2

This chapter explores the relative strengths and *weaknesses of program* theory as *a* tool for *inferring* causality and outlines a five-stage *approach* that makes increased use *of* inductively built *program theories* and takes more deliberate account of the varying *levels* of certainty that are *required for* evaluative conclusions.

Ascertaining Causality in Theory-Based Evaluation

E. Jane *Davidson*

"Causation. The relation between mosquitos and mosquito bites" (Scriven, 1991, **p.** 77). Although causality is easily understood in everyday life, formulating a precise definition that spells out how it must be demonstrated proves to be considerably more difficult (for example, Sosa and Tooley, **1993).** Rather than delving into this difficult philosophical debate, this chapter focuses on the main issue for the practicing evaluator—determining whether observed changes are due to the program (and can correctly be referred to as program effects), are due to some other cause, or are purely coincidental.

Why is causality important? If an evaluator erroneously concludes that **a** program is meritorious (because **it** is thought to have caused some positive changes), resources may be wasted on continuing it or expanding it in its current form. In addition to the obvious monetary costs to funders, there are serious opportunity costs for recipients and program staff, who could be putting their time and efforts into something more worthwhile. Conversely, a good program might be discontinued or altered if negative changes are wrongly attributed to it or if its positive effects are thought to be due to something else. In other words, causality is not merely an issue of relevance to academics; it deeply affects the lives of many stakeholder groups, whether they realize it or not.

The attribution of causality to programs and other types of evaluand is a daunting challenge. However, there are a number of methods available, both traditional and nontraditional, that can help the practitioner address this issue. This chapter explores the relative strengths and weaknesses of program theory as a tool for inferring causality and makes two broad rec-

ommendations: pay greater attention to the decision-making context, and increase the use of inductively built program theories. A set of criteria for inferring causality is outlined, as is a step-by-step strategy intended to guide the practitioner through the process of building an evidence base for causal attributions.

There are a number of terms currently in use that describe the kind of evaluation first proposed by Suchman (1967) and later developed by Weiss (1972), Bickman (1987), Chen (1990), Lipsey (1993), and others. In this chapter, I have used the term *theory-based* rather than *theory-driven* in order to avoid the impression that any particular theory should "drive" the evaluation. I have also avoided the term *program theory evaluation* because it could be construed to imply that the main task is the evaluation of program theory rather than the evaluation of the program itself. As some examples in the next section illustrate, these definitional issues are far more important than they first appear because they may lean the evaluator toward approaches to using program theory that limit his or her ability to infer causality. In this chapter, the term *theory-based evaluation* refers to the use of program theory (or *logic models*) as a framework in the determination of merit or worth.

Testing Causal Mechanisms with Theory-Based Evaluation

The introduction and development of theory-based evaluation and its variants (for example, Bickman, 1987; Chen, 1990; Lipsey, 1993; Suchman, 1967; Weiss, 1972) sparked an increased focus on understanding the mechanisms by which programs produce their effects. This is an important development because we can increase our confidence in a causal claim if we have some understanding of the causal mechanism involved (Sayer, 1992). A claim such as "drug resistance education programs cause increased drug use" becomes considerably more convincing when at least one element in the causal chain is illuminated. For example, Donaldson, Graham, and Hansen (1994) discovered that students who received such a program not only reported higher levels of drug use in later years but also considered drug use in school to be significantly more prevalent than those who had not taken the resistance program. When the links between these elements were tested and found to be significant, the evidence in favor of a causal effect was enhanced still further.

Despite the purported focus of theory-based evaluation on investigating the causal mechanisms by which a program achieves its effects, surprisingly few actually do this. For example, Weiss (1997) inspected some thirty studies that claimed to draw on program theory and found that very few of them measured the mediator variables identified in the model, let alone tested the links between them and the program and its outcomes. However, logic models are often used in other ways.

Many evaluations appear to use program theory as a framework for determining the variables that should be measured in an evaluation or as a means of better understanding the evaluand. For example, an evaluation of Project TEAMS (Technology Enhancing Achievement in Middle School), meta-evaluated by Cooksy (1999), used a logic model to provide a conceptual representation of the program, of which a small fraction of the variables were measured (often necessary when there are budgetary constraints). The model provided a useful conceptualization of the program, although the causal links specified in the model were not tested. Accordingly, Cooksy noted that one of the major weaknesses of the evaluation was "the inability to attribute the changes reported to TEAMS" (p. 135).

