⚜ FIVE ⚜

DEALING WITH THE CAUSATION ISSUE

⎯⎯⸱⬥⸱⎯⎯

No discussion of evaluation nuts and bolts is complete without some mention of the **causation** issue. Although this is a relatively simple concept to grasp in everyday life, causation is both one of the most difficult and one of the most important issues in evaluation. Even if we observe changes that are consistent with the expectations or goals of a program or another evaluand, we cannot correctly refer to these as "impacts" or "outcomes" unless we can demonstrate that the evaluand was at least a primary cause of those changes.

Strategies for inferring causation form a key part of what should be written into the Methodology checkpoint of the Key Evaluation Checklist (KEC) (Exhibit 5.1). The choice of evaluation design affects the evaluation team's ability to make causal inferences. Causation is also relevant for the Outcomes checkpoint because identifying anything as an outcome is saying that it was caused by the evaluand.

One of the great challenges with causation is that the further down the causal chain (toward what we might call "ultimate outcomes") one goes, the more other factors come into play. For example, the career success of a university graduate can be attributed not only to the quality of education he or she received but also to the quality of mentoring and advancement opportunities after graduation, support from family, aptitude or intelligence, and many other factors. The fact that so many variables are in play makes it quite difficult to pin down whether career success (and other downstream changes) is substantially

Exhibit 5.1   The KEC Checkpoints for Which the Causation Issue Is Most
                   Relevant

III. Methodology
What is the overall design of the evaluation (e.g., quasi-experimental,
participatory, goal free)? Explain why (briefly).

---

7. Outcome Evaluation
How good or valuable are the impacts (both intended and unintended) on
immediate recipients and other impactees?

---

due to the evaluand (in this case, the program from which the student gradu-
ated) or can be attributed mostly to other factors.

Although the causation issue is incredibly important, demonstrating causal
links can seem like an impossible task, especially for evaluators with limited time
and resources (most of us likely fall into that category). For this reason, many
people abandon the issue altogether, either by tacking on a bunch of disclaimers
to their evaluations or by downplaying the importance of causal analysis.

Here is the good news: There is some practical light at the end of the cau-
sation tunnel, and the tunnel is not nearly as long and treacherous as legend
has it. To deal with the causation issue, we need to answer four important ques-
tions in the following order:

- How certain does the client need us to be to say that the evaluand
  "caused" a certain change?
- What are the basic principles for inferring causation?
- What types of evidence do we have available to help us identify or rule
  out possible causal links?
- How should we decide what blend of evidence will generate the level
  of certainty needed most cost-effectively?

CERTAINTY ABOUT CAUSATION

Many readers of this book may have noticed that in the academic literature,
research conclusions are often so laced with disclaimers about causation that

one wonders whether it is possible to demonstrate causal links at all. Why would an evaluator on a limited budget even bother trying? The trick here is to understand two things. First, there are some major differences between the standards of proof being used by academics and what may be appropriate for us to use. Second, many of the methods used to address causation in the empirical literature are, quite frankly, pretty weak on the causal inference front.

Every profession has its own "dialect," including special terminology and rules regarding how to talk about things. For academics in the hard sciences and at least most of the social sciences, the norms dictate that even if researchers have evidence that makes them 99% sure that something is true, they still cannot say that they "know" or have "proved" it. Instead, the language is always framed in a cautious way, for example, "The evidence appears to suggest . . ." or "We found tentative support for . . ." This is in sharp contrast to the way in which we (and our clients) use terms such as *know* and *certain* in everyday conversation.

Organizational reality in for-profit, not-for-profit, and many government settings is that most decision makers would say that they *knew* something—and were prepared to make decisions on the basis of that knowledge—if they were, say, 70% or 80% certain based on the evidence. Of course, this varies a bit from setting to setting and from decision to decision, but most would agree that this sounds about right.

Our task as evaluators is to provide timely answers about the quality or value of products, programs, policies, and other evaluands, often to help people make decisions. These may be internal decisions about how to improve something or consumer decisions about which product to buy or which school to attend. Because each decision-making context requires a different level of certainty, it is important to be clear up front about the level of certainty required. Then, rather than throwing in the methodological kitchen sink or skipping the causal inference step, we will be in a much better position to strategically put together a blend of methods that will meet that certainty requirement (Davidson, 2003).

