3

This chapter offers a critical commentary on theory-based evaluation, stressing its utility as a method of program planning and as an adjunct to experiments but rejecting it as an alternative to experiments.

The False Choice Between Theory-Based Evaluation and Experimentation

Thomas D. Cook

It is currently fashionable in many foundation and some scholarly circles to espouse a theory of evaluation for complex, aggregated social settings such as communities and schools that depends on three steps:

Explicating the substantive theory of the program to be evaluated so as to detail all the flow-through relationships that should occur if the intended intervention is to have an impact on major target outcomes. In education such outcomes include achievement gains, and in welfare policy they include stable employment in the labor force.

Collecting data from a relevant sample of units (usually people) and in this way measuring each of the constructs specified in the substantive theory of the program.

Analyzing the collected data in order to assess the extent to which the postulated relationships have actually occurred in the predicted time sequence. If the data collection can cover only part of the postulated causal chain, then only part of the model will be tested. However, the aspiration is to test the complete program theory.

One major reason that this theory of evaluation is currently in vogue is as much because of what it is not as because of what it is. It is not a theory of evaluation that depends solely on qualitative methods. Such a theory would lack credibility in many academic and policy circles if the results from the qualitative studies were used to support inferences about what a program has

achieved for its clients or society at large. By contrast, the theory-based model seems scientific, acknowledging the importance of substantive theory, quantitative assessment, and causal modeling.

The theory-based approach to evaluation is also *not* experimental or even quasi-experimental. Advocates of theory-based evaluation point to how often experimental evaluations have produced disappointing results about community and school effects, leaving it unclear whether the programs are rarely effective or whether the experimental methods are too insensitive to register what a program has actually achieved. It is undeniable that there are very real and widely acknowledged practical difficulties that arise when doing real-world experiments with units higher than individuals, including communities and schools. It is therefore not illogical to want to shoot the messenger bearing the pessimistic message, leaving the message to live another day.

However, there are other reasons for preferring to shoot the messenger. If the message were correct, this would entail restricting or abandoning the cherished belief in *multiplier effects,* community-based forces that create impacts greater than the sum of individual effects. Belief in them sustains (and justifies) much of the funding aimed at intact communities and schools. So believing the messenger could also endanger the interests of the many program developers, researchers, and program funders whose reputations and jobs depend on "proper" evaluation making clear the general effectiveness of community-based interventions. But I do not want to be too cynical. It is also important to note that developers, funders, and researchers often encounter what they genuinely believe are effective projects during their visits to selected communities and schools. They see change occurring in these organizations, and they want a theory of evaluation that will register this change.

It is therefore a relief to them to learn of a theory of evaluation that claims to test causal effects validly, that promises to explain why these effects come about, and that does not concern itself with the inconvenient paraphernalia that experiments require in order to create a valid causal counterfactual against which to evaluate whether a program has caused any of the changes that might have been noted in a program group—namely, random assignment, close matching, pretests, and control groups. Espousing a theory of theory-based evaluation entails justifying research that includes the group experiencing the treatment and no one else. And its results will also seem scientific. If the causal modeling analyses suggest that the obtained data do not differ much from what the program theory predicts, then the presumption is that the validity of the theory and the success of the program have been demonstrated. Even if the time available for research does not permit assessing all the postulated causal links, the incomplete result can still be useful if it is congruent with the first part of the program theory. Such an incomplete result will at least inform staff about the quality of initial program implementation, as implementation variables are usually the first constructs in the causal model of the program. It can also be used against critics to argue for maintaining the program now that a data-

based rationale exists for believing the program could be effective in the future. Moreover if the promised results are not initially obtained, then it is evidently illogical to argue that a program is effective because it sets in motion the postulated mediating processes. These have demonstrably not occurred. All this is the promise of theory-based evaluation—both the positive justification based on what it accomplishes and the negative justification based on avoiding the anticipated pessimism associated with most past experimental evaluations of community interventions.

The question I ask here is, Can theory-based evaluations provide the valid conclusions about a program's causal effects that have been promised? I will reply in the negative, citing at least seven reasons for skepticism. They all have to do with the network of assumptions on which theory-based evaluation is premised—that a highly specific program theory is available, that the measurement is of high quality, that valid analyses of explanatory time-dependent processes have been conducted, and that everyone understands what is logically entailed if only part of a model has been tested in the time frame available. I will then go on to argue that theory-based evaluation techniques are extremely useful when used together with experiments, rather than in opposition to them. When added to experiments, they will focus needed attention on what the program theory is, what level of program implementation is obtained, which presumed causal mediating processes actually change, and how this variation in implementation quality is related to variation in distal outcomes.