Saxe and Tighe's (1999) evaluation of Fighting Back, a series of community programs designed to combat alcohol and drug use, encountered a similar problem. The authors used a very comprehensive logic model to identify a wide range of important variables that were measured as part of the evaluation, thereby ensuring good coverage of key performance criteria. However, Saxe and Tighe also reported difficulty inferring causality, attributing this to their inability to use randomized designs and the unavailability of control groups.

In both of the preceding cases, the evaluators appear to have been constrained in their ability to employ more complex experimental or quasi-experimental designs (Cook and Campbell, 1979), a factor that no doubt contributed to their problems with the causality issue. But how might it have helped to test the linkages among the variables in the program models, and what could they have concluded from such tests?

Reynolds's evaluation (1998) of the High/Scope Perry Preschool Program for disadvantaged children not only specified a comprehensive logic model for the evaluand but also tested the links in the model using structural equation modeling. When testing the causal mechanism, expected changes in the mediators and outcome variables, coupled with a good fit of the data to the hypothesized model, were considered to be strong evidence that the program was indeed causing the effects. However, as Reynolds noted, "Inferences about treatment are largely dependent on the validity of the program theory" (p. 217).

Reynolds's comment hits squarely on the greatest weakness of the theory-testing approach to the use of logic models in evaluation. Even when the model appears to be strongly supported, the fact remains that any number of models might fit the data, making it impossible to conclude that a model is "correct" simply because it "fits" (Cohen and Cohen, 1983).

If alternative logic models are developed exclusively from existing theory or stakeholder input, or both, as recommended by Chen (1990), the possibility remains that one or more important causal chains (or alternative explanations) exist that are not covered by existing theory or did not occur to either stakeholders or evaluators. This may not be fatal in an evaluation whose primary purpose is to build organizational capacity for self-evaluation and inquiry

or in instances where other evidence already provides the required level of certainty about causal attributions. However, failing to check a possible causal chain may be a serious flaw in a high-stakes evaluation, when accurate causal inference is critical to being able to draw defensible conclusions.

Hunting for Causal Mechanisms

The difficulty associated with tracking down potential causes is very similar to that of hunting for side effects. Most evaluators who use program theory would agree that unintended consequences are just as important to track down as goal-related outcomes (for example, Chen, 1990). However, a program model that is generated from program goals (or theory derived from the academic literature or from stakeholders) inevitably focuses primarily on intended outcomes and is therefore in considerable danger of failing to include important potential side effects as variables in the model (Rogers and McDonald, 1999).

The same will be true in the hunt for causal explanations using these theories. A causal model built using social science or stakeholder theory, or both, may fail to include not only potential side effects (and the causal chains leading to them) but also any causal paths not predicted by the program theory.

There has been considerable work in the development of qualitative methods that trace causal chains of events to produce defensible conclusions. Examples include the *modus operandi method* used by detectives to investigate crime (Scriven, 1974a), *process tracing* (Ford and others, 1989), and *qualitative causal modeling* (Miles and Huberman, 1994). Because these methods have strong similarities, it would be redundant to discuss each one separately. Scriven's (1974a) modus operandi method provides the best *conceptual* description of the logic behind these methods and will be outlined in more detail here.

The modus operandi method uses the detective metaphor to describe the way in which potential causal explanations are identified and tested. Scriven describes how chains of causal events often leave signature "traces" that the evaluator tracks down by moving both up and down the causal chain. Starting with the observed effects (the "clues"), one can move up the causal chain, identifying what might have caused them. Using the previous example of the drug resistance training program (Donaldson, Graham, and Hansen, 1994), one could start with the increase in self-reported student drug use and search for possible reasons—for example, by asking students what they think gave rise to the increase or by observing their behavior.

In the opposite direction, one can start with the program itself (the "suspect") and trace down the causal chain to see what impacts it might have had, and through what mechanisms. If evidence is consistent with the expected trace left by a particular causal chain, then confidence in that chain

as the correct causal explanation is increased. Evidence that contradicts the expected trace eliminates that causal chain as a possibility, and missing evidence makes the explanation more doubtful. In the Donaldson, Graham, and Hansen (1994) study, the increased prevalence estimates support the explanation that the program increased self-reported drug use by making students think drug use was more widespread than they had previously believed. Had prevalence estimates not increased, that causal chain would have been eliminated as a possibility.

