive school reform aim to achieve this zone of causality as well.

Causality Zones of the Evaluative Research

Longitudinal studies ask how an outcome changes over time and what influences the pattern of change (Maas and Snijders 2003; Singer and Willett 2003, 7–15). Ideally, a well-designed longitudinal study has three or more repeated measures on the same unit, outcome measures whose values change over time, a sensible metric for assessing time, and an adequate set of covariates to control for the effects of extraneous variables. Longitudinal studies when implemented with an experimental or quasi-experimental difference-in-differences (DID) design and an appropriate multilevel statistical model can provide evidence for their achievement of the zone of *potential outcomes*.

As explicated in Chapter 10, the two chapters on the effects of comprehensive school reforms combine a difference-in-differences (DID) quasi-experimental design with longitudinal data. These studies evaluate the effects of external change agents—consultants who were employed by Co-nect, an educational design and professional development firm then based in Cambridge, Massachusetts. These consultants endeavored to implement whole-school reforms and professional development by providing customized portfolios of educational reforms. For the periods of these studies, depending on the needs of the schools, Co-nect consultants provided in-service workshops and mini-sabbatical sessions that introduced project-based learning, standards-based assessment, the teaching of small groups of students by teams of teachers, the effective use of technology, and de-centralization of leadership.

To control for spuriousness and to reduce bias, these evaluative studies rely on, respectively, matching and propensity scores. Even without direct measures of the numerous aspects of the reforms, because these studies blend DID designs, matching or propensity scores, and multilevel modeling; both studies aim for the zone of *potential outcomes causality*.

Chapter 11, Target, Matched, and Not-Matched Schools

Chapter 11 presents my evaluation of elementary school reforms in Harford County, Maryland. It advances an earlier study by Russell and Robinson (2000) in which they closely matched each of the five target elementary schools that received the reforms with two other elementary schools that did not receive the reforms. They matched these schools on six demographic variables: grades served by the school, the number of students in the school (i.e., school size), the percentage of nonwhite students (i.e., ethnicity), the percentage of students eligible for a free or reduced lunch (i.e., parental socioeconomic status), proficiency in English (i.e., students with Limited English Proficiency, LEP students), and the mobility rate. This earlier evaluation compared the performance of each of the five target schools to its two matched schools; the results were weakly positive.

Contrastively, in this study I borrowed strength by pooling the data and analyzing all 31 elementary schools grouped by their type as target, matched, or not-matched. The use of a not-matched control group follows William Cochran's advice, as related by Rubin ([1984] 2006, 10):

> A second theme in design is the need for a control group, perhaps several control groups (e.g., for a within-hospital treatment group, both a within-in hospital control group and a general-population control group). The rationale for having several control groups is straightforward: If similar estimates of effects are found relative to all control groups, then the effect of the treatment may be thought large enough to dominate the various biases probably existing in the control groups, and thus the effect may be reasonably well estimated from the data

The not-matched group enhances the validities of the study.

Fit validity

My use of administrative data shaped the conceptual scheme of the study, defining the pivotal variables. These data included indicators of the students' social and economic backgrounds, student-to-teacher ratios, and test scores. The study lacked direct measures of classroom processes, teacher quality, the student's family life, and so forth. Consequently, the modeling could only focus on how the intervention influenced aggregated performance measures, controlling for the available covariates on a school at a time point.

Construct validity

The measures of the treatment, the covariates, and the response variables exhibited sufficient construct validity for the multilevel modeling:

Treatments

The target schools, which numerous African-American students attended, experienced the educational reforms for a period of three years; the matched and the not-matched schools did not experience the reforms during that time period. Compared with the not-matched schools, the target and matched schools had similar smaller standard deviations of response variables suggesting that the matching was effective. Because the quality of the implementation of the reforms varied, all nontarget schools received an implementation score of 1 whereas each target school received an additional increment depending on the efficacy of the reform implementations; their scores ranged from 1.52 to 1.86. The multilevel modeling used these scores as a weight variable in Proc Mixed.

Covariates

The target and matched schools had very similar values on four key covariates: the proportion of students receiving free and reduced price lunches (i.e., parental socioeconomic status), the proportion African-American (i.e., student ethnicity), the proportion female (i.e., student gender), and the ratio of students to teachers (i.e., overcrowded classrooms). The not-matched schools had higher proportions of disadvantaged students and African-American students. Additional covariates included the location of the school (urban fringe or rural), the lowest grade of the school (pre-kindergarten or kindergarten), whether the school qualifies for Title-1 benefits, and whether the school is a targeted poverty school; the values of these additional covariates tended to be stable across time. I centered all of these covariates, and the time periods that intervened between pre and post, by their overall sample means; their effects thus appeared in the intercept term. This procedure sharpened the focus of this study to uncovering the *effects of a cause*; the effectiveness of the educational consultants.

Response Variables

The aggregated measures characterized the achievement of the students in a school's third grade and fifth grade. The composite index (CI) summarized a school's performance in six content areas: reading, writing, language usage, mathematics, science, and social studies. The school performance index (SPI) summarized a school's relative distance from satisfactory standards. Because these measures were almost perfectly correlated, I gauged overall school performance by the combination of these two variables. The response variables for third and fifth grade included a school's grade-specific CI, reading, and mathematics test scores.

Internal validity

By pooling the data on all of these schools, this study borrowed strength: it shrunk the diversity of the schools by estimating one equation rather than the ten implicit equations in the earlier un-pooled analysis. To reduce bias, the multilevel models applied both matching and regression methods (Cochran and Rubin [1973] 2006, 57). Moreover, the following aspects of this study contributed to its internal validity: the values of the covariates were stable across time; the reform treatments were not shared by the other schools; the weighting by the implementation scores buttressed the stable-unit-treatment-value assumption (i.e., SUTVA); the comparisons between the target and matched schools, and between the not-matched and matched schools, specified the effects of the reforms.

External validity

In essence, Chapter 11 presents a case study of the effects of comprehensive school reform as provided by Co-nect consultants in Harford County, Maryland for school years 1996–1997 through 1999–2000. Its results should not be generalized much beyond this locale and time period. However, these results were included in a wide-ranging meta-analysis of studies of comprehensive school reform that documented the positive effects of such efforts, and Chapter 12 tests the Co-nect reforms in schools in Houston, Texas.

Statistical conclusion validity

In the multilevel models the yearly measures on an elementary school are at level-1 and the schools are at level-2. The type of school—target, matched or not-matched—is the nesting typology; the schools nested in their type at a time point are the subjects. My careful selection of the appropriate covariance structures for the multilevel modeling and my application of adjustments for the multiplicity of the response variables enhanced the validity of the statistical conclusions.

Covariance structures

Because the covariance structure of a model can profoundly influence the estimates of the statistical parameters and their significance, I conducted a series of tests that pointed to the relatively best covariance model for each response variable. The compound symmetry (CS) covariance structure was the best for six of the eight response variables. For third grade reading the variance components (VC) model fit best, and for fifth grade reading the heterogeneous compound symmetry (CSH) model fit best.

Treatment effect coefficients

The models estimated the effect of the reform treatment as the coefficient on the interaction of the indicator for the post time period and the indicator for target schools. Similarly, the models estimated the null treatment effect of the not-matched schools as the coefficient on the interaction of the indicator for the post time period and the indicator for the not-matched schools. Both treatment effect coefficients were estimated relative to the matched schools, the omitted category. A substantial effect of the reform treatment compared to the matched schools, along with an insubstantial effect of the null treatment of the non-matched schools compared to the matched schools, would indicate a strong causal effect of the reforms.

For a given response variable, the treatment effect coefficient also equals the difference between two differences, as follows: (1) for the target schools calculate the difference between the post-period and pre-period means; (2) for the matched schools calculate the difference between the post-period and pre-period means; then (3) calculate the difference between those two differences. This same logic holds for the null effect of the not-matched schools relative to the matched schools: the DID effect coefficient is the difference between the pre-to-post difference in the means for the not-matched schools and the pre-to-post difference in the means for the matched schools.

Multiplicity corrections

Because the response variables are numerous and confounded (e.g., the comprehensive index includes the reading and mathematics tests), a statistically significant finding could happen by chance. To protect against this threat to statistical conclusion validity, I grouped the change in the outcomes into these families: overall, third grade, and fifth grade change. Then I applied the step-down Bonferroni procedure of SAS's Proc Multtest to calculate the final probability values.

Results

Three research questions asked: Did the Co-nect consultants improve the overall performance of the schools; the performance of students in third grade; and the performance of students in fifth grade? The pattern of the findings is clear: for the two overall performance measures the target schools exhibit a clear advantage in pre-to-post improvement compared with either the matched or not-matched schools. For third grade the performance of the matched schools declined leading to a large increase in the target schools and smaller increases in the not-matched schools. For fifth grade the target schools exhibited improvement; the performance of the not-matched schools declined more than the decline in the matched schools. However, the corrections for multiplicity weakened some of these results and also pointed to some significant sizes of effects.

Because evaluations of educational reforms are best assessed in terms of the effect size statistic *d*, the chapter reports estimates of *d* that I calculated using the correlational method. This method depends on the value of t^2 and the degrees of freedom (*df*) of the treatment effect. Initially, the $df = 27$. Because this relatively small number appears in the formula for the correlation, it results in a large estimate of r and an unrealistically large estimate of the effect size d. However, when the Repeated statement is used, as it was in these models, SAS recommends applying the Kenward–Roger corrections for the standard errors, F-statistics, and degrees of freedom. When either the corrected or uncorrected values of the estimates of d were used, for all eight outcomes the effects of the reform treatment relative to the matched schools were noticeably larger than the effects of the null treatment of the non-matched schools relative to the matched schools. The values of effect size produced by the Kenward–Roger corrections were smaller and more appropriate. Appendix Tables 11.1 and 11.2 compare a range of effect sizes and covariance parameters when Kenward–Roger corrections are in force, and when such corrections are absent. The effect sizes change but the covariance parameters and their statistical significance are unchanged.

Causality argument

For the time period of their active involvement in the schools, the Co-nect consultants and their school reform treatments caused pre-period to post-period overall improvement in student achievement ($d = 0.46$ compared with $d = -0.06$ for the null effect of the not-matched treatment); third grade reading ($d = 0.64$ compared with $d = 0.46$ for the null effect); and fifth grade mathematics ($d = 0.58$ compared with $d = -0.23$ for the null effect). Given that these target schools are closely matched with their comparison schools, and that the favorable results were tested for multiplicity, this longitudinal analysis mostly likely achieved the zone of *potential outcomes* causality for these three findings. Of course, additional controls for propensity scores would have strengthened these claims of causality.

Chapter 12, Using Propensity Scores

This longitudinal study compares the change in test scores for seven elementary schools that received a comprehensive school reform treatment with the change in six nearby elementary schools that did not receive this treatment; these schools in Houston, Texas have many Hispanic students. The tests covered reading, mathematics, the average of reading and mathematics, and writing in fourth grade. During the final school year additional teachers in the low-performing schools improved the ratios of students to teachers, creating a second treatment effect that improved the fifth grade test scores of the comparison schools.

Fit validity

The conceptual scheme was shaped by the available aggregated administrative data on the students for the school in a specific school year (SY) and by the characteristics of the schools. The variables on the students included indicators of their ethnicity (Hispanic, African-American, and so forth), economic status (eligibility for free or reduced-price lunches), limited English proficiency (LEP), grade in school, gender, mobility, and test scores. The variables on schools included their treatment status (exposure to the reforms or not), student-to-teacher ratios, Title-1 school, highest grade of school, school rating, and the presence of Success for All, a comprehensive curriculum reform. The lack of direct measures of classroom processes, teacher quality, the family life of the students, and so forth, limits this study to assessing the effects of the interventions, controlling for the design variables and the available covariates.

Construct validity

Robust treatment effects often reduce the standard deviations of response variables. Making this assumption, I compared the standard deviations of the response variables across the three years of the study. Validating the effect of the reforms, the standard deviations in the schools receiving the reform treatment tended to be smaller and more homogeneous than those in the comparison schools. However, in the comparison schools during the final time period, the standard deviations dropped for fifth grade reading and mathematics but not for the writing test given in fourth grade, thereby validating the effects of the extra teachers in fifth grade.

The covariates in the model are not collinear and appropriately measure what they purport to measure: The proportion of Hispanic students and the proportion of students eligible for free or reduced price lunches tap the ethnic and economic background of the students in the school. The highest grade of the school, the mobility rate, and the student-to-teacher ratio tap salient school characteristics. The propensity scores did indeed influence the relationship between the reform treatment and the outcomes. Conversations with school personnel verified the presence of Success for All in a school.

Internal validity

The values of the covariates were stable across the three years of the evaluation, suggesting that changes in a school's population did not determine the outcomes. The target and comparison schools had very similar characteristics, but there were some differences: The target schools had a higher proportion of Hispanic students and a lower proportion of African-American students. The students in the target schools were slightly less disadvantaged economically. The teachers in the target schools were offered the reforms and they voted to have their schools receive the reform program; the comparison schools were not offered and did not receive the reforms.

Because the target and comparison schools were not matched prior to the reforms, selection bias was an important factor that could have affected the results.

To control for selection bias I estimated propensity scores, which are the predicted probability that a school received the reform treatment. I applied a Firth-corrected logistic regression model to linearized data, regressing the treatment dichotomy—the school received the reforms (1) versus it did not receive the reforms (0)—on a number of antecedent characteristics of 200 Houston elementary schools. The predicted scores for the target and comparison schools were then used as an additional mean-centered covariate in the regression of the response variables on the explanatory variables.

The additional teachers in fifth grade did not threaten the internal validity of the study; rather, they reduced the student-to-teacher ratios creating a second treatment effect. Consequently, this evaluation focuses on change from Time 0 to Time 1 due to the reforms; change from Time 1 to Time 2 due to the extra teachers, and overall change from Time 0 to Time 2 due to the reforms. In the comparison schools in the final year of the study (i.e., fifth grade), the extra teachers boosted the performance of the students on the mathematics and reading tests from the nadir of the previous year, but not on the writing tests administered in fourth grade.

External validity

The results of this case study of elementary schools in a specific feeder system in Houston, from SY 1999–2000 (Time 0) through SY 2001–2002 (Time 2) should not be generalized much beyond this locale and time period. However, these findings do contribute to cumulative social research on the effects of comprehensive school reforms and smaller class sizes, and can be included in meta-analyses on these topics.

Statistical conclusion validity

In these longitudinal analyses the schools are the level-2 units; they are nested by the reform treatment typology, which also is a fixed variable in the model. At the three equally spaced time points, the repeated measures on the covariates and the response variables are the level-1 units. Simplifying the modeling, all of the covariates except for the design variables are centered by their overall means; the modified intercept term expresses their effects. Consequently, the design can be viewed as having observations on two groups of schools for three time periods.

Difference-in-differences

The logic of difference-in-differences (DIDs) defined the treatment effects that quantify the change on a response variable from Time 0 to Time 1, Time 1 to Time 2, and Time 0 to Time 2. The treatment effect coefficients were estimated as follows: (1) calculate the the pre-period to post-period difference between the means in the reform treatment group; (2) calculate the pre-period to post-period difference in the means in the comparison group; (3) calculate the difference between the two differences.

Multilevel logistic regression

The tests scores were reported as the proportions passing a test in a school at a specific grade or time point. These proportions are conceptualized as the number of successes per 100 trials. For these test scores, Proc Glimmix provided the estimates of the parameters of the statistical model assuming that the error was binomial and the link was the logit. The logit is the natural log to the base e of the probability of success (p) divided by the probability of not success ($1 - p$). If 90% of the students pass a test, then the logit $= \ln(0.90/0.10) = 2.197$. All of the test-score proportions were transformed into logits; the statistical model is multilevel logistic regression. The statistical significances of the effects were reported on the logit scale and for odds ratios.

Covariance structures

The covariance structure of a model of a response variable influences the significance of the effects and must be chosen with care. Using the new Covtest statements of SAS, I selected the following structures as preferred: the compound symmetry structure was best for the reading and the average of reading and mathematics tests; the two-banded Toeplitz, for the mathematics tests, and the banded main diagonal, for the fourth grade writing tests.

Corrections for multiplicity

Because there were several response variables, and the average of the reading and mathematics tests was confounded with its two component measures, the treatment effects were corrected for multiplicity. The step-down Bonferroni provided stringent estimates and the false discovery rate provided less stringent estimates.

Effect sizes

Because educational researchers interpret their findings in terms of standardized effects, I calculated the effect sizes as the quotient of the DID effect expressed as a proportion difference divided by the standard deviation pooled across groups and time period; I cross-checked these estimates by applying the correlational method.

Results

The target schools improved their performance between the baseline year (Time 0) and the first full year of the reforms (Time 1); the performance of the comparison schools usually declined. The difference between those differences created statistically significant effects of the reform treatment for reading, mathematics, and the average of reading and mathematics, but not for fourth grade writing. The corrections for multiplicity weakened the significance of the reforms on the reading tests, whereas the significance of the effects on mathematics and on the average of the reading and mathematics tests were maintained. The standardized effect sizes were: reading = 1.3SD; mathematics = 0.79SD; average of reading and mathematics = 1.1SD; and fourth grade writing = 2.05SD.