Reasons for Skepticism

First, it was my experience in coauthoring a paper on the theory of a program with its developer (Anson and others, 1991) that program theories are not always very explicit. More important, in this case the theory could have been made more explicit in several different ways, not just one. Is there a single theory of a program, or are there several possible versions of it? I am definitely inclined from experience to favor the latter. And I do this not because of the obvious point that every program is dynamic and hence changing over time. Rather I would argue that it can be construed in multiple different ways even at any one time. This multiplicity of possible program theories entails a large (but not necessarily insurmountable) problem for a theory of theory-based evaluation.

Second, most of the program theories with which I am acquainted are very linear in their postulated flow of influence. They rarely incorporate reciprocal feedback loops or external contingencies that might moderate the entire flow of influence. Yet we know from bitter experience that how individuals have been affected by a program affects their subsequent exposure to the program, sometimes because they come to need it less and sometimes because they come to need it more. And we also know that programs do not exist in political, social, or cultural vacuums. They are contextually embedded, and these contexts affect how the programs work and how individuals

and groups react to them. To postulate closed systems, clearly differentiated category boxes, and exclusively unidirectional causal arrows is all a little too neat for our chaotic world. It is better to assume constant external perturbations, constructs with fuzzy rather than clear boundaries, and causation that is reciprocal rather than unidirectional. Unfortunately, testing theories based on these more realistic but also more complex assumptions entails many more technical difficulties than testing simple linear models based on clearly independent constructs within a closed explanatory system.

Third, few program theories specify how long it should take for a given process to affect some proximal indicator in the causal chain. But without such specifications, it is difficult to know when a disconfirmation occurs, whether the next step in the model has simply not occurred yet or instead will not occur at all. It is this ambiguity about time lines that allows program developers who have been disappointed by evaluation results to claim that positive results would have occurred had the evaluation lasted longer. Given program theories with specific time lines, this particular argument would never be heard. But because such theories are not typically available, the argument is often heard when developers do not like what the evaluator reports. (This is not the fault of program developers, of course. The problem lies with the quality of our social science knowledge in general).

Fourth, theory-based evaluation places a great premium on knowing not just when to measure but how to measure. When measures are only partially valid, failure to corroborate a model is ambiguous in its implications. Does the failure reflect a program theory that is false—the desired inference—or does it reflect measures that were inadequate for a strong test of the theory? Researchers can protect against this dilemma by explicating constructs better initially, by choosing more reliable single measures, and by using multiple measures of the same construct. Although such procedures are always desirable in social research, they are probably nowhere more necessary than when using a theory-based approach to evaluation. It is unfortunate then that better and more extensive measurement costs money. In addition, it can be burdensome to respondents, including staff and students within communities and schools. Still, this objection based on the quality of measurement is essentially practical rather than theoretically fundamental, given that we can usually improve our measurement if we are willing to pay the opportunity costs. The major of these is that for a fixed budget fewer constructs will tend to be measured if it is important to raise the quality of assessment of individual constructs.

Fifth, there is the epistemological problem that many different models can usually be fit to any single pattern of data (Glymour, Scheines, Sprites, and Kelly, 1987). The causal modeling methods usually espoused by advocates of theory-based program evaluation do not permit falsifying among competing models. They do not allow us to ascertain whether different models with the same (or additional) variables would fit the data at least as well or better than the model under test. This leads to an apparent paradox. Theory-based evaluations are predicated on using theory to predict outcomes and not to explain how they came about, all appearances and rhetoric

to the contrary. To discover a complex, multivariate pattern of data that matches what was predicted provides one plausible model of how the variables are interrelated but not necessarily the correct one.

The sixth and biggest problem with a theory of evaluation that depends on a program's substantive theory alone is that there is no valid counterfactual, no way of knowing what would have happened at any stage in the model had there not been the program. As a result, it is logically impossible to say whether any processes that are observed are genuine products of the intervention or whether they would have occurred anyway, even without the reform. How can we rule out all the threats to internal validity outlined in Cook and Campbell (1979)? The biggest struggle in evaluation is around summative claims—that is, claims that a program has or has not caused some observed consequence. Theory-based evaluation does not take on this central issue; it sidesteps it.