Methods of Inferring Causality

Scriven (1974a) describes two methods for inferring causality using the modus operandi method. The first is *causal list inference:* suppose we have a list of almost all possible causes of a certain effect. If the effect occurred, and only one of the possible causes occurred, then it is very probable that that was the true cause. The second is *modus operandi inference:* if more than one of the possible causes occurred, but only the characteristic causal chain (or modus operandi) for one of those possible causes was present, then that was probably the cause, especially if the modus operandi is highly distinctive.

What other criteria might the evaluator use to help build an evidence base for causal inference? The three most commonly used criteria are those proposed by philosopher David Hume (as cited in Huberman and Miles, 1998). They are *temporal precedence* (A before B), *constant conjunction* (when A, always B), and *contiguity of influence* (a plausible mechanism links A and B). Unfortunately, there are problems with some of these criteria if taken too literally. For example, the constant conjunction criterion might imply that a program should have similar impacts on every type of recipient in every context, and the contiguity of influence criterion might lead one to discount mechanisms that seem implausible because they are not well known. However, all are still useful rules of thumb for checking causal claims, as long as they are not applied too rigidly.

Huberman and Miles (1998) suggest four additional criteria taken from the field of epidemiology. They are *strength of association* (much more B with A than with other possible causes), *biological gradient* (if more A, then more B), *coherence* (the A-B relationship fits with what else we know about A and B), and *analogy* (A and B resemble the well-established pattern noted in C and D). Although not all of these criteria would fit every possible evaluation situation, and they could not alone constitute sufficient evidence to infer causality, they clearly add some useful sources of evidence that could be used by evaluators with access to either qualitative or quantitative data. The following is a list of the *nine potential types of evidence for inferring causality:*

1. Causal list inference (almost all Bs are caused by A, A', A", . . ., or A_n; B has occurred; A did occur; but A', A", . . ., A_n did not occur).

2. Modus operandi inference—use if more than one possible cause occurred (only the characteristic causal chain/modus operandi for one of the possible causes, A, was present; inference strengthened if the modus operandi in question is highly distinctive).
3. Temporal precedence (A before B).
4. Constant conjunction (when A, always B).
5. Contiguity of influence (a plausible mechanism links A and B).
6. Strength of association (much more B with A than with other possible causes).
7. Biological gradient (if more A, then more B).
8. Coherence (the A-B relationship fits with what else we know about A and B).
9. Analogy (A and B resemble the well-established pattern noted in C and D).

Causal Tracing: A Five-Stage Process

The nine potential types of evidence for inferring causality provide a useful starting point for the practitioner attempting to develop a body of evidence that either confirms or refutes whether a program did in fact give rise to observed changes. But how can the practitioner maximize the power of theory-based evaluation to build an evidence base for causal attribution, given the problems highlighted earlier with the traditional theory-testing approach?

Two changes to the way we use theory-based evaluation should help alleviate these problems. The first is gaining a solid understanding of the decision-making context, as well as the information needs of the client. The second is supplementing tests of program theory with a more open-ended causal tracing and inductive approach to theory building, an approach referred to here as *goal-free theory-based evaluation.* To illustrate these recommendations and guide the practitioner, I would like to propose a *five-stage process for causal tracing using theory-based evaluation.*

1. Information needs assessment, which determines required level of certainty
2. Goal-free search for all important changes
3. Inductive hunt for the causes of those changes =, which builds draft logic model, possibly using input from program staff in the later stages
4. Supplementation of inductive program theory with additions of possible effects and causal chains from the literature
5. Test of the revised logic model, preferably using information sources and recipients not used to build the inductive model

In order to conclusively demonstrate causality in a particular case, one would have to eliminate all other possible causal explanations, including those whose mechanisms are not yet understood. In reality, 100 percent certainty is an unachievable (not to mention unnecessary) goal. As

in a criminal or civil trial, the evidence needs to be sufficiently compelling to satisfy the relevant standards of proof in that context (for example, *beyond a reasonable doubt* or *preponderance of the evidence*). Similarly, in evaluation it is necessary to produce a body of evidence that will stand up to the scrutiny to which it will be subjected and that is commensurate with the relative costs of Type I errors (in causality terms, erroneously attributing a coincidental charge to a program) and Type II errors (erroneously concluding that an effect caused by the program was coincidental) in the given context. Accordingly, *stage one* in the causal tracing process involves an information needs assessment, in order to establish the degree of certainty to which conclusions about causality must be drawn. This should preferably be addressed when assessing program evaluability—that is, what information is needed, to what degree of certainty, and within what time and budgetary constraints? Note that it is not simply a matter of asking clients what they *think* is needed. Rather it is up to evaluators to determine this, based on multiple sources of information, using their evaluation expertise.