The comparison schools improved their performance from Time 1 to Time 2, presumably because of the well-trained extra teachers assigned to the schools to prepare the students for the achievement tests in fifth grade. In the comparison schools their presence reduced the student-to-teacher ratios from 17.1 to 14.7, while in the target schools the ratios remained higher. The difference-in-differences estimates resulted in statistically significant increases in mathematics and the average of mathematics and reading, but not for reading and fourth grade writing. However, the step-down Bonferroni weakened the significance of the mathematics and the average of the reading and mathematics tests to $p = 0.069$, whereas the less stringent false discovery rate maintained their significance at $p = 0.039$. The standardized effect sizes were: mathematics = 0.96SD and average of reading and mathematics = 1.19SD.

Because the differences from Time 0 to Time 2 for the treatment and comparison schools were about equal, the overall difference between the differences, which defined the effects of the comprehensive reforms for reading, mathematics, and the average of reading and mathematics, were minimal, even though the target schools maintained their improved levels of performance. However, the fourth grade writing tests that were not affected by the extra teachers allowed the comprehensive reforms to have a positive, statistically significant improvement relative to the change in the comparison schools. The effect size of the comprehensive reforms was 1.29SD.

Causality argument

The statistically significant findings reached the zone of *potential outcomes causality*. For the target schools from Time 0, the baseline period, to Time 1, the reforms caused improvements on all four of the outcomes treated as a composite, and especially for mathematics and the average of the reading and mathematics tests. For the comparison schools from Time 1 to Time 2, the effects of the extra teachers caused improvements in mathematics and the average of the reading and mathematics tests. In the target schools from Time 0 to Time 2 the comprehensive reforms caused improvements in the fourth-grade writing tests.

Causality Zone of the Research Summaries

Chapter 14, Gatekeepers and Sentinels

To consolidate quasi-experimental findings about the effectiveness of precertification nurses functioning as gatekeepers and onsite concurrent review nurses functioning as sentinels, this chapter applied the fixed- and random-effects paradigm of meta-analysis. Integrating the findings from eight studies, it quantified the effect of precertification nurses on rates of inpatient admissions. Consolidating the findings from four studies, it quantified the effects of the onsite nurses on a range of indicators of utilization and costs of inpatient medical care. These response variables included measures of cost savings in total and ancillary expense; reductions in utilization assessed by reduced admissions rates, lengths of stay, and bed days; and medical complications. The conventional program restricted the onsite nurses' discretion by having them check with utilization review physicians; the "cybernetic" program enabled the nurses to exercise their discretion. These consolidations included all of the studies the research group conducted circa 1986–1991 on precertification and onsite review, regardless of any shortcomings in their validities, which I examine next.

Face validity

All of these studies were secondary analyses of administrative data from databases of medical insurance claims. These databases included rich information on insurance plan design, geographic area, patient demographics, ICD-9 disease codes, case mix, seasonality, and so forth. These data were merged with information specifically about the presence of precertification, and information about the presence of precertification and onsite concurrent review.

Construct validity

The researchers checked the presence and timing of the treatment variables very carefully and care was taken to roughly match the comparison group with the treatment group. These two groups were not contiguous so there was little evidence that the treatments influenced the comparison group. The sets of covariates in the models were chosen with care, they included general variables common to all studies and other variables that were specific to a study. The consolidation of data on precertification focused primarily on its effects on admissions rates and explored its effects on indicators of medical complications; the latter measures were state-of-the-art at the time of these studies. The consolidation of data on onsite concurrent review included effects on numerous interrelated response variables; multiplicity was a potential problem. However, the key driver of savings was the onsite nurses' reduction in expense for ancillary services. When this component of expense was removed from overall expense, then the cost savings were minimal.

Internal validity

Almost all of the studies of precertification (and all of the studies of onsite review) whose effects were consolidated applied the basic fourfold differences-in-differences design as exposited in Chapter 10 of Part 3 of this book. Conceptually, the estimate of the treatment effect was based on the difference between the pre-period to post-period difference in means in the target group minus the difference between the pre-period to post-period difference in means in the comparison group. Pragmatically, all of the covariates except the design variables were centered by overall means, thus putting their effects into the intercept. This method combines simplicity with complexity: it produces estimates of the few crucial variables that define the study design along with controls for nuisance covariates that are not the focus of the analysis and could spuriously affect the relationship between treatment and response. These applications thus focused on the *effects of a cause* estimating the coefficients of parameters: the intercept, the pre-to-post trend coefficient, the cross-sectional difference between target and comparison groups, and the program effect coefficient quantified as the product of the target group coefficient and that for the post time period.

External validity

These studies were conducted during the time period of managed indemnity fee-for-service health insurance, which was prior to the current ubiquitous use of networks of physicians and hospitals as utilization and cost-containment methods; the generalizability of the findings are thus limited.

Statistical conclusion validity

The researchers applied standard regression analysis methods to estimate the effects of the treatments on the response variables; this was the accepted approach at the time of these studies. Some analyses included disease codes as a covariate; these could be viewed as intervening variables between the treatment and response, leading to endogeneity bias that could reduce the treatment effects creating Type-I errors. Consequently, less biased estimates of the treatment effects may be larger than the reported sizes of effects. The multiple response variables could be creating significant findings that in fact are due to the multiplicity of responses, thus creating Type-II errors. Hopefully, the meta-analytic procedures will compensate for these potential biases.

The meta-analytic procedures

The consolidations applied the fixed- and random-effects paradigm of meta-analysis to summarize the data from the evaluations:

Data

For each study in the consolidation, a four-fold design table was created that summed the relevant estimates of the coefficients for each cell of the table. Because the intercept mean was common to all four cells of the design and was based on the total number of units, the resulting table expressed these counterfactual "what if" assumptions. If all of the units were in cell 1 (defined by intercept = 1, target group = 0 and post = 0), then the mean value for that cell would be the value of the intercept. If all of the units were in cell 2 (defined by intercept = 1, target group = 1 and post = 0), then the mean value for that cell would be the intercept value plus the value of coefficient on the cross-sectional indicator variable. If all of the units were in cell 3 (defined by intercept = 1, target group = 0 and post = 1) then the mean value for that cell would be the intercept value plus the value of the coefficient on the pre-to-post change indicator variable. Similarly, if all of the units were in cell four (defined by intercept = 1, target group = 1, post = 1, target × post = 1), then the mean value for that cell would be the sum of all four coefficients.

Assuming that the treatment and comparison groups were very closely matched and a range of relevant covariates were controlled and centered by their means, then the estimate of the average causal effect of the treatment for each study would be the difference between these two differences: the pre-to-post difference in means in the target group minus the pre-to-post difference in means in the comparison group. If these conditions were met by each of the individual studies, then the treatment effect coefficient summarizing their effects would achieve at least the zone of *robust dependence* and perhaps the zone of *potential outcomes causality*, depending on the results of the consolidation's statistical modeling.

Fixed and random effects

In a consolidation the fixed-effects model answers this question: "In the studies at hand, *did* the program on average produce an effect?" It assumes that each study provides an estimate of the one unique program effect. For this model to be valid the results should exhibit little variability from study to study—the studies produce homogeneous results that can be combined into summary measures of the true program effect and its standard error.

The random-effects model answers this question: "Based on the studies at hand, *will* the program on average produce an effect?" Because it aims to estimate the grand-mean treatment effect of all studies, the range of inference is not limited to only those studies included in the consolidation. The random-effects model quantifies the variation that is between the studies. It assumes that there is a distribution of true effects—each study has its own true effect that would result if it were replicated an infinite number of times. The results of the set of studies are first combined under the fixed-effects model and the homogeneity of the components of the overall treatment effect is tested using a chi-square test. If the components are homogeneous, then the fixed-effects model fits the data. But if the components are not homogeneous, then the random-effects model needs to be applied. This model takes into account the lack of homogeneity of the components of the fixed-effects estimate of the program effect by adding a parameter to the variance estimate. When the between-study variability is statistically significant and the random-effects estimate of the program effect is estimated, it will have a wider confidence interval than the fixed-effects estimate.

Causality argument for precertification

For the historical time period of the eight studies of precertification, the consolidated finding that the precertification nurse gatekeepers reduced annual admission rates per 1,000 enrollees (EEs) did achieve at least the zone of *robust dependence*; the same result but based on stronger underlying studies would have qualified this finding for the zone of *potential outcomes causality*. This consolidated finding was consistent with the rather limited literature on this topic. It was based on studies that applied the fourfold design comparing change between the groups with the intervention to groups that did not have the intervention, controlling for the effects of mean-centered covariates. Seven of the eight studies reported reductions in admissions. The fixed-effects estimate was statistically significant (−7.6 admissions per 1,000 EEs) with a narrow confidence interval (−4.9, −10.3). But one study had anomalous results necessitating the calculation of random-effects estimates: the treatment effect (−7.3 admissions per 1,000 EEs) was statistically significant—the wider confidence interval estimates did not include zero (−0.8, −13.8).

These reductions in admissions had very little effect on medical complications: A fixed-effects analysis of very limited data found that precertification was associated with increased rates of surgical complications but not with increased rates of obstetrical complications and not with increased rates of surgical and obstetrical complications; when the *p*-values were corrected for multiplicity none of these effects were statistically significant.

Causality argument for onsite review

Of the four onsite concurrent review sites, one program was conventional—physicians specializing in utilization review needed to endorse potentially controversial decisions of the nurse vis-à-vis the attending physician, and three programs were "cybernetic"—the onsite nurse was allowed to negotiate directly with the attending physician. The random-effects estimates across the four sites indicate that for the time period of these studies the nurse sentinels reduced bed days by −47.6 days per 1,000 lives (−47.6; −92.1) and ancillary expense by −$605 per confinement (−$929, −$281); these effects results achieved at least the *robust dependence zone of causality;* the same results but based on stronger underlying studies would have qualified the findings for the *potential outcomes zone of causality*. The activities of these nurses were not associated with increased rates of surgical and obstetrical complications, separately or combined.

Implications

This book has developed multilevel models of social problems bearing on human development. The chapters present contextual studies, evaluative research, and research summaries that clarify substantive problems through the application of fundamental research methods. To encourage social problems researchers to apply multilevel modeling and to think in terms of causal relationships, the book developed notions of causality grouped into three general categories that closely parallel Cox and Wermuth's distinctions (2001, 65–70): *stable association* includes classic causality and robust dependence; *potential outcomes* here includes causality as an effect of an intervention and stochastic causal models for attributes; and *dependency networks* include causality as graphical models, association graphs for log-linear models, generative processes, and causality in policy research.

Based on the explications in this chapter, Table 16.1 summarizes qualitatively the notions of causality achieved by the multilevel models of this book. The causality zones of the findings produced by these models range from *classic causality* through *dependency networks*; the latter is exemplified by the graphical model and generative mechanism producing consideration of use. The analysis of the effects of democracy and slavery on human development, and the intervention studies of the fraud detector and comprehensive school reform achieved the zone of *potential outcomes causality*. The predictive analysis of human development, the modeling of anti-Jewish violence in Europe, and the meta-analytic consolidation of the effects of nurses exhibited *robust dependence*. Whether these causal relationships will hold in the future is a problem for subsequent cumulative social problems research.

Table 16.1 Zones of causality achieved by the multilevel models of this book

		Stable association		Potential outcomes		Dependency networks	
Chapters	Estimating multilevel Models	Classic causality	Robust dependence	Counter-factuals	Effects of interventions	Graphical models	Generative mechanisms
2	Incidents and Permissiveness Influenced Teachers' Apprehension	√					
4	Research Experience Influenced Consideration of Use[a]					√	√
7	Civilizations, Democracy, Slavery, HIPCs Corruption, and disorder Influenced Human Development Rank and Scores		√				
	Democracy and Slavery Caused Human Development			√			
8	Middle East Conflict, Muslim and Jewish Population Size, and Attitudes Influenced Anti-Jewish Violence		√				
9	The Joint Implementation of Two New Computer Systems Caused Discontent				√		
11	Elementary school reforms caused:						
	Overall Improvement in Achievement				√		
	Improvement in third Grade Reading				√		
	Improvement in fifth Grade Reading				√		
12	Elementary School Reforms Caused:						
	Improvement in Math and Reading				√		
	Improvement in Mathematics				√		
	Improvement in Writing				√		
	Fewer Students to Teachers Caused						
	Improvement in Math and Reading				√		
	Improvement in Mathematics				√		
14	Precertification Nurses Reduced Inpatient Admissions		√				
	Onsite Concurrent Review Nurses Reduced:						
	Expense for Ancillary Services		√				
	Bed Days per 1,000 EEs		√				
15	No Scientific Evidence Supports a Causal Effect of Childhood Vaccinations on Autism-Spectrum Disorders.						

[a] These are structural equation models estimated by AMOS. All of the other multilevel models in this book were estimated using SAS's Proc Mixed or Proc Glimmix.

Endnotes

[1] To avoid confusing the levels of causality with the levels of the variables in the multilevel models, I spell their levels of causality (i.e., zero, one, and two) whereas I designate the hierarchical levels of the multilevel models and their variables using numbers (i.e., level-1, level-2, and level-3).

[2] Rubin ([1984] 2006, 27) has reprinted William Cochran's reading list for Statistics 284, circa 1968. The books and studies on this list exemplify classic causality and robust dependence; they were a starting point for Rubin's subsequent advancement of statistical theory and method.

[3] Rubin ([1984] 2006, 7) gives this example of an observational study:

> An analysis of health records for samples of smokers and nonsmokers from the U.S. population is an observational study. The obvious problem created by observational studies is that there may exist systematic differences between the treatment groups besides treatment exposure, and so any observed differences between the groups (e.g., between smokers and nonsmokers) with respect to an outcome variable (e.g., incidence of lung cancer) might be due to confounding variables (e.g., age, genetic susceptibility to cancer) rather than to the treatments themselves. Consequently, a primary objective in the design and analysis of observational studies is to control, through sampling and statistical adjustment, the possible biasing effects of those confounding variables that can be measured: a primary objective in the evaluation of observational studies is to speculate about the remaining biasing effects of those confounding variables that cannot be measured.

[4] Computer simulation studies find that variables that only affect treatment assignment but not the response (i.e., instrumental variables) are best not included as a covariate, but those variables that affect the response but not treatment assignment can be included. (Presentation to the Boston Chapter of the American Statistical Association by Til Stürmer, MD, MPH, May 4, 2010.)

[5] See the data for 2009 and the time series in the Stephen Roth report (2010).

[6] Heckman, Pearl, Robins, and Rubin and their colleagues are developing comprehensive general theories of causality. These general theories would incorporate the three notions of causality as special cases and would extend their conceptualizations and procedures, thus forming a new level-three notion of generalized causal relations.

Glossary

Absolute properties Properties of members of a collective that researchers measure without using information about the characteristics of the collective or about relationships among the members. A person's gender is an absolute property.

Akaike information criterion (AIC) Can test the relative goodness of fit of models estimated by maximum likelihood (ML), or if restricted maximum likelihood (REML) estimation is used, the relative fit of models that have the same covariates but different random effects.

All causes model A view of causality grounded in the production model of traditional economics: it relates inputs to outputs and assesses the causes of various effects.

Analytical properties Properties of a collective that researchers obtain by performing some mathematical operation upon some property of each single member. The proportion female in a collective is an analytical property of the collective.

Association graphs Depict the two-factor relationships of a loglinear model by vertices and edges; the edges that connect the two vertices (A and B) are the two-factor interactions (AB).

Association The correlational relationship between two nominal or ordinal attributes.

Attributable risk The difference between the incidence of cases among those exposed to the risk and the incidence among those not exposed; it is the risk difference.

Autoregressive, AR(1), covariance structure For a repeated measure that exhibits change over time, the first order AR(1) structure posits that the correlation of one measure on a variable with another measure on the same variable that is close in time, is stronger than the correlation of that measure with another measure on that same variable that is distant in time: $r_{12} = r_{23}$ but $r_{12} > r_{13}$; $r_{13} = (r_{12})(r_{12}) = r_{12}^2$.

Backward selection This model-finding procedure begins with a saturated model that includes all possible relationships among the variables. The algorithm then successively eliminates factors that have insignificant effects.

R.B. Smith, *Multilevel Modeling of Social Problems: A Causal Perspective*,
DOI 10.1007/978-90-481-9855-9, © Springer Science+Business Media B.V. 2011

Bayesian information criterion (BIC) Can be used to test the relative goodness of fit of models estimated by maximum likelihood (ML), or if restricted maximum likelihood (REML) estimation is used, the relative fit of models that have the same covariates but different random effects.

Bias Refers to the ways that study design and analytic methods may lead to systematic errors in the estimates of the true effect sizes.

Block A set of variables that have the same priority usually indicated by having the same time-ordering.

Borrowing strength One model is estimated that incorporates relevant information from numerous units that have similar properties rather than estimating models for each individual unit separately. The variability of the individual units is shrunk to a more stable, average value.

Can it? Causality Asks if at least one case unequivocally shows that an outcome was caused by a putative cause.