Some of the research in the academic literature tends to be somewhat lacking in evidence for making causal inferences. There are two reasons for this. One is that many researchers use wholly quantitative or wholly qualitative methods in their studies. The other reason is that in quantitative studies, researchers often lack the opportunity to use large samples, control groups, and random assignment. In such cases, quantitative methods alone tend to be

woefully inadequate for attributing causation, as are many all-qualitative designs. So, those disclaimers about causation that we see in such single-method (i.e., all-quantitative or all-qualitative) research are almost certainly justified. Moreover, they are attributable to shortcomings in the research design itself rather than to the impossibility of solving the problem.

INFERRING CAUSATION: BASIC PRINCIPLES

What are we trying to do when we infer causation? There are two basic principles here. First, look for evidence for and against the suspected cause (i.e., the evaluand). Second, look for evidence for and against any important alternative causes (i.e., rival explanations).

When considering whether the evaluand caused the observed changes, the evaluation team members need to consider what evidence, if present, would help to convince them that this was the case. Conversely, if such evidence were absent, to what extent would that convince them that the evaluand was proba-bly not the cause?

Equally important in causal analysis is the careful consideration of any and all important rival explanations for the observed changes. But how do we know which rival explanations are most important and how many we need to eliminate? That all depends on what level of certainty you need in your decision-making context. Sometimes you will need to eliminate only the "primary sus-pects," that is, the most likely alternative explanations. Sometimes you will need to rule out just about anything that anyone can suggest.

Probably the best way in which to approach this task is with a stepwise process. The first step is to put yourself in the shoes of the harshest critics you can imagine and think what objections they might raise to your claim that the evaluand caused a particular effect. Using a mix of strategies from the next section, gather enough evidence to confirm or rule out that rival explanation. Then consider what the next objection is likely to be. Repeat the process until all remaining alternative explanations are unlikely enough that they do not threaten your conclusions given the level of certainty needed to make them.

Usually, the more politically charged or controversial something is, the more likely it is that there will be opponents, many of whom will attack the methodology of the evaluation if they do not like the conclusions. And the harder people attack, the more solid your answers need to be. For this reason, the level of certainty required may change depending on what you uncover in the evaluation.

Even if a fairly high level of certainty is required, the trick is not to focus on a single "Rolls Royce" method for causal inference (e.g., an elaborate experimental design with multiple controls). Rather, you should use a strategic mix of methods that have different strengths and that *together* will give you enough evidence to be certain enough that the link is (or is not) causal. This principle (using methods with different strengths to complement each other) is called *critical multiplism* (Shadish, 1994).

INFERRING CAUSATION: EIGHT STRATEGIES

Some academics, among others, frequently say that the only way in which to infer causation is with the use of randomized **experimental designs.** This is sometimes met with a response from practitioners that such methods are simply not feasible in real-world settings. This is not true.

There is good news on both fronts for evaluators who need to know whether the changes they are seeing really are outcomes (i.e., changes attributable to the evaluand)—and that is all of us. The fact of the matter is that experimental designs (or at least **quasi-experimental designs**) are actually quite viable more often than we might expect. But even when they are not, there are several other practical strategies, some of which make use of some very powerful qualitative methodologies, that can be used to supplement or even replace the use of experimental and quasi-experimental designs.

The following subsections describe a range of methods for inferring causation, from very simple commonsense strategies to some more complex methods. For a small-scale evaluation, even some modest evidence about causation could prove to be sufficient. For more high-stakes evaluations, the evaluation team will need to draw on a range of methods to attain the level of certainty required.

Strategy 1: Ask Observers

Suppose that someone asked you to name the four or five most important factors that led to the development of your current professional skill set. Most

of us could easily identify which experiences were the most important and which ones had nearly no effect whatsoever. For the powerful learning experiences we have in our careers, there is no doubt in our minds that the experiences were primary *causes* of the learning. In many cases, we can also identify important contextual factors or we can point to a combination of experiences that culminated in a quantum leap in knowledge. And we can just as easily list a number of courses, books, conferences, and work assignments that added very little (i.e., where there was virtually no causal link).