There is one circumstance, though, in which the claim has been made that causal inference can be justified without controlled assignment, control groups, pretests, and the like. This circumstance involves *signed causes* (Scriven, 1976), situations in which the postulated pattern of multivariate relationships is so unique that it could not have occurred other than through the availability of the reform. Unfortunately, signed causes depend on access to considerable well-validated substantive theory (Cook and Campbell, 1979). Detectives can "finger" a suspect because the crime scene provides a multivariate pattern of clues, because they already know the modus operandi (MO) of various suspects, because they presume to know all the relevant suspects using this MO, and because they can use interviews to discriminate among suspects if more than one of them has an MO matching the evidence at the crime scene. Likewise, pathologists can ascertain the cause of death because they have the multivariate evidence laid out before them on the dissecting table as a pattern of effects and because past research in anatomy, physiology, and the like has taught them how to identify the specific pathways through which individual diseases affect some organs but not others. Rarely do social scientists have such specific background information available to them from substantive theory and experience, so discriminating among alternative causes is much more difficult. And rarely is the pattern of effects to be explained as clear-cut as the crime scene that a detective finds or the body that a pathologist dissects. So the theory of signed causes is not likely to be a widely applicable alternative to a valid counterfactual control group. Indeed it is just an earlier form of theory-based evaluation.

So the best safeguard for those who place a high premium on identifying causal effects is to have at least one well-matched comparison group, and the best comparison group is a randomly constructed one. So we are back again with the proposition that theory-based evaluations are useful as complements to experiments but not as alternatives to them, and preferably as complements to randomized experiments rather than quasi-experiments.

My final reason for being against theory-based evaluation is not intrinsic to the method. But I do fear that it could be used to postpone doing

hard-headed experimental work on programs. Many practical-minded advocates of specific reforms realize how difficult it will be to bring about substantial changes in distal outcomes, given inevitable shortfalls in program theory and implementation as well as in evaluation sensitivity, not to speak of the limitations associated with the short time lines within which change is often called for. The advocates' hope is that implementing a reform with vigor and theoretical fidelity will entail little dilution of influence across all the probabilistic links in a program's substantive theory. But their more realistic expectation—promises to funders notwithstanding—is that implementation will be weaker than desired and that some of the planned intervening processes may not come about even if the program theory is true. So it is tempting for program developers and those with a similar stake in the program's success to concentrate on the first steps in the program's theory, steps that refer to implementation issues and therefore seem to need control groups less urgently. However, if the initial steps in the theory do come about as planned, this will surely lead to the temptation to claim (illogically, as it happens) that later effects are more likely to come about because the earlier ones already have. Given the stakes and the probability of demonstrating success with distal-outcome criteria, it is easy to see how the advocacy of theory-based evaluation could become an excuse not to evaluate reforms by traditional summative means.

Theory-Based Measurement and Analysis Within Experiments

I am resolutely in favor of evaluators measuring and analyzing theoretically specified mediating processes. If all the theories of a program that one can construct postulate relationships one knows to be generally false, then this indicates that the program is not likely to be worth much and is certainly not worth squandering evaluation resources on. Attention to program theory also helps place special emphasis on implementation quality. This is because the first variables in the causal sequence are the most often assessed, and they are usually tapped into implementation. It is my belief that evaluators with summative aspirations do not spend enough time dealing with implementation issues, even though implementation shortfalls are one important reason why results are often disappointing. Finally, I think that we need to know why programs are or are not effective. To learn this absolutely requires the measurement and analysis of data that are subsequent to implementation but prior to distal outcomes. So I am a fan of theory-based evaluation and have recently deliberately used the expression in the title of two articles (Cook and others, 2000; Cook, Hunt, and Murphy, 2000) evaluating Comer's School Development Program (Comer, 1980).

But these articles are about randomized experiments with whole schools as the unit of assignment and analysis. The studies were designed both to describe and to explain the program's consequences for school staff and stu-

dents, using the randomized experiment part to describe causal relationships and using the theory-based part to help explain the pattern of results obtained. Thus I have tried to model ways of conducting evaluations that combine experimental designs and the analysis of such intervening processes as program implementation quality and early substantive process effects.