Causal inference, the fundamental problem Prior to assignment to a treatment, a unit (i.e., person) has potential outcomes under each of several treatments. After receiving one treatment, the realized outcome under that treatment can be observed for that unit; but the outcomes under the other treatments not received cannot be directly observed; such data are missing. Consequently, the individual causal effect cannot be calculated, but an average causal effect of the treatment can be estimated between the two (or more) groups if they are closely matched.

Causal transience A unit's response to the treatment t at an earlier time does not affect the unit's response to a treatment c at a later time.

Causality, notions of Stable association, potential outcomes, dependency networks, and general theories of causality.

Causality Holding all relevant factors constant except the implied cause x, the change in the outcome y due to the manipulated change of x, is referred to as the causal effect of the manipulated factor x.

Class variable In SAS the CLASS statement designates a continuous variable or a nominal attribute as a typology and creates binary (0,1) indicator variables.

Clustering Observations on individual units contained in one macrounit most probably are more similar than observations on individual units contained in another macrounit; the observations are clustered and not independent.

Collectives Are macrounits that contain members; that is, the level-2 units contain level-1 units; a university (level-2) contains professors (level-1).

Comparative analysis A comparison between measurements on a microvariable in different macrocontexts. For example, the countries' scores on the human development index (level-1) vary from region to region of the world (level-2).

Comparative fit index (CFI) Assesses the relative improvement of the current structural equation model M compared with the baseline model B that assumes independence among the observed variables (covariances = 0). The CFI = $1 - \hat{\delta}_M$ divided by $\hat{\delta}_B$ where $\hat{\delta}_M$and $\hat{\delta}_B$ are estimates of the noncentrality parameters of noncentral χ^2 distributions. A CFI > .90 indicates that M improves the fit.

Comparative properties Properties of a member of a collective that researchers derive by comparing a member's value on an absolute or relational property to the distribution over the entire collective. A student's percentile in the graduating class would be a comparative property of the student.

Comparison group Does not receive the treatment that the target group receives; it may receive a null treatment or an alternative treatment.

Compound symmetry (CS) covariance structure For a repeated measure that exhibits change over time, this structure posits that all on-diagonal variances are equal and all off-diagonal covariances are equal.

Conditional odds ratios The odds ratios in the conditional tables of a control variable.

Confounder A "third variable" that is prior to and associated with the putative cause and its effect; it may explain all or some of the observed correlation.

Constant effects The assumption of constant effects implies that the effect of t on every unit is the same: $\delta = Y_t(u) - Y_c(u)$ for all u in U, and that the treatment t adds a constant amount δ to the amount of the response in the control for each such unit.

Contextual effects Cross-level interactions between variables on the macrolevel and variables on the microlevel.

Contextual property The description of a member by a property of the collective; the professor teaches at a high quality academic institution.

Contextual study A multilevel analysis of macro- and microunits that quantifies the effects of level-2 variables, level-1 variables, and their cross-level interactions.

Continuation-ratio logits In a cross tabulation of a dichotomous response variable with an ordinal categorical variable of three or more categories, the algorithm for this model first calculates the effect of the first category (1) relative to all others (0). Then, it deletes the first category and calculates the effect of the second category (1) relative to the remaining categories (0), and so forth.

Continuous variables Properties of a unit that are measured by ordinal rankings or interval scales.

Control group These units do not receive the treatment the experimental group receives; they may receive a null treatment or an alternative treatment.

Controlling for a variable An $x \rightarrow y$ relationship is tested by examining that relationship in the conditional tables of a third variable t associated with x and y. Controlling for t implies that the units that have the same values of t are kept separate (i.e., held constant), and the $x \rightarrow y$ relationship is quantified for each value of t. An average value of the $x \rightarrow y$ relationship is calculated across these conditional tables and is compared to the original value of the $x \rightarrow y$ relationship. If the averaged value is much less than the original value of $x \rightarrow y$, and t is prior to x and y, then the control for t implies that the original relationship was spurious; if t intervenes between x and y, then t interprets that relationship, $x \rightarrow \mathrm{t} \rightarrow y$.

Counterfactual A hypothetical event, outcome, or relationship that could have but did not actually take place.

Covariance structure Models the random-effects parameters of the multilevel model.

Covariates Are the control variables in the structural part of a multilevel model.

Dependence, robust A variable y is robustly dependent on a prior variable x, if their relationship is maintained when a number of relevant test factors t are simultaneously controlled; that is, $yx.t_1\ t_2\ t_3\ t_4, \ldots, t_n > 0$.

Dependency networks Models of systems of potentially causal relationships among a number of blocks of variables that have different priorities usually determined by their different time orderings.

Design variables Formalize the study design so that the effect of the treatment can be estimated taking into account differences between the target and comparison groups, the passage of time, and the effects of the covariates.

Deviance Measures the difference between the -2 × log-likelihood of a less complicated current model and the saturated model that includes it as a special case.

Did It? Causality Asks if the evidence strongly suggests that the putative cause *did cause* an outcome in one or more cases.

Difference-in-differences The average effect of a treatment is determined by taking the difference between (1) the pre to post period difference between the means of a response variable in the target group, and (2) the pre to post period difference between the means of the response variable in the comparison group.

Dispersion scale factor Gauges the extent to which the response variable is either over- or under-dispersed. In a Poisson model it is the quotient of the sample variance of the response variable to its sample mean.

do(.) operator is a fundamental concept of Pearl's theory of causation. Let do $(X_i = x_i)$ or do(x_i) denote the simplest external intervention in which the variable X_i is forced to take on the value x_i. Then, the causal effect of the intervention is the difference between the naturally occurring state of X_i and the state of X_i under the external intervention, *do* (x_i). For example, let there be two perfectly matched groups of schools that are isolated from each other. Let one group not experience an intervention like comprehensive school reform, that is *do* $(X_i = X_i)$; this group produces the naturally occurring average test scores, say Y". Let the other group experience the external intervention, comprehensive school reform, *do* $(X_i = x_i)$; this group produces the forced by the intervention average test scores, say Y'. Then the average causal effect $= \mathrm{E}\,(do\,(X_i = x_i)) - \mathrm{E}\,(do\,(X_i = X_i)) =\ Y' - Y''$, which expresses the Rubin causal model in terms of Pearl's *do* (x) operator.

Dummy variables Binary indicator variables coded 1 or 0.

Edge matrix, ancestor See Variables, ancestors.

Edge matrix, parental See Variables, parents and children.

Edge, absence of Graphical models depict the absence of a direct effect of x on y by the absence of a directed arrow that would connect the variables.

Edge, directed Graphical models depict a direct linkage between two variables y and x in different blocks of variables as an arrow (i.e., a directed edge) that signifies a relationship that is conditionally dependent, when prior variables in the system are controlled (e.g., $b_{\mathrm{yx.abcde}} \neq 0$).

Edge, straight line Graphical models depict a symmetric relationship between two variables in the same block by a straight line that links the variables. In path-analytic models a curved two-headed arrow depicts a symmetric relationship.

Elaboration procedure The systematic examination of the $x \rightarrow y$ relationship under test factors t that may be prior to both x and y or intervening between x and y.

Equal footing Variables at the same level of priority; that is, they have the same time ordering or structural level and are in the same block of variables.

Events/trials format In a logistic regression the response variable may be the count of the number of successful events to the total number of events, creating a proportion of successes: 110 passed a test out of 200 people taking the test.

Expected mean squares Are used to determine the devisor for the F-statistics of the effects in an analysis of variance.

Exposure odds ratio The odds of exposure among the cases divided by the odds of exposure among the controls; it estimates the true relative risk when adverse events are rare.

Fixed part of the equation Defines the model's structural components and not its stochastic components.

Fixed-effects models Estimate the parameters of the statistical model so that inferences can be made to the data at hand. In a meta-analysis the fixed-effects model ignores inter-study variation. It assumes that there is one constant program effect and each study provides an estimate of that effect.

Fundamental problem of causal inference See Causal inference, fundamental problem of.

Gamma (γ) A measure of association between two nominal attributes that is calculated as the quotient of the difference between the concordant and discordant pairs of observations divided by their sum.

General linear models Are best suited for the analysis of continuous response variables having a normal probability distribution. These models include analysis of variance, analysis of covariance, and regression procedures.

Generalized linear mixed model (GLMMs) These generalized linear models can include both fixed and random effects in models of response variables that have either normal or non-normal probability distributions.

Generalized linear models (GLMs) These models enable the analysis of response variables that have either normal or non-normal probability distributions. The mean of the population is modeled as a function of a fixed linear predictor through a nonlinear link function that allows the response probability distribution to be any member of an exponential family of distributions. These procedures include general linear models, logistic regression, Poisson regression, probit models, and so forth.

Generative mechanism (in agent-based modeling) The interactions of a number of microlevel agents who are acting in pursuit of their own goals produce macrolevel outcomes as consequences of those interactions.

Generative mechanism (in social research) A pattern of intervening interpretive variables that link x to y: $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \rightarrow y$.
Genuine correlation A correlation that is still strong after numerous allowable and relevant test factors have been simultaneously controlled.
Global properties Properties of a collective that are not based on properties of its members.
Granger causality A view of causality specific to time-series analysis: the hypothesized cause X can predict a future event or value Y, even when the information in the allowable test factor Z is taken into consideration.
Granger noncausality A view of spuriousness specific to time-series analysis: the hypothesized cause X cannot predict a future event or value Y, when the information in the allowable prior test factor Z is taken into consideration.
Graphs, interpretive Depict the results of statistical analyses using such visual aids as boxes that signify blocks of variables; points that signify variables; and arrows, straight lines, and curved two-headed arrows that signify relationships.

Hierarchical loglinear models Are nested so that that a less complex model is composed of some but not all of the elements of a more complex model.
Hierarchical models A synonym for multilevel models and for mixed models: there is a specific equation for level-2 variables and a specific equation for level-1 variables, which can be combined and estimated, either separately or in combination.
Homogeneous subgroups Can reduce the bias when means y are compared in target and comparison groups. The bias results because y may be related to a variable t whose distributions vary in the two groups. By examining the difference between the means in the target and comparison groups in six subgroups that are formed so that they have the same value of t, and then averaging these differences across these homogeneous subgroups, the bias in the overall estimate will be reduced. See Propensity scores for an application of this procedure. Also see Nesting.
Hypothetical See Counterfactual.

Ignorability A stipulation that the treatment assignment mechanism does not affect the outcomes; selection bias is zero.
***iid* — independent and identically distributed** An assumption that the errors each have the same probability distribution (e.g., are normally distributed), are uncorrelated with each other, and are independent.
Incidence rate The number of new cases of a disease occurring in a specified time period divided by the total number in the population at risk during that time period.
Independence A variable y is independent of a prior variable x if their initial marginal relationship is null. If their relationship becomes null when a number of prior test factors t are simultaneously controlled, then the two variables are conditionally independent: $yx.t_1\, t_2\, t_3\, t_4, \ldots, t_n = 0$. However, two marginally independent variables may become conditionally associated if they jointly determine a third variable.

Information criteria Statistics such as the AIC and BIC that are used to determine the relative goodness of fit of alternative models of the data. These statistics adjust the maximized log likelihoods by a penalty factor that favors more parsimonious models; the BIC's penalty usually is more severe than the AIC's.

Interaction effect The coefficient on the product of two variables (*xz*) that also have singular influences on *y*. That is, $y = b_0 + b_1x + b_2z + b_3(xz)$.

Interpretation If the hypothesized $x \rightarrow$ y relationship disappears when a test factor *t* that intervenes between *x* and *y* is controlled, then the initially observed relationship between *x* and *y* is interpreted by *t*: $x \rightarrow t \rightarrow y$.

Intraclass correlation coefficient, *ρ* (rho) The intraclass correlation coefficient quantifies the amount of clustering. It is calculated as the ratio of the unexplained level-2 variance of the response variable to the sum of the level-1 and level-2 unexplained variances of the response variable.

Kendall's tau (*τ*) A measure of symmetric association for the analysis of ordinal variables, $\tau^2 = (d_{\mathrm{yx}})(d_{\mathrm{xy}})$; Somers's d_{yx} is a component of Kendall's tau.

Level-zero causality A synonym for stable association; here it includes classic causality, robust dependence, Granger causality, results from meta-analyses, and so forth.

Level-one causality A synonym for causality under the potential outcomes perspective; here it includes causality as an effect of an intervention and a stochastic causal model for attributes.

Level-two causality A synonym for multivariate dependencies that explain processes. Here it includes graphical models, association graphs for loglinear models, generative processes, and all-causes structural equation models.

Level-three causality General models of causality that contain the first three levels as special cases.

Likelihood-ratio test Can compare two models in which one is nested within the other. For example, the significance of a level-2 variance component can be tested by holding constant the fixed covariates and comparing that model to a similar model in which the level-2 variance parameter is absent. If the difference in $-2LL_R$ between the two models is statistically significant using a likelihood ratio χ^2 (chi square) test with one degree of freedom (for the missing parameter), then the null hypothesis H_0 of no difference is rejected. Rejection of H_0 implies that the alternative hypothesis H_1 is preferred; that is, the better model includes a level-2 variance component.

Log odds The natural logarithm (ln) of the odds ratio: $\ln(\theta) = \ln(\mathrm{odds}_1 / \mathrm{odds}_2) = \ln[\pi_1 / (1 - \pi_1)] / [\pi_2 / (1 - \pi_2)]$.

Logit model A logistic regression model in which the dichotomous response variable is transformed into a logit by taking the natural log of its odds. The transformed response variable is then regressed on a set of covariates whose effects can be interpreted on the logit scale, as odds ratios, or as differences in proportions.

Logit The natural logarithm (ln) of the odds of success: $\ln(\pi / (1 - \pi))$ where π is the probability of success and $(1 - \pi)$ is the complementary probability of failure.

Macrolevel Variables that contain other variables at a lower level in the data hierarchy.

Macrounits Are higher level units in a multilevel data structure. In a structure with two levels they are at level-2 and they contain level-1 units.

Marginal correlation The correlation between two variables x and y when there are no control variables.

Marginal odds ratio The odds ratio in a bivariate table that is formed by summing over the conditioned (i.e., partial) cells of a control variable. Alternatively, it is the odds ratio in a bivariate table that subsequently is elaborated by controls.

Matching A form of control for spurious factors in which very similar units are paired and one unit in the pair is given the treatment of interest and the other is given a null or alternative treatment.

Maximum likelihood estimation (ML) Involves the observed data, a likelihood function, and maximum likelihood estimators. The likelihood function expresses the probability of the observed data as a function of its unknown parameters. ML estimation finds the parameters of the likelihood function that maximize the probability of obtaining the observed data.

Mechanism In social research a mechanism is a chain of relationships linking the response variable to a number of prior variables: $x \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \rightarrow y$; or it is a contextual effect — a professor's apprehension is a function of her permissiveness, (level-1), incidents (level-2), and their cross-level interaction.

Members Are microunits contained by macrounits; that is, level-1 units grouped within level-2 units. Synonyms of grouped are contained, nested, and included.

Meta-analysis A statistical study that combines information from several other studies to determine an overall summary effect δ of a treatment or program and its statistical significance. If the results of the studies are homogeneous, then δ is properly a fixed-effects estimator. If the results of the studies exhibit significant between-study variability, then the random-effects estimate of δ should be determined along with its significance.

Microunits In a data hierarchy of two levels, a microunit is at level-1 and it is grouped by a level-2 variable.

Mixed models A synonym for multilevel models with fixed and random effects. Such models have the form $\mathrm{Y} = \mathrm{X}\beta + \mathrm{Z}\gamma + \varepsilon$ where the term $\mathrm{X}\beta$ models the fixed effects, $\mathrm{Z}\gamma$ models the random effects, and ε models the residual errors.

Model chi square This basic measure of fit for a structural equation model tests the χ^2_M of the current somewhat parsimonious model M that has some unused degrees of freedom ($df > 0$) against the fit of a "just identified" saturated model that fits the data perfectly with no spare degrees of freedom ($\chi^2 = 0$, $df = 0$). The null hypothesis H_0 posits that the current model fits the data as well as the saturated model. Aχ^2_M probability of fit that is greater than 0.05, ideally much greater (say 0.25 or so), is consistent with the view that the parsimonious model fits the data; whereas a probability less than 0.05 suggests that the hypothesized model does not fit the data; here non rejection of H_0 implies that model M fits.

Mutual independence loglinear models Contain only single-factor terms, which are also referred to as the main effects. Such models are not composed of two-factor or higher interactions, each pair of variables is independent. Thus, these are "main-effects only" models.

Nested models A complex model nests a less complex models when some but not all of the parameters that appear in the more complex model are shared by the less complex model and that model has no other parameters that do not appear in the more complex model. A complex model with a fixed set of covariates and two covariance parameters nests a simpler model with the same set of covariates and only one covariance parameter.

Nesting For example, the nesting of the global region of a country within the categories of the amount of democracy of the country is denoted as region(democracy). This nesting of region by democracy creates region × democracy subgroups that may have smaller values of random effects than the random effects of the different regions when region is not nested by democracy. The nested variable region never appears as a main effect in the multilevel model.