It is amazing how "arm's length" we are as we look at the impacts of things on people's lives, especially in quantitative research. We gather pre- and posttest measures, and then use regression and other statistical tools to partial out the extraneous effects of this and that, without ever considering that perhaps we should start by just asking the question directly. In qualitative research, such evidence is perhaps more likely to be collected, but it is often not treated as explicit evidence of causation.

The "ask observers" strategy includes two possibilities. The first is to directly ask people who were supposedly affected by the evaluand (i.e., actual or potential impactees). The second possibility is to ask those who were in a position to observe the effects on impactees (e.g., coworkers, parents, teachers, trainers).

There are two ways in which to infer causation by just asking the people who were supposedly affected by the evaluand. One is to first gather some data about changes in outcome variables (e.g., reduced absenteeism, improved performance) and then to identify those people who experienced (a) little change, (b) some change, and (c) substantial change (positive or negative). In a follow-up interview or survey, the evaluation team members could ask, for example, "We noticed that you have had a substantial decrease in the number of times you were absent from or late to work during the past few months. Can you tell us a little about why that is?" The answer would tell you whether the individual believed that the evaluand was the primary cause or not and/or the extent to which other factors (e.g., contextual factors, other events) might also have contributed to the change. Note that the use of an open-ended question here allows respondents to list other causes that the evaluation team might not even have considered.

The other way in which to gather causal information from those directly affected by the evaluand is to actually work causation into the survey or interview questions themselves. So, instead of asking people to rate their level of

knowledge before and after completing a training or educational program, you might ask directly, "How much has your knowledge increased *as a result of* participating in this program?" (Include the italics in the survey item to make sure that respondents pay attention to it.) To probe other causes of knowledge gain, you might ask, "Did anything else besides the program increase your knowledge in this area over the same period of time?" To get at side effects, you might ask, "Please describe anything else that has happened to you or someone you know *as a result of* participating in this program." This way, you are not simply asking what has changed since before the program; instead, you are asking directly about the things that people know or believe were caused by the program.

Some researchers may argue that causation-rich questions such as these are leading, that is, that they implicitly direct the respondent to answer in a particular way (usually positive). It is true that we need to be careful about question wording when designing interview instruments or questionnaires, bearing in mind that in most cases, it is quite obvious what the evaluation team is trying to get at. But do not forget that these same questions, if well constructed, can also provide the opportunity for the respondent to say, "My knowledge of X increased during that time, but not because of that program." The arm's-length pre- and postquestionnaire that does not ask about causation eliminates the opportunity for people to even mention this.

A great example that incorporates both of these "just ask people" strategies just described is Brinkerhoff's (2003) Success Case Method. All participants in a particular program are given a 5-minute questionnaire on which they are asked whether or not they have been able to achieve enhanced performance as a result of the program and, if so, to give an example. Claims of dramatic improvement are then cross-checked against hard data to identify the true success cases, and a sample of these individuals are then interviewed in-depth to find out what it was that allowed them to get so much out of the program. In this case, the causation question is not just *whether* the program produced the effect but also *what other factors* enabled or inhibited the effect.

Some might argue that the individual might not be a reliable witness to help answer the causation question. In rare cases, this may be true. However, there are few evaluands that are so subtle in their effects that the recipient does not even notice their influence, so it seems remiss to exclude the views of the very people who likely saw things happen with their own eyes or experienced change directly. Of course, in most cases, other evidence will also be required to make justifiable causal inferences.

The "ask observers" method is not limited to those who were themselves changed. Often it is possible to identify people who observed a cause produce an effect in someone (or something) else. For example, parents of very young children can often directly observe the influence of particular experiences on their children (e.g., whether children mimic violent acts after watching a certain television show). A spouse might be in a good position to observe whether a violent offender's behavior was affected by a counseling session. Or an observer in a mathematics class might be able to see directly whether children learn faster and are more engaged when a new practical exercise is used to illustrate a concept.