It is not easy to do experiments with intact communities and whole schools. There are many reasons for this, having to do with the sample size of units that one can afford to study and that are willing to participate, the highly variable program exposure within a setting, the treatment crossovers, the differential attrition, the politics and ethics of gaining access, the limitations of generalizations that arise from dealing only with settings willing to volunteer to be in the study, and so forth. I have yet to meet a perfect community or even school-based experimental evaluation, and my own studies certainly do not merit such an appellation. Particularly worrisome, in my view, is that experimental work with intact communities will often be very expensive if a sufficiently large number of communities is to be included in the design. One might even argue that the depressing picture of community-level effects that has emerged from experimental evaluations is deceptive, based on studies so small as to have little chance of showing effects. So experimental evaluation needs to be undertaken more often, taking advantage of all that has been learned about implementing randomized experiments over the last twenty years.

The great advantage of experiments (or of close approximations) is that the test, from the intervention to the individual intervening processes, is unbiased or involves less bias than the alternative approaches to evaluation. This is because experiments are designed to examine whether each step in the causal model is related to the planned treatment contrast. But a causal model involves other tests, especially of the path from intervening processes to the planned distal outcomes. These tests are potentially biased. In essence, they depend on stratifying units by the extent to which the postulated theoretical processes are faithfully reproduced before examining how this variation in implementation is related to variation in the outcome. Still, these second-stage observational analyses are well worth doing, though their results should be clearly labeled as more tentative than the results of any planned experimental contrast.

In this context, it is interesting to note that Angrist, Imbrens, and Rubin (1996) have argued that it is possible to obtain unbiased estimates of the consequences of intervening processes—but only when there is random assignment. This is because such assignment can function as an unbiased instrumental variable. So if they are correct, unbiased causal inferences are sometimes possible both from the treatment to the intervening variables and from the intervening variables to the distal outcomes. Unfortunately, the method of Angrist, Imbrens, and Rubin has not yet been generalized to handle the multiple different intervening variables that a program theory postulates will change at different times in a causal sequence. Hence we still

need to conduct traditional causal modeling analyses of the pattern of influence from the intervention to the various mediating variables and then from these mediators to a distal outcome.

Few evaluators will argue against the more frequent and sophisticated use of substantive theory to detail intervening processes. Probably the sole exceptions are those who believe that the act of measuring process creates conditions different from those that would apply in the actual policy world. Few evaluators argue that it is not possible to collect measures of intervening processes. So it should be possible to construct and justify a theory-based form of evaluation that complements experiments and is in no way an alternative to them. It would prompt experimenters to be more thoughtful about how they conceptualize, measure, and analyze intervening process. It would also remind them of the need to first probe whether an intervention leads to changes in each of the theoretically specified intervening processes and then explore whether these processes could plausibly have caused changes in the more distal outcomes of policy interest. I want to see theory-based methods used within an experimental framework and not as an alternative to it.

References

Angrist, J. D., Imbrens, G. W., and Rubin, D. B. "Identification of Causal Effects Using Instrumental Variables." Journal of the American Statistical Association, 1996, 91, 444-462.

Anson, A., and others. "The Comer School Development Program: A Theoretical Analysis." Journal of Urban Education, 1991, 26, 56–82.

Comer, J. P. School Power. New York: Free Press, 1980.

Cook, T. D., and Campbell, D. T. *Quasi-Experimentation:* Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin, 1979.

Cook, T. D., Hunt, H. D., and Murphy, R. M. "Comer's School Development Program in Chicago: A Theory-Based Evaluation." American Education *Research Journal*, forthcoming, summer 2000.

Cook, T. D., and others. "Comer's School Development Program in Prince George's County, Maryland: A Theory-Based Evaluation." American Educational *Research Journal*, forthcoming, winter 2000.

Glymour, C., Scheines, R., Sprites, P., and Kelly, K. Discovering Causal Structure: *Artificial* Intelligence, Philosophy of Science, and Statistical Modeling. Orlando, Fla.: Academic Press, 1987.

Scriven, M. "Maximizing the Power of Causal Investigation: The Modus Operandi Method." In G. V. Glass (ed.), Evaluation Studies Review Annual. Thousand Oaks, Calif.: Sage, 1976.

THOMAS D. COOK is professor of sociology at Northwestern University.