No three-factor interaction loglinear models These contain no three-factor interactions and by implication no four-factor or higher-factor interactions as well. Such models assume homogeneous odds ratios for each pair of variables. For example, no three factor interaction among vote, the character issue, and the environmental issues implies that the association between vote and character is the same in the conditional sub-tables defined by the categories of the environmental issue. Similarly, the association between vote and the environmental issue is the same in the conditional sub-tables defined by the categories of the character issue. Contrarily, a three-factor interaction of the vote with the character issue and with the environmental issue implies that the association between vote and character varies in the sub-tables conditioned on the environmental issue. It also implies that the association between vote and the environmental issues varies in the sub-tables conditioned on the character issue.

Nominal attributes Properties of a unit that are measured by attributes, classifications, and typologies whose categories have no intrinsic order.

Non centrality parameter Is designated by δ; it expresses the degree of misspecification of a structural equation model by taking the difference between the model chi square χ^2_M and its degrees of freedom df_M: $\hat{\delta}_M = \max\,(\chi^2_M - df_M\,, 0)$. The estimate of $\hat{\delta}_M$ is the larger of zero or the difference. Smaller values of δ are preferred.

Null hypothesis If a researcher believes that men are on average taller than women, then a null hypothesis is $H_{0:}$ height (men) = height (women). If appropriate samples of data indicate that height (men) is significantly greater than height (women), then the null hypothesis is rejected, implying that height (men) > height (women).

Observational study Captures the effect of a naturally occurring treatment, or that of a manipulated treatment in which units are not randomly assigned, by distinguishing between the target group of units that received this treatment and a comparison group of units that did not receive that treatment. The measure on a response variable in the target group is compared to the measure on that response variable in the comparison group, with controls for spuriousness via matching and regression analysis.

Odds ratio A measure of association for a four-fold table that is the ratio of two odds: $\theta = \text{odds}_1 / \text{odds}_2 = [\pi_1 \ (1 - \pi_1)] / [\pi_2 \ (1 - \pi_2)]$. Such statistical methods as loglinear models and logistic regression utilize odds and odds ratios rather than proportions because they have more desirable statistical properties for the analysis of dichotomies and trichotomies.

Odds The probability of success π divided by the probability of not success $(1 - \pi)$; odds $= \pi / (1 - \pi)$. Odds are always a positive number: odds greater than one indicate that success is more likely than not success; odds less than one indicate the opposite. Odds equal to one indicate that the two probabilities are equal.

Over-dispersion In a Poisson process the mean of the distribution should equal the variance, so their quotient Variance/Mean equals unity. When the variance is greater than the mean, their quotient will be greater than unity, thereby exhibiting over-dispersion in the data. If the quotient is less than unity, then the data are under-dispersed.

Overlap A stipulation of the potential outcomes perspective that observations on each covariates appear about equally in the comparison and target groups.

Partial correlation A correlation between two variables x and y when there is at least one control variable z; there usually will be more than one control variable. When there is only one control, a synonym is first-order correlation; when there are three controls, the correlation is referred to as third-order, and so forth.

Path diagram Depicts variables as boxes, asymmetric relationships among variables by straight arrows, and unanalyzed correlations usually between exogenous variables as two-headed arrows. The fully standardized path coefficients (β weights) are entered near the appropriate arrow. Above the box that is the target of the arrow, the R^2 of that regression is entered. The effect of extraneous variables on that response variable is depicted by the arrow from a circular unmeasured variable to that response variable. The residual path coefficient equals the square root of the quantity $(1 - R^2)$ for that response variable.

Path regression diagram Depicts variables as boxes, asymmetric relationships among variables by straight arrows, and unanalyzed covariances usually between exogenous variables as two-headed arrows. The unstandardized path coefficients (b weights) are entered near the appropriate arrows. The effect of extraneous variables on a response variable is depicted by the arrow from a circular unmeasured variable to that response variable. Above the circular unmeasured variable the diagram reports the variance estimate of that variable.

Potential Outcomes A notion of causality grounded in experimental studies in which prior to assignment to a treatment a subject has potential outcomes under each of the alternative treatments. After assignment to a specific treatment the subject has a realized outcome under that treatment and counterfactual outcomes under the other treatments. The causal effect of the treatment on the subject cannot be calculated because only the realized outcome is observed. This fact illustrates the fundamental problem of causal inference. However, average treatment effects can be calculated based on the mean responses to the various treatments of closely matched groups of subjects, with or without randomized assignment to the treatments.

Precision The accuracy of a parameter as estimated by the inverse of its variance. When the variance is estimated from empirical data, the precision is estimated by the inverse of its squared standard error.

Prima facie cause A self-evident or apparent causal relationship x → y that is sufficiently documented so that it can be assumed to hold until it is disproved.

Probit model Is a special case of generalized linear models that can be used to model the groups of units classified on a response variable, the classification of which is based on a threshold value. This model assumes that the response is a manifestation of a continuous unobserved implicit variable. The cut point forming the groups is the threshold. For example, political partisanship can be viewed as a continuous unobserved variable on which a value at or above the threshold value indicates Democrat partisanship (1) and a value below the threshold indicates Republican partisanship (0). This dichotomy can be regressed on a set of covariates by specifying the probit link as an option in programs for estimating generalized linear models or generalized linear mixed models using the events/trials format.

Propensity scores Allow the control for numerous prior variables as they may affect the $x \rightarrow y$ relationship. Membership in the target group is coded 1 and membership in the comparison group is coded 0. Applying an appropriate statistical model (usually logistic or probit regression) this binary indicator is then regressed on a large number of confounding variables that are prior to a unit's treatment assignment and to the response variable. The probability of a unit being in the target group is the unit's propensity score. Units are then matched on their propensity score forming a number of groups (say 5 to 10). Then, the $x \rightarrow y$ relationship is quantified in each of these groups and averaged across these groups. That value of that average effect is the putative average causal effect of x on y. Propensity scores can also be used as a covariate in the regression of y on x, along with other covariates, if the probabilities are not too extreme.

Proportional-odds model In a cross tabulation of a dichotomous response variable with an ordinal variable of three or more categories, this model holds that the odds ratios in successive fourfold tables are equal. The model's goodness of fit is tested against the null hypothesis $H_0 = \beta_k = \beta$ for all k; that is, the log odds ratios are the same in all fourfold tables. Probabilities less than 0.05 reject this hypothesis.

Random shocks In Coleman's causal model these are the effects impinging on the transition rates between the two states of a response variable that are due to extraneous variables that are not explicitly measured. When no causal variables are operating, the random shocks in the positive direction equal the intercept.

Random-effects models Estimate the parameters of the statistical model so that inferences can be made from the data at hand to the population from which the data are (at least ideally) a random sample. The effects of a factor of the model may vary across the units. In a meta-analysis, for example, the random-effects model quantifies interstudy variation. It assumes that there is a distribution of true effects and that each study has its own unique true effect that would result if it were replicated an infinite number of times. Since the random-effects model aims to estimate the grand-mean treatment effect, the range of inference is not limited to only those studies included in the consolidation. It suggests whether such a program *will* have an effect.

Recursive strategy Given that there are several blocks of variables a, b, c, and d, start with the ultimate response variable in block a and regress it on all of the variables in the prior blocks, b, c, and d. Then, regress variables in block b on the variables in blocks c and d. Then, regress variables in block c on those in block d. Display the results in tables and in an interpretive regression graph.

Relational properties Are properties of members of a collective that researchers derive from information about relationships among the members. For example, based on sociometric matrices of interactions in housing units, a person may belong to a cohesive housing unit or to a unit that is not cohesive.

Relative risk The ratio of the incidence of cases among those exposed to the risk to the incidence among those not exposed.

Restricted maximum likelihood (REML) When the response variable is normally distributed, the default method of estimation in Proc Mixed is REML. To estimate the parameters of the multilevel model, REML minimizes the likelihood of residuals from fitting the fixed-effects portion of the model, or equivalently, by minimizing the negative of the log likelihood. When REML is used rather than maximum likelihood (ML), such goodness-of-fit statistics as the AIC and BIC are best applied to compare models that have the same fixed variables but different random-effect parameters.

Risk difference (or attributable risk) The difference between the incidence of cases among those exposed to the risk and the incidence among those not exposed.

Robust dependence A variable y is robustly dependent on a prior variable x, if their relationship is maintained when a number of prior test factors t are simultaneously controlled. The two variables are conditionally dependent: $yx.t_1\, t_2\, t_3\, t_4, \ldots, t_n > 0$.

Root Mean Square Error of Approximation (RMSEA) A statistic used to assess the fit of structural equation models. Taking sample size into account, it estimates the amount of error of approximation per model degree of freedom. The formula divides the square root of the non centrality parameter for the model by the product of the model's degree of freedom times the sample size minus 1: RMSEA $= (\delta_M \,/\, df_M\, (N-1))^{1/2}$. Values close to zero (RMSEA < 0.05) and a confidence interval between 0 and 0.10 indicate a close fit of the model to the data.

Root mean square residual (RMR) This measure of fit for structural equation models is based on the differences between actual and predicted covariances; it is the square root of the squared values of these differences; smaller values indicate a better fit. When this statistic is based on actual and predicted correlations among the variables, then values less than 0.10 are indicative of a close fit.

Saturated loglinear model Includes the maximum possible model parameters and fits the data perfectly; its deviance is zero.

Scaled deviance Is the quotient when the deviance of the model is divided by the parameter that gauges the over- or under-dispersion of the model. Over-dispersion (or under-dispersion, as the case may be) is the quotient of the sample variance divided by the sample mean. When that quotient is greater than unity the model is fitted to over-dispersed data. For example, a Poisson model with a deviance = 50 and an over-dispersion parameter of 3.81 has a scaled deviance = 50/3.81 = 13.1.

Selection bias Occurs when the subjects in the treatment arms of an experiment or an observational study have prior characteristics that are associated with the outcomes and the treatment arms are not balanced on these confounders.

Serially uncorrelated random shocks In Coleman's causal model the error terms are independent from one observation to the other. In Rubin's causal model extraneous variables (random shocks) should not influence the assignment of a unit to the treatment or control groups, or the response variable.

Shrinkage When a statistical procedure borrows strength, the estimates for the individual units exhibit shrinkage: they are shrunk to an average value for all of the units.

Simpson's paradox The observation that a marginal association between x and y can have a different sign of direction than that of a conditional association between x and y, when other variables are controlled.

Somers's d_{yx} Is a measure of asymmetric association for the analysis of ordinal variables; it is conceptualized as an analog of an unstandardized regression coefficient.

Spurious correlation If the hypothesized $x \rightarrow$ y relationship disappears when a test factor t that is prior to both x and y is controlled ($yx.t = 0$), then the initially observed relationship between x and y is spurious. That is, x and y are conditionally independent given a control for a prior t.

Stable association If the $x \rightarrow$ y relationship is still robust after a number of antecedent test factors are controlled sequentially or simultaneously, then the association (i.e., correlation) between x and y is thought to be stable. If a number of appropriate test factors are controlled simultaneously and the $x \rightarrow$ y relationship is still robust, then the dependence is robust.

Stable-Unit-Treatment-Value Assumption (SUTVA) In the potential outcomes perspective the response of a unit u when exposed to a treatment t should be the same regardless of the mechanism used to assign the treatment t to the unit u and regardless of what treatment the other units receive. If in fact there are two treatments applied to some units in the single-treatment group, or if there is leakage of the treatment to the units in the control group, then SUTVA is violated. SUTVA is violated if the assignment to the treatment group is confounded with the outcomes. In an observation study with treatment and comparison groups composed of schools, and with the test scores of the students being the response variable, if school administrators or teachers in a treatment school exclude students with low grades from taking the criterion achievement test, then the school's performance would be artificially inflated; this violates SUTVA because the outcome is confounded with assignment to the treatment group.

Stochastic part of an equation The section of an equation for a multilevel model that contains its random effects but not its structural, fixed components.

Structural properties Properties of a collective that researchers obtain by performing some operation on data about the relations of each member to some or all of the others. Based on sociometric matrices of interactions of residents of a housing unit, the unit can be classified as cohesive or not cohesive.

Temporal stability In Coleman's causal model, for a fixed set of explanatory variables the transition rates do not change over time. Similarly, Rubin's model assumes constancy of a response to a treatment.

Typology Classifies units as belonging to one of a number of qualitative categories.

Unconfoundedness A stipulation of the potential outcomes perspective that the assignment of a unit to either the target or comparison treatment should not be confounded with the potential outcomes.

Under-dispersion The data of a Poisson regression model, for example, exhibits under-dispersion when the quotient of the sample variance of the response variable divided by its sample mean is less than unity.

Unit homogeneity In Coleman's model individual's characterized by the same pattern of measures on explanatory variables (say + + -) will have the same transition rates, which will differ from individuals with a different pattern (say - + -).

Units Analytic entities that have properties; a unit may be a member of a collective (i.e., a microunit) or the collective (i.e., a macrounit).

Un-nested random variable A random variable that is not nested by a fixed covariate.

Variables, ancestor In a graphical model in which there is a precedence ordering among the variables, a parent variable has direct descendants that are referred to as its children, but the parent may be a child of another parent variable, which is referred to as an ancestor of the child: $x \rightarrow y \rightarrow$ z. Thus z is a child of y and x the parent of y is an ancestor of z. An ancestor edge matrix has rows and columns for each variable in the graphical system. Each row signifies a possible response variable, a child, and each column a possible parent or ancestor variable. In the row for a child a 1 is entered for each direct parent or ancestor, and 0 for each variable that is not a parent or an ancestor of that child.

Variables, parents and children In a graphical model in which there is a precedence ordering among the variables, a parent variable has direct descendants referred to as children; thus x $\rightarrow y$ implies that x is a parent of the child y. A parental edge matrix has rows and columns for each variable in the graphical system. Each row signifies a possible response variable or child and each column a possible parent variable. In the row for a specific child, a 1 is entered for each direct parent and a 0 for each variable that is not a parent.

Variance Components Is the default covariance structure for normally distributed, continuous variables when modeled by Proc Mixed. "Variance components" also refers to the covariance structures that can model the random effects.

These include, in addition to variance components, AR(1), banded main diagonal, compound symmetry, Toeplitz, unstructured, and so forth. One variance component gauges the variance between the macrounits and the other gauges the variance among the microunits contained by the macrounits.

Will it? Causality Asks how frequently a putative cause *will cause* a specific outcome in individuals or populations.

Zone of Causality Because the causal level of findings in the social sciences is at best ambiguous, it is more efficacious to assign a study's results to one of three approximate zones of causality rather than to a precise level of causality. One can then explicate the shortcomings of the study with the goal of improving subsequent studies and perhaps moving their results to a more precise, enhanced level of causality.

Acknowledgements These definitions, which the author derived, synthesized, and edited from passages in the chapters of this book and from other sources, aim to facilitate the reader's comprehension of multilevel modeling. The chapters of this book cite the various sources, and these sources appear in the References along with other cited works.

References

Aaron, Henry J., and Robert D. Reischauer. 1995. The medicare reform debate: What is the next step. *Health Affairs* 14(4): 8–30.

Aaron, Henry J., and Robert D. Reischauer. 1998. Comment on 'Rethinking medicare reform'. *Health Affairs* 17(Jan/Feb): 69–71.

Achen, Christopher H. 1986. *The statistical analysis of quasi-experiments*. Berkeley: University of California Press.

Adelman, Irma. 2001. Fallacies in development theory and their implications for policy. In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 103–134. New York: Oxford University Press.

Agresti, Alan. 1996. *An introduction to categorical data analysis*. New York: Wiley.

Alwin, Duane F., and Robert M. Hauser. 1975. The decomposition of effects in path analysis. *American Sociological Review* 40: 33–47.

Aladjem, Daniel K., Kerstin Carlson LeFloch, Yu Zhang, Anja Kurki, Andrea Boyle, James E. Taylor, Suzannah Herrmann, Kazuaki Uekawa, Kerri Thomsen, and Olatokunbo Fashola. September 2006. *Models matter—the final report of the national longitudinal evaluation of comprehensive school reform*. Washington, DC: American Institutes of Research.

Ansolabehere, Stephen, Shanto Iyengar, and Adam Simon. 1999. Replicating experiments using aggregate and survey data: The case of negative advertising and turnouts. *American Political Science Review* 93: 901–909.

Anti-Defamation League. 2002a. *European attitudes toward Jews, Israel and the Palestinian-Israeli Conflict*. New York, NY: 823 United Nations Plaza. June 27.

Anti-Defamation League. 2002b. *European attitudes toward Jews: A five-country survey*. New York, NY: 823 United Nations Plaza. October.

Anti-Defamation League. 2004. *European attitudes toward Jews, Israel and the Palestinian-Israeli Conflict in ten European Countries*. New York, NY: 823 United Nations Plaza. April.

Anti-Defamation League. 2005. *Attitudes toward Jews in twelve European countries*. New York, NY: 823 United Nations Plaza. April.

Anti-Defamation League. 2009. *Attitudes toward Jews in seven European countries*. New York, NY: 605 Third Avenue. February.

Arbuckle, James L., and Werner Wothke. 1999. *Amos 4.0 user's guide*. Chicago, IL: SmallWaters Corporation.