Strategy 2: Check Whether the Content of the Evaluand Matches the Outcome

Here is another super simple commonsense strategy for inferring causation. Suppose that a treatment program for alcoholics taught participants several very specific strategies they could use to avert potential relapses. Also, suppose that participants in this program really did have very few relapses after completing the program. If the program were truly the cause of the lack of relapses, the evaluation team would expect to find that the alcoholics who avoided relapses used the strategies they had been taught in the treatment program rather than other strategies they knew previously or had picked up elsewhere. In other words, the content of the evaluand should quite often be reflected in some of the outcomes themselves if the evaluand did indeed cause the observed change.

When using this method, it is equally important to look for counterexamples. In this case, that means other strategies that were not learned in the program but that were used successfully to avert relapses. Where (or from whom) were these strategies learned? This information may point to one or more additional causes of alcoholics' success that were not attributable to the specific program. The existence of these additional causes does not negate the value of the program. However, if all potential relapses were prevented using strategies other than those taught in the program, and especially if relapses were not prevented in several cases where the taught strategies were used, this would call into question the value of the relapse avoidance strategies (and perhaps of the entire program).

Strategy 3: Look for Other Telltale Patterns
That Suggest One Cause or Another

In addition to looking to see whether early outcomes (e.g., the behavior changes just described) match the content of the evaluand, it is often possible to identify other telltale patterns that suggest a particular cause. These patterns, or "signature traces," are described by Scriven as the key to making causal inferences using the *modus operandi method.* This 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, or "clues," one can move up the causal chain, identifying what might have caused them.

In the opposite direction, one can start with the evaluand itself, or 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, 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.

The modus operandi method works best for evaluands that have highly distinctive patterns of effects. For example, a faith-based marriage counseling program, if effective, not only would result in partners using strategies taught within the sessions to improve their marriages but also would be likely to yield a telltale pattern of distinctive side effects. We might expect participants to report increased spiritual enlightenment and stronger connections with the relevant faith community. We might also expect to see less tolerance of attitudes and behaviors that are inconsistent with participants' faith. In contrast, improvements in marital relationships that were due not to the faith-based element but rather to the regular counseling would not be expected to yield such a pattern.

In some cases, there is not a great deal known about the patterns we should expect if a certain evaluand is likely to cause a particular effect. In such cases, it can be useful (albeit a weaker option) to draw an analogy with what is known about something similar. So, if the pattern observed closely resembles a known pattern in an analogous case, this can be interpreted as at least partial evidence for a causal link.

Let's use an example to illustrate the use of analogy as partial evidence of causation. Suppose that you had been asked to evaluate a cutting-edge intervention that helped teams of people to critically reflect on their work and generate new ways of doing things. Also, suppose that there was virtually no documentation about what happens when such interventions are successful. As an alternative, the evaluation team might dig for similar interventions that had not been used on teams. Previous research shows that when individuals are taught to critically reflect on their own work (e.g., in executive coaching), they can make transformational improvements in their own performance. The team learning intervention seeks to translate this idea for an interactive team setting. By examining the patterns in executive coaching success cases and seeing whether they are mirrored in the team intervention, it may be possible to use this as indirect evidence for a causal link by drawing an analogy with the individual-level version of the intervention. This evidence alone will not allow the evaluation team to make causal inferences, but it is certainly one additional piece of evidence to add to the pool.

Strategy 4: Check Whether the
Timing of Outcomes Makes Sense

In nearly all cases, an outcome should appear only at the same time as or after whatever caused it.[1] With distal outcomes in particular (i.e., those quite far downstream in the causal chain), the evaluation team should expect a considerable delay between the introduction of the evaluand and the appearance of outcomes. In general, the further downstream the outcomes, the longer they should take to appear.

For example, suppose that we were evaluating a community health intervention that focused on improving diet and exercise. We should probably expect to see the following:

- Fairly immediate knowledge and skill gain relating to the subject matter taught as part of the intervention (i.e., we should be able to detect this during and/or immediately after any health education component)
- A short delay (days to weeks) before the knowledge and skills are translated into changed behavior such as improved eating habits and exercise

- A moderate delay (weeks to months) before we could expect to see changes in individual health indicators, such as cholesterol, weight, and blood pressure, as a result of sustained behavior change
- A long delay (probably years) before these changes could be expected to have become widespread enough in the community to affect community-level health statistics such as the incidence of diabetes and heart disease and average life expectancy

Information about expected time frames for outcomes may be found in the relevant literature and from experts in the field. But in many cases, the evaluation team's logic might not be too far off target, so just taking the time to think through the timing issue will probably pay dividends.