Bearing in mind the standard of proof needed for a particular evaluation, *stage two* begins with an open-ended and goal-free effort to detect all important outcomes of positive or negative value (Scriven, 1974b). Ideally, this should be carried out by someone thoroughly trained in the discipline of evaluation, but with minimal substantive knowledge of the particular type of evaluand and no knowledge of the program's specific goals. This ensures that the search does not focus primarily on outcomes that would have been predicted by theory or that were intended by staff, thereby reducing the chance that something important will be missed. The scope of this task will be dictated by the time and budgetary constraints associated with the evaluation, so it need not necessarily be hugely time consuming.

Stage three involves the goal-free theory-based approach—the use of inductive techniques to trace the causes of the effects uncovered in stage one. Here it would be ideal to use an individual or team that was highly skilled in the qualitative methods described earlier, with at least one member who had little knowledge of academic theories pertaining to the evaluand. At this point, the evaluation may start to incorporate the implicit theories held by program recipients or staff. However, it is not necessary (and may not be desirable) to fully involve them in the theory development process unless there are compelling reasons for doing so (for example, a lack of buy-in for evaluation results could cause serious problems in trying to implement improvements). For the evaluator on a limited budget and a tight time line, this step may involve some extended questioning of program recipients about what they attribute certain changes to. For a high-stakes evaluation, the approach will need to be considerably more thorough. In each case, the depth and breadth of the hunt for causes will need to match the required standards of proof established in stage one.

By the end of stage three, there should be a draft program theory created through a fully inductive process. This is developed further in *stage four*

to incorporate any relevant theories from the literature that might shed light on missing links or puzzling data and that might identify new mechanisms that could explain the observed results. Although a comprehensive literature review would be ideal for the large-scale evaluation, the evaluator on a tight budget may find that perusal of the draft logic model by a subject matter expert would be sufficient. Program goals may also be incorporated at this point to make sure that the evaluation covers all the information the client needs.

In *stage five,* the model is tested using a combination of qualitative and quantitative methods. The focus here should be not so much on the overall statistical fit of the model, but on the testing and elimination of alternative causal explanations for the observed effects, using both qualitative and quantitative methods. This should continue until there is a sufficient body of evidence to satisfy the appropriate standards of proof for the given situation. For the evaluator on a shoestring budget, this may involve some observation or interviewing to test any potentially suspect causal paths.

Conclusions

This chapter has noted some of the difficulties associated with the attribution of causality to programs and other types of evaluand and the strengths and limitations of theory-based evaluation as a tool for doing so. The main weakness of theory-based evaluation in this respect was its overreliance on the validity of a program theory that rested on prior knowledge, either from the social science literature or from program staff.

The hunt for alternative causal explanations in addition to predicted explanations is as important and as difficult as the hunt for side effects in addition to intended outcomes. Missing just one serious alternative could spell the difference between a valid and an invalid conclusion, and there are no predetermined road maps for ensuring that all possibilities have been covered. In an attempt to guide the practitioner through this challenging task, a five-stage strategy is outlined in this chapter. This combines goal-free, inductive, and theory-testing modes of investigation and offers options for practitioners operating under a range of budgetary and time constraints.

A second point noted in this chapter is that the weight and quality of evidence required to infer causality varies dramatically depending on the context in which the evaluation is being conducted. The usual standards from the social sciences (which are roughly equivalent to beyond a reasonable doubt) will be too lenient in some situations and too stringent in others. What is crucial is that the required level of certainty is ascertained by the evaluator early on in the evaluation planning stage, that estimates are made of what evidence is required to meet that standard, and that decisions are made as to whether theory-based or other approaches are used.

Finally, as Cook (Chapter Three) notes, theory-based evaluation should not be seen simply as a replacement for experimental and quasi-experimental designs. For high-stakes evaluations with large budgets and extended time lines, the two may be used in conjunction to allow virtually bulletproof causal attributions, provided they are used skillfully. For the everyday evaluator under more serious time and budgetary constraints, ideas from both methodologies should be considered in order to build evidence for inferring causality (see the list of potential types of evidence for inferring causality earlier in this chapter). The depth and breadth of the required evidence base is a key consideration in evaluation planning and should be based on a thorough assessment by the evaluator of stakeholder information needs. This not only will help with budgeting the evaluation more accurately but also will facilitate any up-front discussions with the client about the trade-offs between budgets, time lines, and the certainty of conclusions.