Arbuckle, James L. 2005. *Amos 6.0 user's guide*. Chicago, IL: SPSS Inc.

Ashaf, Haroon. 2001. US expert group rejects link between MMR and autism. *The Lancet* 357 (9265, 28 April):1341.

Back, Kurt W. 1951. The exertion of influence through social communication. *Journal of Abnormal and Social Psychology* 46: 9–23.

Bailit, Howard, John Federico, and William McGivney. 1995. Use of outcomes studies by a managed care organization: Valuing measured treatment effects. *Medical Care* 33(4): AS216–AS225.

Bainbridge, William Sims, Kathleen M. Carley, David R. Heise, Michael W. Macy, Barry Markovsky, and John Skvoretz. 1994. Artificial social intelligence. *Annual Review of Sociology* 20: 407–36.

Bales, Kevin. 1999. *Disposable people*. Berkeley: University of California Press.

Bales, Kevin. 2000. *New slavery*. Santa Barbara, CA: ABC-Clio.

Bales, Kevin. 2002. The social psychology of modern slavery. *Scientific American* 286(4, April): 82–88.

Bales, Kevin. 2004. *International labor standards: Quality of information and measures of progress in combating forced labor*. Washington, DC: Free the Slaves.

Barro, Robert. 1999. Determinants of democracy. *Journal of Political Economy* 107(6): S158–S183.

Barro, Robert, and Rachel M. McCleary. 2003. Religion and economic growth across countries. *American Sociological Review* 65: 760–781.

Basu, Kaushik. 2001. On the goals of development. In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 61–86. New York: Oxford University Press.

Beckfield, Jason. 2003. Inequality in the world polity: The structure of international organization. *American Sociological Review* 68: 401–424.

Beinart, Peter. 2010. The failure of the American Jewish establishment. *The New York Review of Books* 57(10, June 10):16, 18, 20.

Berends, Mark, Susan Bodily, and Sheila Nataraj Kirby. 2002. *Facing the challenges of whole school reform: New American schools after a decade*. Santa Monica, CA: RAND.

Berg-Schlosser, D. 2003. Comment on Welzel, Inglehart and Klingemann's 'The theory of human development: A cross-cultural analysis'. *European Journal of Political Research* 42: 381–386.

Berk, Richard A. 1988. Causal inference for sociological data. In *Handbook of sociology*, ed. Neil J. Smelser, 155–172. Newbury Park, CA: Sage.

Bernard, Tara S. 2010. For consumers, clarity on health care reform. *New York Times*, March 21, downloaded from the "your-money" section.

Bickerton, Ian. 2004. From tolerance to division? Murder and extremism shake Dutch faith in an open society. *Financial Times*, November 17:15.

Blalock Jr., Hubert M. [1961] 1964. *Causal inferences in nonexperimental research*. Chapel Hill: University of North Carolina Press.

Blau, Peter M. 1960. Structural effects. *American Sociological Review* 25: 178–193.

Blauner, Robert. 1964. *Alienation and freedom*. Chicago, IL: University of Chicago Press.

Blendon, Robert J., and John M. Benson. 2009. Understanding how Americans view health care reform. *New England Journal of Medicine* 361(9): e13(1-4). Downloaded from www.nejm.org on September 5, 2009.

Blumer, Herbert. 1956. Sociological analysis and the variable. *American Sociological Review* 21 (December): 683–690.

Blumer, Herbert. 1969. *Symbolic interactionism: Perspective and method*. Englewood Cliffs, NJ: Prentice-Hall.

Bollen, Kenneth A., and R.W. Jackman. 1985. Political democracy and the size distribution of income. *American Sociological Review* 50: 438–457.

Borman, Geoffrey D., and Gina M. Hewes. 2002. The long-term effects and cost-effectiveness of success for all. *Educational Evaluation and Policy Analysis* 24: 243–266.

Borman, Geoffrey D., Gina M. Hewes, L.T. Overman, and S. Brown. 2003. Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research* 73: 125–230.

Boudon, Raymond. 1996. The 'cognitivist model': A generalized rational choice model. *Rationality and Society* 8(May): 123–150.

Boudon, Raymond. 2003. Beyond rational choice theory. *Annual Review of Sociology* 29: 1–22.

Boyle, Richard P. 1966. Causal theory and statistical measures of effect: A convergence. *American Sociological Review* 31: 843–841.

Brenner, M. Harvey. 1971. *Time series analysis of relationships between selected economic and social indicators*, vol. 1: Text and Appendices. New Haven, CT: Yale University School of Medicine and Department of Sociology, March.

Brenner, M. Harvey. 1973. *Mental illness and the economy*. Cambridge, MA: Harvard University Press.

Brenner, M. Harvey. 1989a. Economic change and mortality in first world countries: Post war to mid 1980s. In *Did the crisis really hurt? Effects of the 1980–1982 recession on satisfaction, mental health and mortality*, ed. Ruut Veenhoven, 174–220. The Netherlands: Universitaire Pers Rotterdam.

Brenner, M. Harvey. 1989b. Mortality and economic change in post-war Netherlands: 1950–1985. In *Did the crisis really hurt? Effects of the 1980–1982 recession on satisfaction, mental health and mortality*, ed. Ruut Veenhoven, 237–258. The Netherlands: Universitaire Pers Rotterdam.

Brenner, M. Harvey. 2005. Commentary: economic growth is the basis of mortality rate decline in the 20th century—experience of the United States 1901–2000. *International Journal of Epidemiology* 34: 1214–1221.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Bryk, Anthony S., Valerie E. Lee, and Peter B. Holland. 1993. *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.

Brym, Robert, and Ajay Bader. 2006. Suicide bombings as strategy and interaction: The case of the Second Intifada. *Social Forces* 84: 1967–1985.

Burd-Sharps, Sarah, Kristen Lewis, and Eduardo Borges Martins. 2008. *The measure of America: American human development report 2008–2009*. New York: Columbia University Press.

Burd-Sharps, Sarah, Kristen Lewis, and Eduardo Borges Martins. 2009. *A Portrait of Mississippi: Mississippi human development report 2009*. Brooklyn, NY: American Human Development Project.

Burris, Beverly H. 1998. Computerization of the workplace. *Annual Review of Sociology* 24: 141–157.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carley, Kathleen M. 1988. Formalizing the social expert's knowledge. *Sociological Methods and Research* 17: 165–232.

Carr, Gregory J., Kerry B. Hafner, and Gary G. Koch. 1989. Analysis of rank measures of association for ordinal data from longitudinal studies. *Journal of the American Statistical Association* 84(407): 797–1989.

Central Intelligence Agency. 2003. *CIA world factbook*. Washington, DC.

Christensen, Ronald. 1997. *Log-linear models and logistic regression*, 2nd ed. New York: Springer.

Cicourel, Aaron Victor, and John I. Kitsuse. 1963. *The educational decision-makers*. Indianapolis, IN: Bobbs-Merrill.

Claeskens, Gerda, and Nils Lid Hjort. 2008. *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.

Clark, Ann Michele. 2003. Trafficking in persons: An issue of human security. *Journal of Human Development* 4(2, July): 247–263.

Cochran, William G. 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* (June): 205–313.

Cochran, William G., and Donald B. Rubin. 1973. Controlling bias in observational studies: A review. *Sankya* Series A, 35: 417–446.

Coleman, James S. 1964. *Introduction to mathematical sociology*. New York: Free Press.

Coleman, James S. 1970. Multivariate analysis for attribute data. In *Sociology Methodology 1970*, ed. Edgar F. Borgatta and George W. Bohrnstedt, 217–245. San Francisco, CA: Jossey-Bass.

Coleman, James S. 1972. *Policy research in the social sciences*. Morristown, NJ: General Learning Press.

Coleman, James S. [1973] 2006. *The mathematics of collective action*. New Brunswick, NJ: Aldine Transaction.

Coleman, James S. 1981. *Longitudinal data analysis*. New York: Basic Books.

Coleman, James S. 1990. *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Comprehensive School Reform Quality Center. 2006. *CSQR Center report on elementary school comprehensive school reform models*. Washington, DC: American Institutes for Research.

Cook, Deborah J., Gordon H. Guyatt, Gerard Ryan, Joanne Clifton, Lisa Buckingham, Andrew Willan, William McIlroy, and Andrew D. Oxman. 1993. Should unpublished data be included in meta-analyses? *Journal of the American Medical Association* 269 (No. 21, June 2): 2749–2753.

Cook, Thomas D., Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller. 1992. *Meta-analysis for explanation: A casebook*. New York: Russell Sage Foundation.

Cooper, Gregory F. 1999. An overview of the representation and discovery of causal relationships using Bayesian networks. In *Computation, causation & discovery*, ed. Clark Glymour and Gregory F. Cooper, 3–62. Cambridge, MA: MIT Press.

Cooper, Joel, and Kimberly A. Kelly. 2004. Attitudes, norms, and social groups. In *Social cognition*, ed. Kimberlee Weaver, Marilynn B. Brewer, and Miles Hewstone, 244–267. Malden MA: Blackwell Publishing.

Cox, D.R. 1992. Causality: some statistical aspects. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 155: 291–301.

Cox, D.R., and Nanny Wermuth. 1996. *Multivariate dependencies: Models, analysis and interpretation*. London, UK: Chapman & Hall.

Cox, D.R., and Nanny Wermuth. 2001. Some statistical aspects of causality. *European Sociological Review* 17: 65–74.

Cox, D.R., and Nanny Wermuth. 2004. Causality: A statistical view. *International Statistical Review* 72: 285–305.

Crafts, Nicholas. 2001. Historical perspectives on development. In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 301–334. New York: Oxford University Press.

Crain, Robert L., and Rita E. Mahard. 1983. The effect of research methodology on desegregation-achievement studies: A meta-analysis. *American Journal of Sociology* 88: 839–854.

Cutler, David M. 2000. Medicare reform: Fundamental problems, incremental steps. *The Journal of Economic Perspectives* 14(2, Spring): 21–44.

D'Agostino, Ralph B., and Heidy Kwan. 1995. Measuring effectiveness: What to expect without a randomized control group. *Medical Care* 33: AS95–AS105. Supplement.

Dahl, Robert A. 1989. *Democracy and its critics*. New Haven, CT: Yale University Press.

Dahl, Robert A. 1998. On counting democratic countries, appendix c. In *On Democracy*. New Haven, CT: Yale University Press.

Darley, John M. 2000. Bystander phenomena. In *Encyclopedia of psychology*, vol. 1, ed. Alan Kazden, 493–495. Washington, DC: American Psychological Association.

Darley, John M., and Bibb Latané. 1968. Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* 8: 377–383.

Daschle, Tom. (with Scott S. Greenberger and Jeanne M. Lambrew). 2008. Critical: What we can do about the health-care crisis. New York: Saint Martin's Press.

Davis, James A. 1971. *Elementary survey analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Davis, James A. 1975. Analyzing contingency tables with linear flow graphs: *d* systems. In *Sociological methodology 1976*, ed. David R. Heise, 111–145. San Francisco, CA: Jossey-Bass.

Davis, James A. 1980. Contingency table analysis: Proportions and flow graphs. *Quality & Quantity* 14: 117–153.

Davis, James A. 1985. *The logic of causal order*, Sage University Paper series on Quantitative Applications in the Social Sciences, vol. 55. Beverly Hills, CA: Sage Publications.

Davis, Karen, Gerard F. Anderson, Diane Rowland, and Earl P. Steinberg. 1990. *Health care cost containment*. Baltimore, MD: Johns Hopkins University Press.

Dawid, A.P. 2002. Influence diagrams for causal modeling and inference. *International Statistical Review* 70: 161–189.

Deaton, Angus. 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. National Bureau of Economic Research: Working Paper #14690, 1–54, Princeton University.

Deer, Brian. 2009. Hidden records show MMR truth. TimesOnline, from the *Sunday Times*, February 8: e1–e7 (Downloaded from www.briandeer.com, 9/2009).

Demicheli, Vittorio, Tom Jefferson, Alessandro Rivetti, and Deirdre Price. 2005. Vaccines for measles, mumps, and rubella in children. *Cochrane database of systematic reviews*, Issue 4, Art. No: CD004407

DeNavas-Walt, Carmen, Bernadette D. Proctor, and Jessica C. Smith. 2008. *U.S. Census Bureau, Current Population Reports, P60-235, Income, Poverty, and Health Insurance Coverage in the United States: 2007*. Washington, DC: U.S. Government Printing Office.

DerSimonian, Rebecca, and Nan Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7: 177–188.

DeStefano, Frank, and Robert T. Chen. 1999. Negative association between MMR and autism. *The Lancet* 353(12 June): 1987–1988.

DeStefano, Frank, T.K. Bhasin, W.W. Thompson, M. Yeargin-Allsopp, and C. Boyle. 2004. Age at first measlemumpsrubella vaccination in children with autism and school-matched control subjects: A population-based study in metropolitan Atlanta. *Pediatrics* 113(2): 259–266.

DuMouchel, William and Christine Waternaux. 1992. Combining information workshop. Boston Chapter American Statistical Association, November 14.

Duncan, Otis Dudley. 1975. *Introduction to structural equation modeling*. New York: Academic Press.

Editorial. 2002. Time to look beyond MMR in autism research. *The Lancet* 359(9307, 23 (February)): 637.

Editors. 2010. *The Lancet* (February 2): Published Online February 2, 2010. DOI:10.1016/ S0140-6736(10)60175-4.

Engelberg Center for Health Care Reform at Brookings. 2009. Bending the curve: Effective steps to address long-term health care spending growth. Washington, DC: Brookings Institution, August, 1–8. Downloaded from www.brookings.edu/healthreform, September 7, 2009.

Erlanger, Steven. 2008. Belgium charges 6 in terror investigation. *New York Times*, December 13 (Basil Katz contributed reporting).

Etzioni, Amitai. 1968. *The active society: A theory of societal and political process*. New York: Free Press.

Etzioni, Amitai. 1983. On policy research. In *A handbook of social science methods*, *An introduction to social research*, vol. 1, ed. Robert B. Smith, 77–92. Cambridge, MA: Ballinger Publishing Company.

Etzioni, Amitai. 1988. *The moral dimension: Toward a new economics*. New York: Free Press.

Etzioni, Amitai. 2001. *The monochrome society*. Princeton, NJ: Princeton University Press.

European Union Monitoring Centre on Racism and Xenophobia. 2004a. *Manifestations of Antisemitism in the EU 2002–2003*, Executive Summary. March.

European Union Monitoring Centre on Racism and Xenophobia. 2004b. *Manifestations of Antisemitism in the EU 2002–2003*. March.

European Union Monitoring Centre on Racism and Xenophobia. 2004c. *Perceptions of Antisemitism in the European Union: Voices from members of the European Jewish communities*. March.

European Union Monitoring Centre on Racism and Xenophobia. 2004d. *Manifestations of Antisemitism in the EU 2002–2003*. EUMC Press Summary. March.

European Union. 2004. Website reports on total population. http://europa.eu.int.

Farrington, C. Paddy, Elizabeth Miller, and Brent Taylor. 2001. MMR and autism: Further evidence against a causal association. *Vaccine* 19: 3632–3635.

Fein, Helen. 1979. *Accounting for genocide*. New York: The Free Press.

Feldstein, Paul J., Thomas M. Wickizer, and John R.C. Wheeler. 1988. Private cost containment: The effects of utilization review programs on health care use and expenditures. *New England Journal of Medicine* 318: 1310–1315.

Festinger, Leon. 1950. Informal social communication. *Psychological Review* 57: 271–282.

Festinger, Leon. 1957. *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson and Company.

Festinger, Leon, Stanley Schachter, and Kurt Back. 1950. *Social pressures in informal groups*. New York: Harper & Brothers.

Finn, J.D., and C.M. Achilles. 1999. Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis* 21: 97–109.

Firebaugh, G. 2003. *The new geography of global income inequality*. Cambridge, MA: Harvard University Press.

Firth, David. 1993. Bias reduction of maximum likelihood estimates. *Biometrika* 80: 27–38.

Fleiss, Joseph L. 1981. *Statistical methods for rates and proportions*. New York: Wiley.

Fleiss, Joseph L., and Alan J. Gross. 1991. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique. *Journal of Clinical Epidemiology* 44: 127–139.

Fogelson, Robert M., and Robert B. Hill. 1968. Who riots? A Study of participation in the 1960s riots. In *Supplemental studies for the National Advisory Commission on Civil Disorders*, 217–248. Washington, DC: U.S. Government Printing Office.

Fox, Peter D. 1997. Applying managed care techniques in traditional Medicare. *Health Affairs* 16(5): 44–57.

Fraser, Steven (ed.). 1995. *The bell curve wars: Race, intelligence, and the future of America*. New York: Basic Books.

Freedman, David A. 2005. *Statistical models*: *Theory and practice*. New York: Cambridge University Press.

Freedom House. 2000. *Freedom in the world 1999–2000*. New York: Freedom House.

Fuchs, Victor R. 2000. Medicare reform: The larger picture. *The Journal of Economic Perspectives* 14(2, spring): 57–70.