There are three ways in which this information can be used to help confirm or disconfirm causal links. First, each identified outcome should be checked to ensure that it did not occur either before the evaluand was introduced or unrealistically quickly afterward. In fact, this is one good reason to check on some of those downstream outcomes at points in time when it should be too early to detect any change.

Second, outcomes should also be checked to see whether the timing of their appearance would be more (or equally) logical relative to other possible causes. For example, suppose that on-the-job performance improved following a well-executed training program that also coincided with the introduction of a performance-linked bonus system. In this case, the evaluation team would look at the timing of the improvements relative to the introduction of the two interventions to try to work out whether one or both of these (and/or something else) were likely to have been a substantial cause of the improvement.

The third strategy for using information about the timing of outcomes is to check whether outcomes further downstream in the logic model did not occur out of sequence, that is, before the outcomes that were expected to lead to them. In the earlier example of a community health program, if participant cholesterol and blood pressure dropped prior to any change in eating or exercising behavior, this makes it unlikely that the observed improvements in health indicators were caused by the program.

For those readers interested in exploring the timing of outcomes in more depth, Lipsey (1989) presents a very useful set of graphs that show different patterns of responses to interventions, including a delayed reaction and an initial response followed by a decay.

Strategy 5: Check Whether the "Dose" Is Related Logically to the "Response"

In the messy real world of evaluation, we are often faced with situations where an evaluand has been implemented inconsistently. For example, the author was once asked to evaluate the effectiveness of a new **management-by-objectives** (MBO) and reward system. A year after the system was rolled out organization-wide, it turned out that approximately a quarter of all staff still had no objectives in place and that the range in the quality of performance objectives was extremely variable across the organization for those who had objectives in place. Although the social scientists in us might throw our hands in the air in frustration in this kind of situation, the shrewd evaluators in us should instantly spot this as an excellent opportunity to check the causal link between the evaluand and its suspected effects.

The dose–response idea (i.e., if more A, then more B) comes from the medical metaphor of drug testing—the higher the dose, the greater the response should be (up to a point). For a performance management system such as the one just described, the more completely and effectively the system had been implemented in a particular work unit, the higher the "dose" (of MBO) for that unit and the greater the expected improvement in performance. If we found that performance had improved more dramatically in units where the system had been poorly implemented (or not implemented at all), this would be evidence that the new performance appraisal system was probably not the cause of the improvement.

When looking at the relationship between the "dose" of the evaluand and the "response" (magnitude of the outcome), it is important to bear in mind that this might not necessarily be a linear relationship. It is very common to have a "ceiling effect" where longer duration or more intensive exposure starts adding little or nothing in incremental value beyond a lower dose. Also, in many cases, there might be an "overdose" where excessive exposure or duration backfires and produces a less than optimal (or a very negative) result. As a simple example, schoolchildren will probably tolerate only so many hours per week of extracurricular reading before they develop a loathing for the activity.

An extension of the dose–response relationship is the situation where multiple doses are given and multiple responses are observed. Evidence for causation is strengthened if the evaluand is implemented in several different contexts and if the effect is observed every time (or nearly every time) the cause is introduced (i.e., when A, always B).

Strategy 6: Make Comparisons
With a "Control" or "Comparison" Group

The dichotomous (on/off) version of the dose–response relationship is the comparison between people who have been recipients of an evaluand and those who have not. This relationship forms the basis for the classic experimental design. In a fully randomized experimental design, participants would be randomly assigned to either a treatment group (receive the evaluand) or a control group (receive nothing or an alternative intervention). Provided that sampling is done carefully and that sample sizes are large enough, randomization helps to make sure that there are no systematic differences between the evaluand recipients and nonrecipients. It is rather like thoroughly shuffling a deck of cards to minimize the chance that one player gets all of the high cards.