References

Bickman, L. "The Functions of Program Theory." In L. Bickman (ed.), Using Program Theory in Evaluation. New Directions for Program Evaluation, no. **33.** San Francisco: Jossey-Bass, 1987.

Chen, H. T. *Theory-Driven* Evaluations. Thousand Oaks, Calif.: Sage, 1990.

Cohen, J., and Cohen, C. *Applied Multiple Regression/Correlation* Analysis for the *Behavioral* Sciences. (2nd ed.) Hillsdale, N.J.: Erlbaum, 1983.

Cook, T. D., and Campbell, D. T. Quasi-Experimentation: Design and Analysis **Issues** for Field Settings. Skokie, Ill.: Rand McNally, 1979.

Cooksy, L. J. "The Meta-Evaluand: The Evaluation of Project Technology Enhancing Achievement in Middle School." *American Journal of* Evaluation, 1999, *20,* 123–136.

Donaldson, S. I., Graham, J. W., and Hansen, W. B. "Testing the Generalizability of Intervening Mechanism Theories: Understanding the Effects of Adolescent Drug Use Prevention Interventions." Journal *of* Behavioral Medicine, 1994, 17, 195–216.

Ford, J. K., and others. "Process Tracing Methods: Contributions, Problems, and Neglected Research Questions." *Organizational* Behavior and Human Decision *Processes,* 1989, *43,* 75–117.

Huberman, A. **M.,** and Miles, M. B. "Data Management and Analysis Methods." In N. K. Denzin and **Y.** S. Lincoln (eds.), Collecting ***and*** Interpreting Qualitative Materials. Thousand Oaks, Calif.: Sage, 1998.

Lipsey, M. "Theory as Method: Small Theories of Treatments." In L. Sechrest and **A.** Scott (eds.), Understanding Causes and Generalizing About Them. New Directions for Program Evaluation, no. 57. San Francisco: Jossey-Bass, 1993.

Miles, **M.** B., and Huberman, **A.** M. Qualitative Data Analysis: An Expanded *Sourcebook.* (2nd ed.) Thousand Oaks: Calif.: Sage, 1994.

Reynolds, A. J. "Confirmatory Program Evaluation: A Method €orStrengthening Causal Inference." *American Journal of* Evaluation, 1998, 19, 203–221.

Rogers, P. J., and McDonald, B. "Three Achilles Heels of Program Theory Evaluation." Paper presented at the international conference of the Australasian Evaluation Society, Perth, Australia, Oct. 1999.

Saxe, L., and Tighe, E. "The View from Main Street and the View from 40,000 Feet: Can a National Evaluation Understand Local Communities?" In J. Telfair, L. C. Leviton, and J. S. Merchant (eds.), Evaluating Health **and** Human Service Programs in Community Settings. New Directions for Evaluation, no. **83.** San Francisco: Jossey-Bass, 1999.

Sayer, A. Method in Social Science: A Realist Approach. London: Routledge, 1992.
Scriven, M. "Maximizing the Power of Causal Investigation: The Modus Operandi Method." In W. J. Popham (ed.), Evaluation in Education: Current Applications. Berkeley, Calif.: McCutchan, 1974a.
Scriven, M. "Prose and Cons About Goal-Free Evaluation." In W. J. Pophsm (ed.), Evaluation in Education: Current Applications. Berkeley, Calif.: McCutchan, 197413.
Scriven, M. Evaluation Thesaurus. (4th ed.) Thousand Oaks, Calif.: Sage, 1991.
Sosa, E., and Tooley, M. (eds.). Causation. New York: Oxford University Press, 1993.
Suchman, E. A. Evaluative Research: Principles and Practice in Public Service and Social Action Programs. New York: Russell Sage Foundation, 1967.
Weiss, C. H. Evaluation Research: Methods of Assessing Program Effectiveness. Englewood Cliffs, N.J.: Prentice Hall, 1972.
Weiss, C. H. "Theory-BasedEvaluation: Past, Present, and Future." In D. J. Rog and D. Fournier (eds.), Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods. New Directions for Evaluation, no. 76. San Francisco: Jossey-Bass, 1997.

E. JANE DAVIDSON teaches consulting psychology at Alliant University/California School of Professional Psychology in San Diego and is completing a doctorate in evaluation and organizational behavior at Claremont Graduate University.