Fukuda-Parr, Sakiko. 2003. New threats to human security in the era of globalization. *Journal of Human Development* 4(2, July): 167–179.

Galtung, Johan. 1992. The emerging conflict formations. In *Restructuring for world peace*, ed. Katherine Tehranian and Majid Tehranian, 23–44. Cresskill, NJ: Hampton Press.

Gardner, Howard. 1983. *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Geier, D. A., and M. R. Geier. 2003a. An assessment of the impact of thimerosal on childhood neurodevelopmental disorders. *Pediatric Rehabilitation* 6(2): 97–102.

Geier, M. R., and D. A. Geier. 2003b. Neurodevelopmental disorders after thimerosal-containing vaccines: a brief communication. *Experimental Biology and Medicine* 228(6): 660–664.

Geier, M. R., and D. A. Geier. 2003c. Pediatric MMR vaccination safety. *International Pediatrics* 18(2): 203–208.

Geier, M. R., and D. A. Geier. 2003d. Thimerosal in childhood vaccines, neuro-developmental disorders, and heart disease in the United States. *Journal of American Physicians and Surgeons* 8(1): 6–11.

Geier, D. A., and M. R. Geier. 2004. A comparative evaluation of the effects of MMR immunization and mercury doses from thimerosal-containing childhood vaccines on the population prevalence of autism. *Medical Science Monitor* 10(3): 33–39.

Gelman, Andrew, and Xiao-Li Meng. 2004. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*. West Sussex, England: Wiley.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, Andrew, Daniel Lee, and Yair Ghitza. 2010. Public opinion on health care reform. *The Forum* 8, issue 1, article 8.

George, Alexander L., and Andrew Bennett. 2004. *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Glass, Gene, and Mary Smith. 1978. *Meta-analysis of research on the relationship of class size and achievement*. San Francisco, CA: Far West Laboratory for Educational Research and Development.

Glass, Gene, and Mary Smith. 1979. *Relationship of class size to classroom processes, teacher satisfaction, and pupil affect: A meta-analysis*. San Francisco, CA: Far West Laboratory for Educational Research and Development. July.

Glock, Charles Y. 1967. Survey design and analysis in sociology. In *Survey research in the social sciences*, ed. Charles Y. Glock, 1–62. New York: Russell Sage Foundation.

Glock, Charles, and Rodney Stark. 1966. *Christian beliefs and antisemitism*. New York: Harper & Row Press.

Goldberg, Bruce, and John Richards. 1996. The Co-NECT design for school change. In *Bold plans for school restructuring: The New American Schools designs*, ed. Samuel C. Stringfield, Steven M. Ross, and Lana Smith, 75–108. Mahwah, NJ: Lawrence Erlbaum.

Goldberger, Arthur. 1964. *Econometric theory*. New York: Wiley.

Goldhagen, Daniel J. 1996. *Hitler's willing executioners*. New York: Knopf.

Goldstein, Harvey. 1987. *Multilevel models in educational and social research*. New York: Oxford University Press.

Goldstein, Harvey. 1995. *Multilevel models*, 2nd ed. New York: Oxford University Press.

Goldstein, Harvey. 2003. *Multilevel models*, 3rd ed. New York: Oxford University Press.

Goldthorpe, John H. 1998. *Causation, statistics and sociology*. Dublin, Ireland: The Economic and Social Research Institute.

Goldthorpe, John H. 2007. Causation, statistics and sociology. In *On sociology*, 2nd edition, 190–216. Stanford, CA: Stanford University Press.

Goodman, Leo. 1972a. A modified multiple regression approach to the analysis of dichotomous variables. *American Sociological Review* 37: 28–46.

Goodman, Leo. 1972b. A general model for the analysis of surveys. *American Journal of Sociology* 77: 1035–1086.

Goodman, Leo. 1973. Causal analysis of data from panel studies and other kinds of data. *American Journal of Sociology* 78: 1135–1191.

Goodman, Leo. 1978. *Analyzing qualitative/categorical data (ed. Jay Magidson)*. Cambridge, MA: Abt Books.

Goodman, Leo. 1984. *The analysis of cross-classified data having ordered categories*. Cambridge, MA: Harvard University.

Goodwin, Doris Kerns. 2005. *Team of rivals: The political genius of Abraham Lincoln*. New York: Simon & Schuster.

Granger, Clive. 1986. Comment. *Journal of the American Statistical Association* 81: 967–968.

Greenland, Sander. 2004. An overview of methods for causal inference from observational studies. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, ed. Andrew Gelman and Xiao-Li Ming, 3–13. West Sussex, England: Wiley.

Greenland, Sander, Judea Pearl, and James M. Robins. 1999a. Causal diagrams for empidemiologic research. *Epidemiology* 10: 37–48.

Greenland, Sander, James M. Robins, and Judea Pearl. 1999b. Confounding and collapsibility in causal inference. *Statistical Science* 14: 29–46.

Gross, Neil, and Solon Simmons. 2007. *The social and political views of American professors*. Working paper. Cambridge, MA: Sociology Department, Harvard University, September.

Hage, Jerald. 1974. *Communication and organizational control: Cybernetics in health and welfare settings*. New York: Wiley.

Hage, Jerald T. 1999. Organizational innovation and organizational change. *Annual Review of Sociology* 25: 597–622.

Haq, Mahbub ul. 1995. *Reflections on human development*. New York, NY: Oxford University Press.

Harding, David J. 2003. Counterfactual models of neighborhood effects: The effect of poverty on dropping out and teenage pregnancy. *American Journal of Sociology* 109: 676–719.

Harkins, Nancy. 2001. *Co-nect schools: Informational report to the Superintendent*. Harford County, Maryland: Public Schools Office. July.

Hasselblad, Vic, Frederick Mosteller, Benjamin Littenberg, Thomas C. Chalmers, Maria G.M. Hunink, Judith A. Turner, Salley C. Morton, Paula Diehr, John B. Wong, and Neil R. Powe. 1995. A survey of current problems in meta-analysis. *Medical Care* 33: 202–220.

Heckman, James J. 2005a. The scientific model of causality. *Sociology Methodology* 35: 1–97.

Heckman, James J. 2005b. Rejoinder: Response to Sobel. *Sociology Methodology* 35: 135–162.

Heckman, James J., and Sergio Urzua. 2009. *Comparing IV with structural models: What simple IV can and cannot identify, NBER working papers 14706*. Cambridge, MA: National Bureau of Economic Research.

Heller, Patrick. 1999. *The labor of development: Workers and the transformation of capitalism in Kerala, India*. Ithaca: Cornell University Press.

Hellevik, Ottar. 2009. Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity* 43: 59–74.

Herrnstein, Richard J., and Charles Murray. 1994. *Bell curve: Intelligence and class structure in American Life*. New York: Free Press.

Higgins, Julian P.T., and Sally Green. (eds.). 2008. Cochrane handbook for systematic reviews of interventions, version 5.0.1 (updated September 2008). *The Cochrane Collaboration*: Available from www.cochrane-handbook.org.

Hill, Austin Bradford. 1965. The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine* 58: 295–300.

Hirsch Jr., E. D. 2010. How to save the schools. *The New York Review of Books* 67(8, May 13): 16–19.

Hoff, Karla, and Joseph E. Stiglitz. 2001. Modern economic theory and development. In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 389–459. New York: Oxford University Press.

Holland, Paul. 1986. Statistics and causal inference (with comments). *Journal of the American Statistical Association* 81: 945–970.

Homans, George. 1975. What do we mean by social "structure.". In *Approaches to the study of social structure*, ed. Peter M. Blau, 53–65. New York: The Free Press.

Hoxby, Caroline M. 2000. The effect of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics* 115(4): 1239–1285.

Huband, Mark. 2004. Europe ponders roots of radicalisation. *Financial Times*, November 17, 15.

Huntington, Samuel P. 1996. *The clash of civilizations and the remaking of world order*. New York: Simon and Schuster.

Huntington, Samuel P. 1968. *Political order in changing societies*. New Haven, CT: Yale University Press.

Hviid, A., M. Stellfeld, J. Wohlfahrt, and M. Melbye. 2003. Association between thimerosal-containing vaccine and autism. *Journal of the American Medical Association* 290(13): 1763–6.

Hyman, Herbert H. 1955. *Survey design and analysis*. New York: The Free Press.

Imbens, Guido I. 2009. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). Department of Economics, Harvard University, 1–30.

Imbens, Guido I., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47(1): 5–86.

Inglehart, Ronald. 2003. *Human values and social change: Findings from the values surveys*. Boston, MA: Brill.

Institute of Medicine. 1994. *Adverse events associated with childhood vaccines: Evidence bearing on causality*. Washington, DC: National Academy Press.

Institute of Medicine. 2001a. *Immunization safety review: Measles-Mumps-Rubella and Autism*. Washington, DC: National Academy Press.

Institute of Medicine. 2001b. *Immunization safety review: Thimerosal-containing vaccines and neurodevelopmental disorders*. Washington, DC: National Academy Press.

Institute of Medicine. 2004. *Vaccines and autism*. Washington, DC: National Academy Press.

Jacoby, Russel, and Naomi Glauberman (eds.). 1995. *The bell curve debate: History, documents, opinions*. New York: Times Books.

Jepson, Christopher, and Steven Rivkin. 2009. Class size reduction and student achievement: A potential tradeoff between teacher quality and class size. *Journal of Human Resources* 44(1): 223–250.

Jöreskog, Karl G., and Dag Sörbom. 1979. *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Jöreskog, Karl G., and Dag Sörbom. 1993. *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Julius, Anthony. 2010. *Trials of the diaspora: A history of anti-Semitism in England*. New York: Oxford University Press.

Kaiser Family Foundation. 2009. Medicare advantage fact sheet. Menlo Park, CA: April, 1–2.

Kane, Robert L., and Bruce Friedman. 1997. State variations in Medicare expenditures. *American Journal of Public Health* 87(10): 1611–1619.

Kaplan, Edward H., and Charles A. Small. 2005. Anti-Israel sentiment predicts Anti-Semitism in Europe. *The Journal of Conflict Resolution* 50: 548–561.

Karney, B.R., and T.N. Bradbury. 1997. Neuroticism, marital interaction, and the trajectory of marital satisfaction. *Journal of Personality and Social Psychology* 72: 1075–1092.

Kendall, Patricia L., and Paul F. Lazarsfeld. 1950. Problems of survey analysis. In *Continuities in social research*, ed. Robert K. Merton and Paul F. Lazarsfeld, 133–196. Glencoe: The Free Press.

Kenny, David A. 1979. *Correlation and causation*. New York: Wiley.

Kentor, Jeffrey, and Terry Boswell. 2003. Foreign capital dependence and development: A new direction. *American Sociological Review* 68: 301–313.

Khandker, Rezaul K., and Willard G. Manning. 1992. The impact of utilization review on costs and utilization. In *Health economics worldwide*, ed. P. Zeifel and H.E. Frech III, 47–62. The Netherlands: Kluwer.

King, Gary. 1997. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton, NJ: Princeton University Press.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

King, Gary, Emmanuela Gakidou, Kosuke Imai, Jason Lakin, Ryan T. Moore, Clayton Nall, Nirmala Ravishankar, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. 2009. Public policy for the poor? A randomised assessment of the Mexican universal health insurance program. *The Lancet* 373(April 25): 1447–1454.

Kline, Rex B. 2005. *Principles and practice of structural equation modeling*, 2nd ed. New York: Guilford Press.

Kohn, Melvin L. 1969. *Class and conformity, a study of values*. Homewood, IL: Dorsey.

Kohn, Melvin L. 1977. *Class and conformity, with a reassessment*, 2nd ed. Chicago, IL: University of Chicago Press.

Kramer, Michael S., and David A. Lane. 1992. Causal propositions in clinical research and practice. *Journal of Clinical Epidemiology* 45(6): 639–649.

Kreft, Ita, and Jan DeLeeuw. 1998. *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Krishnan, T.R. 1997. The route to social development in Kerala: Social intermediation and public policy. In *Development with a human face: Experiences in social achievement and economic growth*, ed. Santosh Mehrotra and Richard Jolly, 204–234. New York: Oxford University Press.

Kristoff, Nicholas D. 2005. The American witness. *New York Times*, March 2, 19.

Kunda, Ziva. 1990. The case for motivated reasoning. *Psychological Bulletin* 108: 480–498.

Ladd Jr., Everett Carll, and Seymour Martin Lipset. 1976. *The divided academy*. New York: W.W. Norton.

Land, Kenneth. 1969. Principles of path analysis. In *Sociology methodology 1969*, ed. Edgar F. Borgatta, 3–37. San Francisco: Jossey-Bass.

Landes, David S. 1998. *The wealth and poverty of nations: Why some are so rich and some so poor*. New York: W.W. Norton.

Langwell, Kathryn. 1992. *The effects of managed care on use and costs of health services. CBO staff memorandum*. Washington, DC: Congressional Budget Office. June.

Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. The effects of negative political advertising: A meta-analytic assessment. *American Political Science Review* 93: 851–875.

Lauritzen, Steffen L. 2001. Causal inference from graphical models. In *Complex stochastic systems: Monographs on statistics and applied probability 87*, ed. Ole E. Barndorff-Nielsen, David R. Cox, and Claudia Klüppelberg, 63–107. Boca Raton, FL: Chapman & Hall, CRC.

Latané, Bibb, and John M. Darley. 1970. *The unresponsive bystander: Why doesn't he help?* New York: Appleton-Century-Crofts.

Lau, Richard R., Lee Sigelman, Caroline Heldman, and Paul Babbitt. 1999. The effects of negative political advertising: A meta-analytic assessment. *American Political Science Review* 93: 851–875.

Lazarsfeld, Paul F. 1955a. Interpretation of statistical relations as a research operation. In *The language of social research*, ed. Paul F. Lazarsfeld and Morris Rosenberg, 115–125. New York: The Free Press.

Lazarsfeld, Paul F. 1955b. Foreword. In Herbert H. Hyman, *Survey design and analysis*. New York: Free Press. v–xi.

Lazarsfeld, Paul F. 1959. Problems in methodology. In *Sociology today: problems and prospects*, ed. Robert K. Merton, Leonard Broom, and Leonard S. Cottrell Jr., 39–78. New York: Basic Books.

Lazarsfeld, Paul F. [1946] 1972. Mutual effects of statistical variables. In *Continuities in the language of social research*, ed. Paul F. Lazarsfeld, Ann K. Pasanella, and Morris Rosenberg, 388–398. New York: Free Press.

Lazarsfeld, Paul F., Bernard Berelson, and Hazel Gaudet. [1944] 1948. *The people's choice*. New York: Columbia University Press.

Lazarsfeld, Paul F., and Wagner Thielens Jr. 1958. *The academic mind*. New York: Free Press.

Lazarsfeld, Paul F., and Herbert Menzel. [1961] 1972. On the relation between individual and collective properties. In *Continuities in the language of social research*, ed. Paul F. Lazarsfeld, Ann K. Pasanella, and Morris Rosenberg, 219–237. New York: The Free Press.

Lazarsfeld, Paul F., Ann K. Pasanella, and Morris Rosenberg (eds.). 1972. *Continuities in the language of social research*. New York, NY: Free Press.

Lewin, Kurt. [1942] 1999. Personal adjustment and group belongingness. In *The complete social scientist: A Kurt Lewin reader*, ed. Martin Gold, 327–332. Washington, DC: American Psychological Association.

Lewin, Kurt. [1948] 1997. Self-hatred among Jews. In *Resolving social conflicts: Selected papers on group dynamics*, ed. Gertrud Weiss Lewin, 133–142. Washington, DC: American Psychological Association.

Lewin, Kurt. [1951] 1997. *Field theory in social science*. Washington, DC: American Psychological Association.

Lewis, Kristen, and Sarah Burd-Sharps. 2010. *A century apart*. Brooklyn, NY: American Human Development Project.

Lieberson, Stanley. 1997. The big broad issues in society and social history: Application of a probabilistic perspective. In *Causality in crisis: Statistical methods and the search for causal* *knowledge in the social sciences*, ed. Vaughn R. McKim and Stephen P. Turner, 359–385. South Bend, IN: University of Notre Dame Press.

Liker, Jeffrey K., and Ahmed A. Sindi. 1997. User acceptance of expert systems: A test of the theory of reasoned action. *Journal of Engineering and Technological Management* 14: 147–173.

Liker, Jeffrey K., Carol J. Haddad, and Jennifer Karlin. 1999. Perspectives on technology and work organization. *Annual Review of Sociology* 25: 575–596.

Lindemann, Albert S. 2000. *Anti-Semitism before the Holocaust*. Essex, UK: Pearson.

Linz, Juan J. 2000. *Totalitarian and authoritarian regimes*. Boulder, CO: Lynne Rienner.

Lippmann, Walter. 1922. *Public opinion*. New York: Harcourt Brace.

Lipset, Seymour Martin. [1959] 1981. Economic development and democracy. In *Political man: the social bases of politics*. Baltimore, MD: Johns Hopkins University Press.

Lipset, Seymour Martin, and Gabriel Salman Lenz. 2000. Corruption, culture, and markets. In *Culture matters: How values shape human progress*, ed. Lawrence E. Harrison and Samuel P. Huntington. New York: Basic Books.