In a quasi-experimental design, groups would not be randomly assigned, but the evaluation team would seek out a closely similar comparison group with which to compare results. Careful matching of treatment and comparison groups eliminates or greatly reduces the likelihood that rival explanations exist (e.g., the groups were different from the start). For example, studies of the effectiveness of the death penalty have compared crime rates in adjacent counties across state lines where one state introduces or abolishes the death penalty but the other state does not. Researchers carefully check to ensure that prior crime rates are similar and that the inhabitants of each county are similar demographically, socioeconomically, and in any other important respects to make sure that the comparison is reasonable to make.

Strategy 7: Control Statistically for Extraneous Variables

In the statistical analysis of data from experimental, quasi-experimental, and even single group (dose-response) designs, it is often possible to "control for" certain characteristics of the recipients and/or the contexts that are suspected of being correlated with the outcomes. This is particularly useful in cases where the evaluation team cannot be certain that the control or comparison group (if any) is truly similar in these respects.

For example, suppose that you were evaluating an innovative new method for teaching mathematics in a high school and that you had decided to use a comparison group of classes that were not exposed to the new technique. Even if you were able to randomly assign students to the classes that used and did not use the method, it might still be useful to make sure that prior aptitude in

math was not causing the results to look better or worse than they really were. The simple way in which to check this is to compare the treatment and control classes on prior math performance or scores on an aptitude or achievement test to ensure that there was no significant difference. But another more sophisti-cated strategy is to use a statistical technique called *regression analysis* to "partial out" the effect of prior aptitude in math so that any differences observed were not due to that factor. In this way, the evaluation team can statistically control for characteristics that might cloud the results.

Options and strategies available in the area of experimental, quasi-experimental, and related designs and associated data analysis are very numerous indeed. For some evaluations, these designs are essential. In such cases, if the evaluation team members do not have a specialist to help with the design, they would be well advised to find one. And in the meantime, there are many resources available to give the beginner a simple overview of the principles and enough know-how to design simple experimental studies.
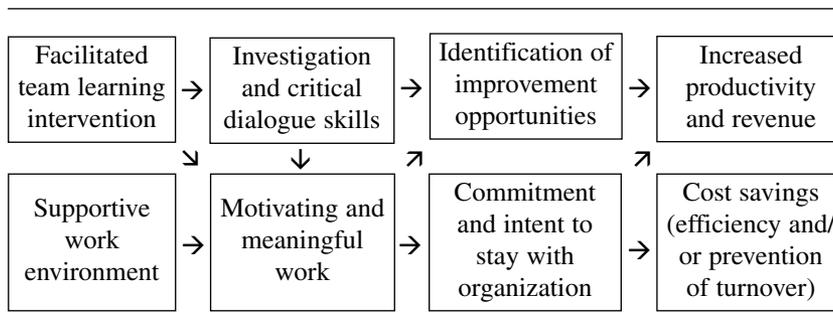
Strategy 8: Identify and Check the Underlying Causal Mechanism(s)

Another commonsense strategy we use a lot in everyday life is to look for an underlying mechanism that will help to make the case for causation more or less convincing. For example, the link between cigarette smoking and lung cancer was for years argued to be purely correlational. However, when research identified several substances known to be carcinogenic in cigarette smoke, it became more difficult to argue that there was not a causal link.

As a second example,[2] suppose that a team of consultants had been brought into an organization to facilitate a team learning intervention. The organization has shown an increase in profitability for the past quarter, and management wants to know whether this was due to the team learning intervention or to something else. How might the evaluation team use causal mechanisms to trace potential causal links?

The logic model in Exhibit 5.2 shows how this hypothetical team learning intervention would probably affect the bottom line. Evidence in favor of a causal link would include (a) an increase in investigation and critical dialogue skills during the intervention and (b) evidence that cost-saving improvements were identified or implemented during the intervention itself. Note that the logic model also includes the important contextual factor of a supportive work environment, which would be required for the success of the intervention.

Exhibit 5.2   Logic Model Connecting a Team Learning Intervention to the Bottom Line

Facilitated team learning intervention	→	Investigation and critical dialogue skills	→	Identification of improvement opportunities	→	Increased productivity and revenue
↘		↓		↗		↗
Supportive work environment	→	Motivating and meaningful work	→	Commitment and intent to stay with organization	→	Cost savings (efficiency and/ or prevention of turnover)

Evidence against a causal link would include (a) no evidence of improved investigation or critical dialogue skills, (b) no evidence that the intervention had a motivating effect (with most participants complaining that it was boring), and (c) most employees attributing their improved performance to the new incentive system rather than the team learning intervention.

Where would a logic model like this come from, and what would make it more or less useful as a source of evidence for causal inference? An evaluation team with knowledge of team learning interventions (and access to the relevant literature) would be able to create a model that is consistent with cutting-edge knowledge about team learning.

CHOOSING A BLEND OF STRATEGIES TO ADDRESS THE CAUSATION ISSUE

As mentioned earlier, there are times when one must build a virtually bullet-proof case for causation, whereas there are other times when such a high level of certainty is not required. Do you need *all* of the previously discussed evidence in hand to demonstrate causation in a particular case? Usually not. Again, it is prudent to put yourself in the shoes of a tough critic. Identify the most potentially threatening rival explanation and then choose the types of evidence that will most quickly and cost-effectively confirm or dispel that rival explanation. Bear in mind that your analysis could show that something else was in fact a major cause of the observed change(s). Make sure that you hunt specifically for evidence that would *confirm* such a rival explanation; do not

just look for evidence that would confirm what you hope to find. This hunt for disconfirming evidence as well as confirming evidence will make your conclusions stronger and more defensible.

The elimination of rival explanations is an iterative process and is one that can be greatly assisted by having a group of "devil's advocates" to help you. Once the first round is complete, identify the next most likely alternative explanation and repeat the process just described. Continue until you have amassed a body of evidence that provides you with enough certainty to draw causal inferences given the political pressure your findings will encounter as well as the decision-making or reporting context you face.

Does the evaluator need to show that there were absolutely no other influences affecting the bottom line at the time? Certainly not; there are always other influences at work in a complex system. The main issue is whether the intervention you are evaluating added a *practically significant impact* above and beyond whatever else was happening at the time that was large enough to justify its cost.

NOTES

1. There are rare exceptions to this, for example, when a premonition is caused by a future event. However, these cases rarely apply in professional evaluations.

2. This example was also used in an earlier article about linking organizational learning to the bottom line (Davidson, 2003).

ADDITIONAL READINGS

Entries in Scriven's (1991) *Evaluation Thesaurus:*
- Causation
- Constructionism
- Etiology
- Evaluability assessment
- Illuminative evaluation
- Naturalistic
- Positivism
- Postpositivism
- Program theory
- Quasi-experimental design

- Relativism
- Theory
- True experiment

Davidson, E. J. (2000). Ascertaining causality in theory-based evaluation. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities* (New Directions for Evaluation, No. 87, pp. 17–26). San Francisco: Jossey-Bass.

Huberman, A. M., & Miles, M. B. (1998). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Collecting and interpreting qualitative materials* (pp. 179–210). Thousand Oaks, CA: Sage.

Hume, D. (2000). An essay concerning human understanding. In *Essays and treatises on several subjects* (Vol. 2). Leeds, UK: Leeds Electronic Text Centre. (Original work published 1777)

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Scriven, M. (1974). Maximizing the power of causal investigations: The modus operandi method. In W. J. Popham (Ed.), *Evaluation in education: Current applications* (pp. 68–84). Berkeley, CA: McCutchan.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

EXERCISES

1. Suppose that you are conducting an evaluation of a training program for long-term unemployed individuals, defined as people who had been out of work for at least 2 years. As part of the same evaluation, you find that 50% of the participants got full-time jobs within 3 months of completing the program. This, of course, is good news. But then a cynical friend of yours points out that the local unemployment rate dropped over the same period of time, so that a general improvement in the job market could just as easily be the reason why these people found jobs.

   a. Which two complementary sources of evidence pertaining to causation would together provide the most powerful counterargument to your friend's claim at the lowest cost? Justify your choices (on three quarters of a page or less).

b. For each source of evidence, describe what you would expect to find if the program was the primary cause of the participants' finding jobs. What evidence would you expect to see if the general change in economic conditions was the primary cause?

2. For your own evaluand, list the top three rival explanations that might be suggested if you find evidence that needs were met. Lay out which causal inference strategies you will include in your evaluation design to make sure that you can check and rule out (or confirm) these rival explanations. (This information should be incorporated into your Methodology checkpoint.)