Lipsey, Mark W., and David B. Wilson. 2001. *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Littell, Ramon C., George A. Milliken, Walter W. Stroup, and Russell D. Wolfinger. 1996. *SAS for mixed models*. Cary, NC: SAS Institute.

Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. 2006. *SAS for mixed models*, 2nd ed. Cary, NC: SAS Institute.

Lucas, William A., Sarah Y. Cooper, and Elena M. Rodriquez-Falcon. 2006. On the recognition of venturing opportunities in science and technology. Cambridge-MIT Institute, March.

Maas, C.J.M., and T.A.B. Snijders. 2003. The multi-level approach to repeated measures for complete and incomplete data. *Quality & Quantity* 37: 71–89.

Madsen, K.M., A. Hviid, M. Vestergaard, D. Schendel, J. Wohlfahrt, P. Thorsen, J. Olsen, and M. Melbye. 2002. A population-based study of measles, mumps, and rubella vaccination and autism. *New England Journal of Medicine* 347(19): 1477–1482.

Magidson, Jay. 1978. An illustrative comparison of Goodman's approach to logit analysis with dummy variable regression analysis. In *Analyzing Qualitative/Categorical Data*, ed. Jay Magidson, 27–54. Cambridge, MA: Abt Books.

Manning, Willard G. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* 17: 283–295.

Manning, Willard G., and John Mullahy. 1999. Estimating log models: To transform or not to transform? Technical working paper 246. Cambridge, MA: National Bureau of Economic Research. November.

Marcus, Steven. 1974. *Engels, Manchester, and the working class*. New York: Norton.

Marmor, Theodore, and Jonathan Oberlander. 1998. Rethinking Medicare reform. *Health Affairs* 17(Jan/Feb): 52–68.

Marsden, Peter V. 1985. D-systems and effect parameters: Some complementarities. In *A handbook of social science methods, quantitative methods: Focused survery research and causal modeling,* vol. 3, ed. Robert B. Smith, 367–389. New York: Praeger.

Mason, William M., George Y. Wong, and Barbara Entwisle. 1983. Contextual analysis through the multilevel linear model. *Sociology Methodology 1983–1984*. San Francisco: Jossey-Bass Publishers.

Mays, Glen P., Gary Claxton, and Justin White. 2004. Market watch: Managed care rebound? Recent changes in health plan's cost containment strategies. Health Affairs-Web Exclusive, August 11: W4-427-W436, downloaded October 24, 2009.

Mazur, Allan. 1972. The causes of black riots. *American Sociological Review* 37: 490–93.

Mazur, Allan. 2008. *Implausible beliefs: The Bible, astrology, and UFOs*. New Brunswick, NJ: Transaction Publishers.

McCartney, Margaret. 2009. Averting outbreaks. *Health, Financial Times,* issue three, September 18: 10.

McClellan, Mark. 2000. Medicare reform: fundamental problems, incremental steps. *The Journal of Economic Perspectives* 14(2, Spring): 21–44.

McCullagh, P., and J.A. Nelder. 1989. *Generalized linear models*, 2nd ed. London, UK: Chapman & Hall.

Media Tenor. 2006. *The portrayal of the war in the Middle East*. Bonn, DE, August 18.

Media Tenor. 2007. Preview of the final annual dialogue report. Bonn, DE, September 29.

Mehrotra, Santosh. 1997a. Health and education policies in high-achieving countries: Some lessons. In *Development with a human face: Experiences in social achievement and economic growth*, ed. Santosh Mehrotra and Richard Jolly, 63–110. New York: Oxford University Press.

Mehrotra, Santosh. 1997b. Social development in high-achieving countries: Common elements and diversities. In *Development with a human face: Experiences in social achievement and economic growth*, ed. Santosh Mehrotra and Richard Jolly, 21–61. New York: Oxford University Press.

Meier, Gerald M. 2001. The old generation of development economists and the new. In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 13–50. New York: Oxford University Press.

Mello, John P., Jr. 2000. In user-friendly era, tech-challenged still see PC as foe. *The Boston Sunday Globe*, November 19, K9.

Merton, Robert K. 1957a. Social structure and anomie. Social theory and social structure, second revised and enlarged edition, chapter 4, pp. 131–160. New York: The Free Press.

Merton, Robert K. 1957b. The machine, the worker, and the engineer. Social theory and social structure, second revised and enlarged edition, chapter 17, pp. 562–573. New York: The Free Press.

Merton, Robert K. 1984. Socially expected durations: A case study of concept formation in Sociology. In *Conflicts and consensus*, ed. Walter W. Powell and Richard Robbins. New York: Free Press.

Meyer, John W., John Boli, George M. Thomas, and Francisco O. Ramirez. 1997. World society and the nation-state. *American Journal of Sociology* 103(1, July): 144–181.

Miller E. 2004. Presentation to the Immunization Safety Review Committee. Thimerosal and developmental problems including autism. Washington, DC.

Miller, Robert H., and Harold S. Luft. 1997. Does managed care lead to better or worse quality of care? *Health Affairs* 16 (Sept/Oct): 7–25.

Mills, C. Wright. 1956. *White collar*. New York: Oxford University Press.

Minder, Raphael. 2004. Belgian far right highlights the dilemma for democracy. *Financial Times*, July 21.

Moore, David S., and George P. McCabe. 1989. *Introduction to the practice of statistics*. New York: W.H. Freeman.

Moreno, Rafael, and Rafael J. Martínez. 2008. Causality as validity: Some implications for the social sciences. *Quality & Quantity* 42: 597–604.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: Methods and principles for social research*. New York: Cambridge University Press.

Mosteller, Frederick. 1995. The Tennessee study of class size in the early grades. *Critical Issues for Children and Youth* 5(Summer/Fall): 113–127.

Murch, S.H., A. Anthony, D.H. Casson, M. Malik, M. Berelowitz, A.P. Dhillon, M.A. Thomson, A. Valentine, S.E. Davies, and J.A. Walker-Smith. 2004. Retraction of an interpretation. *The Lancet* 363: 750.

Murphy, Dan, and Bella Rosenberg. 1998. Recent research shows major benefit of small class size. *Educational Issues Policy Brief 31*: Washington DC, American Federation of Teachers, 1–3.

Murray, Charles. 2008. *Real education: Four simple truths for bringing America's schools back to reality*. New York: Crown Forum.

Myrdal, Gunnar. [1944] 1964. *An American dilemma*. New York, NY: McGraw-Hill.

Myrdal, Gunnar. 1971. *Asian drama: An inquiry into the poverty of nations*. New York: Pantheon.

Nagel, Ernest. 1961. *The structure of science: Problems in the logic of scientific explanation*. New York: Harcourt, Brace & World.

Neumayer, Eric. 2005. Commentary: the economic business cycle and mortality. *International Journal of Epidemiology* 34: 1221–1222.

Nunn, Nathan. 2008. Long term effects of Africa's slave trades. *Quarterly Journal of Economics* 123: 139–176.

Nyman, John A., Roger Feldman, Janet Shapiro, Colleen Grogan, and David Link. 1990. Changing physician behavior: Does medical review of part B Medicare claims make a difference? *Inquiry* 27(Summer): 127–137.

Obama, Barack. 2009. Health care speech, September, 9. Downloaded from the *Huffington Post*, September 11, 2009.

Operario, Don, and Susan T. Fiske. 2004. Stereotypes. In *Social cognition*, ed. Marilynn B. Brewer and Miles Hewstone, 120–141. Malden, MA: Blackwell.

Organisation for Economic Co-Operation and Development. 2000. *Health at a glance: OECD indicators*. Paris: OECD Publishing.

Organisation for Economic Co-Operation and Development. 2007. *Health at a glance: OECD indicators*. Paris: OECD Publishing.

Pape, Robert A. 2005. *Dying to win: The strategic logic of suicide terrorism*. New York: Random House.

Parsons, Talcott, Robert F. Bales, and Edward A. Shils. 1953. *Working papers in the theory of action*. Glencoe IL: The Free Press.

Parsons, Talcott. 1966. *Societies: Evolutionary and comparative perspectives*. Englewood Cliffs, NJ: Prentice-Hall.

Patterson, Orlando. 1982. *Slavery and social death: A comparative study*. Cambridge, MA: Harvard University Press.

Pawson, Ray. 1989. *A measure for measures: A manifesto for empirical sociology*. London: Routledge.

Pawson, Ray. 2006. *Evidence-based policy: A realist perspective*. London: Sage.

Pawson, Ray, and Nick Tilley. 1997. *Realistic evaluation*. London: Sage.

Paxton, Pamela. 2002. Social capital and democracy: An interdependent relationship. *American Sociological Review* 67: 254–277.

Peabody, Zanto. 2003. Sharpstown numbers game shows flaw in school reform. Houston Chronicle, July 5.

Pearl, Judea. [2000] 2009. *Causality: Models, reasoning, and inference*, 2nd ed. New York: Cambridge University Press.

Pearl, Judea. 2002. Comments on seeing and doing. *International Statistical Review* 70: 207–209.

Pearl, Judea. 2009. Causality in the social and behavioral sciences. University of California, Los Angeles, Computer Science Department: Technical Report R-355.

Porat, Dina. 2004. From Europe to the Muslim world and back again. *Haaretz*, April 5.

Public Policy Polling. 2010. *Country divided on Obama, Congress*. Raleigh, NC: Public Policy Polling.

Prezeworski, Adam, Michael E. Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and development: Political institutions and well-being in the world, 1950–1990*. New York: Cambridge University Press.

Project GRAD: Working to Close the Academic Achievement Gap. Undated, circa 2002.

Ramsey, Sarah. 2001. Controversial MMR-autism investigator resigns from research post. *The Lancet* 358 (issue 9297, 8 December): 1972.

Rapoport, Anatol. 1950. *Science and the goals of man*. New York: Harper & Brothers.

Rapoport, Anatol. 1953. *Operational philosophy*. New York: Harper & Brothers.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*, 2nd ed. Newbury Park, CA: Sage.

Raudenbush, Stephen W., Anthony S. Bryk, Yuk Fai Cheong, and Richard T. Congdon Jr. 2000. *HLM 5 hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Ravitch, Diane. 2010. *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Reinhardt, Uwe E. 2000. Health care for the aging baby boom: Lessons from abroad. *The Journal of Economic Perspectives* 14(2, Spring): 71–84.

Restuccia, J.D. 1983. The effect of concurrent feedback in reducing inappropriate hospital utilization. *Medical Care* 20: 46–62.

Reynolds, H.T. 1974. Ordinal partial correlation and causal inferences. In *Measurement in the social sciences*, ed. H.M. Blalock Jr., 399–423. Chicago: Aldine.

Robins, James M., and Larry Wasserman. 1999. On the impossibility of inferring causation from association without background knowledge. In *Computation, causation, & discovery*, ed. Glymour Clark and Gregory F. Cooper, 303–321. Cambridge, MA: MIT Press.

Rosenbaum, Paul R., and Donald B. Rubin. [1983] 2006. The central role of the propensity score in observational studies for causal effects. In *Matched sampling for causal effects*, ed. Donald B. Rubin, 170–184. New York: Cambridge University Press.

Rosenbaum, Paul R., and Donald B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39: 33–38.

Rosenthal, Robert. 1991. *Meta-analytic procedures for social research*, Applied Social Research Methods Series, vol. 6, revised ed. Newbury Park, CA: Sage.

Rosenthal, Robert, and Donald B. Rubin. 1994. The counternull value of an effect size: A new statistic. *Psychological Science* 5(6 November): 329–334.

Rosenthal, Robert, Ralph L. Rosnow, and Donald B. Rubin. 2000. *Contrasts and effect sizes in behavioral research*. Cambridge, UK: Cambridge University Press.

Ross, Dennis. 2004. *The missing peace: The inside story of the fight for Middle East peace*. New York: Farrar, Straus & Giroux.

Ross, H. Laurence. 1970. *Settled out of court*. Chicago: Aldine.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology* 66: 688–701.

Rubin, Donald B. 1986. Comment: Which ifs have causal answers. *Journal of the American Statistical Association* 81: 961–962.

Rubin, Donald B. 2003. *Basic concepts of statistical inference for causal effects in experiments and observational studies*. Cambridge, MA: Harvard University, Department of Statistics.

Rubin, Donald B. [1984] 2006. William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In *Matched sampling for causal effects*, ed. Donald B. Rubin, 7–29. New York: Cambridge University Press.

Rubin, Donald B. 2006. *Matched sampling for causal effects*. New York: Cambridge University Press.

Russell, Michael, and Ryan Robinson. 2000. *Co-nect retrospective outcomes study. Center for the study of testing, evaluation and educational policy*. Chestnut Hill, MA: Boston College. April.

Sachs, Jeffery D. 2005. *Investing in development: A practical plan to achieve the Millennium Development Goals (overview)*. New York: United Nations Development Program.

Samarah, Adel. 2001. *Epidemic of globalization: Ventures in world order, Arab nation and Zionism*. Glendale, CA: Palestine Research & Publishing Foundation.

Sampson, Robert, Stephen Raudenbush, and T. Earls. 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277: 914–918.

SAS Institute Inc. 1990. Chapter 24, the GLM procedure. In SAS/STAT® User's Guide Volume 2, GLM-VARCOMP: Version 6, fourth edition, 891–996. Cary, NC: SAS Institute.

SAS Institute. 1997a. *SAS/STAT® Software: changes and enhancements through release*, vol. 6.12. Cary, NC: SAS Institute.

SAS Institute Inc. 1997b. Chapter 10, the GENMOD procedure. In *SAS/STAT® software: Changes and enhancements through release*, vol. 6.12, 245–348. Cary, NC: SAS Institute.

SAS Institute. 1997c. Chapter 20, the MULTTEST Procedure. In *SAS/STAT ® Software: changes and enhancements through release*, vol. 6.12, 777–821. Cary, NC: SAS Institute.

SAS Institute. 2005. *The Glimmix procedure, November*. Cary, NC: SAS Institute.

Saving, Thomas R. 2000. Making the transition to prepaid Medicare. *The Journal of Economic Perspectives* 14(2, Spring): 85–98.

Schemo, Diana J. 2003. Questions on data cloud luster of Houston schools. *New York Times*, July 11.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Selvin, Hanan. 1960. *The effects of leadership*. New York: Free Press.

Seffrin, John R. 2009. Letter to the Editor: Re "In Long Drive to Cure Cancer, Advances Have Been Elusive" ("Forty Years' War" series, front page, April 24). *New York Times*, published April 27.

Semmelweis, Ignaz. [1861] 1983. In *The etiology, concept, and prophylaxis of childbed fever*, ed. K. Codell Carter. Madison, WI: University of Wisconsin Press.

Sen, Amartya. 1999. *Development as freedom*. New York: Alfred A. Knopf.

Sen, Amartya. 2001. What is development about? In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 506–513. New York: Oxford University Press.

Sen, Amartya. 2002. Cultural Imprisonment. *The New Republic*, 10 June: 28–33.

Seward, William H. 1877. *An autobiography of William H. Seward from 1801 to 1834*. New York: D. Appleton.

Sewell, William H., and J. Michael Armer. 1966. Neighborhood context and college plans. *American Sociological Review* 31: 159–168.

Shastri, Lal Badadur. 1990. *Incidence of bonded labour in India: Area, nature and extent*. Mussoorie: National Academy of Administration.

Shils, Edward A., and Morris Janowitz. 1948. Cohesions and disintegration in the Wehrmacht in World War II. *Public Opinion Quarterly* 12: 280–315.

Simon Wiesenthal Center. 2004. *Response*. Los Angeles, CA.

Simon, Herbert A. [1953] 1957. Causal ordering and identifiability. In *Models of man*, 10–36. New York: Wiley.

Simon, Herbert A. [1954] 1957. Spurious correlation: A causal interpretation. In *Models of man*, 37–49. New York: Wiley.

Simon, Herbert A. 1957. *Models of man*. New York: Wiley.

Singer, Judith D. 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics* 24(4): 323–355.

Singer, Judith D., and John B. Willett. 2003. *Applied longitudinal data analysis*. New York: Oxford University Press.

Siska, Andrea, C. Truffer, S. Smith, S. Keehan, J. Cylus, J.A. Poisal, M.K. Clemens, and J. Lizonitz. 2009. Health spending projections through 2018: Recession effects add uncertainty to the outlook. *Health Affairs* 28(2): w346–w357.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized latent variable models: Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall.

Slezkine, Yuri. 2004. *The Jewish century*. Princeton, NJ: Princeton University Press.

Smeeth, Liam, Claire Cook, Eric Fombone, Lisa Heavy, Laura C. Rodrigues, Peter G. Smith, and Andrew J. Hall. 2004. MMR vaccination and pervasive development disorders: A case-control study. *The Lancet* 364 (issue 9438, 11 September): 963–969.

Smelser, Neil J. 1998. The rational and the ambivalent in the social sciences. *American Sociological Review* 63(February): 1–15.

Smith, Craig. 2004. Thwarted in Germany, Neo-Nazis take Fascism to France. *New York Times* on the Web, August 13.

Smith, David G. 1992. *Paying for Medicare: The politics of reform*. New York: Aldine De Guyter.

Smith, Robert B. 1972. Neighborhood context and college plans: An ordinal path analysis. *Social Forces* 51: 200–229.

Smith, Robert B. 1974. Continuities in ordinal path analysis. *Social Forces* 53: 200–229.

Smith, Robert B. 1978. Nonparametric path analysis: Comments on Kim's "Multivariate Analysis of Ordinal Variables". *American Journal of Sociology* 84: 437–448.

Smith, Robert B. 1985a. Focused survey research: Aspects of the perspective and method. In *Handbook of social science methods, Quantitative methods: Focused survey research and causal modeling*, vol. 3, ed. Robert B. Smith, 59–96. New York: Prager.

Smith, Robert B. 1985b. Spearman's rho-b, $\bar{r}$ and the path analysis of contingency tables. *Quality & Quantity* 20: 53–74.

Smith, Robert B. 1986. The performance of rho-b statistics. *Quality & Quantity* 20: 53–74.

Smith, Robert B. 1993. Health care reform now. *Society* 30(March/April): 56–65.

Smith, Robert B. 1998. Anti-Semitism and Nazism. *American Behavioral Scientist* 41: 1324–1362.

Smith, Robert B. 1999a (Unpublished). Don't hassle me! Why white collar workers opposed computerized fraud detection: Using Coleman effect parameters to draw causal inferences. Cambridge, MA: Social Structural Research Inc.

Smith, Robert B. 1999b. Untangling political ideology and party identification in the United States. *Quality & Quantity* 33: 27–44.

Smith, Robert B. 2000. *A computer simulation model of opinion change in Nazi Germany*. Cambridge: Social Structural Research.

Smith, Robert B. 2001. Gatekeepers and Sentinels: Their consolidated effects on inpatient care. *Evaluation Review* 25: 288–330.

Smith, Robert B. 2002. Will claims workers dislike a computerized fraud detector? *Evaluation Review* 26: 3–39.

Smith, Robert B. 2003a. Inferential causal models: Integrating quality and quantity. *Quality & Quantity* 37: 337–361.

Smith, Robert B. 2003b. Political extremism: Left, Center and Right. *The American Sociologist* 34:70–80. Reprinted, 2004. In *Civil society and class politics,* ed. Irving Louis Horowitz, 107–121. New Brunswick, NJ: Transaction, 2004.

Smith, Robert B. 2003b. *Effects of Co-nect's comprehensive school reform in Houston, Texas*. Cambridge, MA: Social Structural Research. September 19.

Smith, Robert B. 2004. *Knowledge about the Holocaust, Holocaust remembrance, and antisemitism*. Cambridge, MA: Social Structural Research. April 17.

Smith, Robert B. 2006. Introduction. In James S. Coleman, *The mathematics of collective action*. New Brunswick, NJ: AldineTransaction. vii–li.

Smith, Robert B. 2008a. A globalized conflict: Anti-Jewish violence during the second intifada. *Quality & Quantity* 42: 135–180.

Smith, Robert B. 2008b. *Cumulative social inquiry: Transforming novelty into innovation*. New York: Guilford Press.

Smith, Robert B. 2009a. Global human development: Accounting for its regional disparities. *Quality & Quantity* 43: 1–34.

Smith, Robert B. 2009b. Issues matter: A case study of factors influencing voting choices. *Case Studies in Business, Industry, and Government Statistics* 2: 127–146.

Smith, Robert B. 2010a. The Academic Mind revisited: Contextual analysis via multilevel modeling. In *Applications of Mathematics in Models, Artificial Neural Networks and Arts: Mathematics and Society*, ed. Vittorio Capecchi, Massimo Buscema, Pierluigi Contucci, and Bruno D'Amore, 163-193. Dordrecht, NL: Springer.

Smith, Robert B. 2010b. Why Nazified Germans killed Jewish people: Insights from agent-based modeling of genocidal actions. In *Current Perspectives in Social Theory*, vol. 27, ed. H.Dahms and L.Hazelrigg, 275–342. Bingly, UK: Emerald.

Smith, Robert B., and Thomas D. Gotowka. 1991. Onsite concurrent review: Impacts on utilization, medical complications and expense. *Benefits Quarterly* 7: 81–90.

Smith, Robert B., and William A. Lucas. 2005. On consideration of use. Cambridge-MIT Institute, July 15.

Smith, Robert B., and William A. Lucas. 2005. Determinants of task self-efficacy: Technical, scientific, and venturing. Cambridge-MIT Institute, August 12.

Smith, Robert B., and William A. Lucas. 2005. Basic research skills drive consideration of use. Cambridge-MIT Institute, November 7.

Smith, Robert B. and William A. Lucas. 2006. On the consideration of use of MIT and University of Cambridge exchange students. Cambridge-MIT Institute, June 19.

Snijders, T.A.B., and R.J. Bosker. 1994. Modeled variance in two-level models. *Sociological Methods and Research* 22: 342–363.

Sobel, Michael E. 1995. Causal inference in the social and behavioral sciences. In *Handbook of statistical modeling for the social and behavioral sciences*, ed. Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel, 1–38. New York and London: Plenum Press.

Sobel, Michael E. 1996. An introduction to causal inference. *Sociological Methods & Research* 24 (3, February): 353–379.

Sobel, Michael E. 2005. Discussion: The scientific model of causality. *Sociological Methodology* 35: 99–133.

Somers, Robert H. 1974. Analysis of partial rank correlation measures based on the product-moment model: Part 1. *Social Forces* (December): 229–246.

Spilerman, Seymour. 1970. The causes of racial disturbances. *American Sociological Review* 35: 627–649.

Spilerman, Seymour. 1971. The causes of racial disturbances: Tests of an explanation. *American Sociological Review* 36: 427–442.

Spilerman, Seymour. 1972. Strategic considerations in analyzing the distribution of racial disturbances. *American Sociological Review* 37: 493–499.

Spilerman, Seymour. 1976. Structural characteristics of cities and the severity of racial disorders. *American Sociological Review* 41: 771–793.

SPSS Incorporated. 2002. *Mixed models white paper, Linear mixed effects modeling in SPSS*. Chicago, IL: 233 S. Wacker Drive.

Stampp, Kenneth. [1959] 1991. *The causes of the civil war*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.

Steele, Claude M. 1997. A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist* 52: 613–629.

Stehr-Green, P., P. Tull, M. Stellfeld, P.B. Mortenson, and D. Simpson. 2003. Autism and thimerosal-containing vaccines: Lack of consistent evidence for an association. *American Journal of Preventive Medicine* 25(2): 101–6.

Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism, Tel Aviv University. 2004. *Antisemitism worldwide 2003/4: General analysis*. Tel Aviv: Tel Aviv University.

Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism, Tel Aviv University. 2007. *Antisemitism worldwide 2006: General analysis*. Tel Aviv: Tel Aviv University.

Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism, Tel Aviv University. 2009. *Antisemitism worldwide 2008/9: General analysis*. Tel Aviv: Tel Aviv University.

Stephen Roth Institute for the Study of Contemporary Antisemitism and Racism, Tel Aviv University. 2010. *Antisemitism worldwide 2009: General analysis*. Tel Aviv: Tel Aviv University.

Stokes, Maura E., Charles S. Davis, and Gary G. Koch. 2000. *Categorical data analysis using the SAS® system*, 2nd ed. Cary, NC: SAS Institute.

Stokes, Donald. 1997. *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution.

Stouffer, Samuel A. 1955. *Communism, conformity, and civil liberties: A cross-section of the nation speaks its mind*. New York: Doubleday.

Stouffer, Samuel A. 1992. Communism, conformity, and civil liberties: A cross-section of the nation speaks its mind (with a new introduction by James A. Davis). New Brunswick, NJ: Transaction

Study Panel on Fee-for-Service Medicare. 1998. *Final Report, from a generation behind to a generation ahead: Transforming traditional Medicare*. Washington, DC: National Academy of Social Insurance. January.

Sudman, Seymour. 1976. *Applied sampling*. New York: Academic Press.

Summers, Lawrence. 2008. How to build a U.S. recovery: The big freeze, part 4. policy response. *Financial Times*, August 7, 5.

Susser, Mervyn. 1973. *Causal thinking in the health sciences*. New York: Oxford University Press.

Suppes, Patrick C. 1970. *A probabilistic theory of causality*. Amsterdam: North-Holland.

Taylor, Brent, Elizabeth Miller, C.Paddy Farrington, Maria-Christina Petropoulis, Isabelle Favot-Mayaud, Jun Li, and Pauline A. Waight. 1999. Autism and measles, mumps, and rubella vaccine: No epidemiological evidence for a causal association. *The Lancet* 333(12 June): 2026–2029.

The Economist. 2009. Reforming American health care: Heading for the emergency room. June 27, 75–77.

Thomas, Vinod. 2001. Revisiting the challenge of development. In *Frontiers of development economics*, ed. Gerald M. Meier and Joseph E. Stiglitz, 149–182. New York: Oxford University Press.

Thomson, Elizabeth A. 2005. Laureate credits basic research as catalyst to success. MIT News Office, October 5.

Tilly, Charles. 1978. *From mobilization to revolution*. Reading, MA: Addison-Wesley.

Tilly, Charles. [1995] 1997. Democracy is a lake. In *Roads from past to future*, ed. Charles Tilly, 193–215. Lanham, MD: Rowman & Littlefield.

Tilly, Charles. 2000. Processes and mechanisms of democratization. *Sociological Theory* 18: 1–16.

Tilly, Charles. 2004. *Contention and democracy in Europe*. New York: Cambridge University Press.

Tilly, Charles. 2007. *Democracy*. New York and Cambridge, UK: Cambridge University Press.

Toner, Robin, and Janet Elder. 2007. Poll shows majority back health care for all. New York Times, March 1.

Tukey, John W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

United Nations Development Program. 1990. *Human development report 1990*. New York: Oxford University Press.

United Nations Development Program. 2001. *Human development report 2001: Making new technologies work for human development*. New York: Oxford University Press.

United Nations Development Program. 2002. *Human development report 2002: Deepening democracy in a fragmented world*. New York: Oxford University Press.

United Nations Development Program. 2002. *Arab human development report 2002: Creating opportunities for future generations*. New York: Oxford University Press.

United Nations Development Program. 2003. *Human development report 2003: Millennium development goals; A compact among nations to end human poverty*. New York: Oxford University Press.

United Nations Development Program. 2010. *Human development report 2010: 20th anniversary edition*. New York: Palgrave MacMillan.

United Nations Statistics Division, Department of Economic and Social Affairs. 2001. *Statistical yearbook*. New York: United Nations Publications.

United Nations, Department of Public Information. 2002. *Yearbook of the United Nations 2000*, vol. 54. New York.

Useem, B. 1998. Breakdown theories of collective action. *Annual Review of Sociology* 24: 215–38.

Vaux, Tony, and Francie Lund. 2003. Working women and security: Self employed women's association's response to crisis. *Journal of Human Development* 4: 265–287.

Vedantam, Shanker. 2009. Court rules autism not caused by childhood vaccines. *Washington Post*, downloaded from internet February 12.

Verba, Sidney. 1961. *Small groups and political behavior*. Princeton, NJ: Princeton University Press.

Verstraeten, T., R.L. Davis, F. DeStefano, T.A. Lieu, P.H. Rhodes, S.B. Black, H. Shinefield, R.T. Chen, and Vaccine Safety Datalink Team. 2003. Safety of thimerosal-containing vaccines: a two-phased study of computerized health maintenance organization databases. *Pediatrics* 112(5): 1039–48.

Wakefield, A.J., S.H. Murch, A. Anthony, J. Linnell, D.M. Casson, M. Malik, M. Berelowitz, A.P. Dhillon, M.A. Thomson, P. Harvey, A. Valentine, S.E. Davies, and J.A. Walker Smith. 1998. Ileal-lmphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet* 351(28 February): 637–641.

Wallerstein, Immanuel. 2002. The itinerary of world-systems analysis. In *New directions in contemporary sociological theory*, ed. Joseph Berger and Morris Zelditch Jr., 358–376. Lanham, MD: Rowman & Littlefield.

Washington Post-ABC News Poll, 2009, conducted August 13–17, 2009.

Watson, James L. 2004. Globalization in Asia: Anthropological perspectives. In *Globalization: Culture and education in the new millennium*, ed. Marcelo M. Suárez-Orozco and Desirée Baolin Qin-Hilliard, 141–172. Berkeley: University of California Press.

Weber, Max. 1947. The theory of social and economic organization, translated by A. M. Henderson and Talcott Parsons. Glencoe, IL: The Free Press.

Weisberg, Herbert I. 2010. *Bias and causation: Models and judgment for valid comparisons*. Hoboken, NJ: Wiley.

Welzel, C., R. Inglehart, and H. Klingemann. 2003. The theory of human development: A cross-cultural analysis. *European Journal of Political Research* 42: 341–379.

Wennberg, John E., and Megan M. Cooper. 1999. *The quality of health care in the United States: A report on the Medicare program*. Chicago, IL: AHA Health Forum.

Wennberg, John E., Jean L. Freeman, and William J. Gulp. 1987. Are hospital services rationed in New Haven or Over-Utilized in Boston? *The Lancet*, May 23:1185–1189.

Wermuth, Nanny. 2003. Analysing social science data with graphical Markov models. In *Highly structured stochastic systems*, Oxford statistical science series, vol. 27, ed. Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, 47–53. New York: Oxford University Press.

Wheeler, John R.C., and Thomas M. Wickizer. 1990. Relating health care market characteristics to the effectiveness of utilization review programs. *Inquiry* 27: 344–354.

White, Joseph. 1999. Uses and abuses of long-term Medicare cost estimates. *Health Affairs* 18(1): 63–79.

White, Ralph K., and Ronald Lippitt. 1960. *Autocracy and democracy: An experimental inquiry*. New York: Harper.

Wickizer, Thomas M. 1990. The effect of utilization review on hospital use and expenditures: A Review of the literature and an update on recent findings. *Medical Care Review* 47: 327–363.

Wickizer, Thomas M. 1991. Effect of hospital utilization review on medical expenditures in selected diagnostic areas: An exploratory study. *American Journal of Public Health* 81: 482–484.

Wickizer, Thomas M., John R.C. Wheeler, and Paul J. Feldstein. 1989. Does utilization review reduce unnecessary care and contain costs? *Medical Care* 27: 632–647.

Wickizer, Thomas M., John R.C. Wheeler, and Paul J. Feldstein. 1991. Have hospital inpatient cost containment programs contributed to the growth in outpatient expenditures? *Medical Care* 29: 442–451.

Wikipedia. 2009. Entry for Mark Geier (Downloaded October 3, 2009 from http://en.wikipedia.org/wiki/Mark_Geier).

Wilensky, Gail R., and Joseph P. Newhouse. 1999. Medicare: What's right? What's wrong? What's next. *Health Affairs* 18(1): 92–106.

Winship, Christopher, and Robert D. Mare. 1983. Structural equations and path analysis for discrete data. *The American Journal of Sociology* 89: 54–110.

Winship, Christopher, and Stephen L. Morgan. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology* 25: 659–707.

Winship, Christopher, and Michael Sobel. 2004. Causal inference in sociological studies. In *Handbook of data analysis*, ed. Melissa Hardy and Alan Bryman. Newbury Park, CA: Sage.

Wolf, Fredric M. 1986. *Meta-Analysis*: *Quantitative methods for research synthesis.* Sage University Paper series on Quantitative Applications in the Social Sciences, Newbury Park, CA: Sage.

Wolfinger, Russell D., and M. O'Connell. 1993. Generalized linear models: a pseudo-likelihood approach. *Journal of Statistical Computing and Simulation* 48: 233–243.

Wood, Elizabeth, et al. 1990. *Project STAR final executive summary report: Kindergarten through third grade (1985–1989)*. Nashville, TN: Tennessee State Department of Education.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2006. *Introductory econometrics: A modern approach*, 3rd ed. Mason, OH: Thomson/South-Western.

World Jewish Congress. 2004. *Website on Jewish Communities of the World*. New York: New York.

York, Richard, Eugene A. Rosa, and Thomas Dietz. 2003. Footprints of the earth: The environmental consequences of modernity. *American Sociological Review* 68: 279–300.

Zald, Mayer N., and John D. McCarthy. 2002. The resource mobilization research program: Progress, challenge, and transformation. In *New directions in contemporary sociological theory*, ed. Joseph Berger and Morris Zelditch Jr., 147–171. Lanham, MD: Rowman & Littlefield Publishers.

Zeisel, Hans. 1985. *Say it with figures*, 6th ed. New York, NY: Harper & Row.

Zeldin, A., and F. Pajares. 2000. Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal* 37(1): 215–246.

Zheng, Tian, Matthew J. Salganik, and Andrew Gelman. 2006. How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association* 101: 409–423.

Zuboff, Shoshana. 1988. *The age of the smart machine*. New York: Basic Books.

Zweifel, Thomas D., and Patricio Navia. 2001. Democracy, dictatorship, and infant mortality. *Journal of Democracy* 11: 98–114.

Index