The Canadian Journal of Program Evaluation Vol. 15 No. 1 Pages 1–24
ISSN 0834-1516

ANSWERING THE *WHY* QUESTION IN EVALUATION: THE CAUSAL-MODEL APPROACH

Anthony Petrosino
American Academy of Arts & Sciences
Center for Evaluation, Initiatives for Children Program
Cambridge, Massachusetts USA

Abstract: Theorists and researchers have urged that evaluation test the causal assumptions about *why* an intervention should work to achieve its goals. Though the terminology is different, these writers agree that the assumed pathway from program to ultimate outcome be made explicit and tested empirically. The causal-model approach, because it is geared toward providing evidence on *why* a program worked or failed, is advocated whenever time, money, and experience permit. After defining casual-model evaluation, an example from the crime prevention literature is used to contrast the approach with traditional evaluations. Benefits, limitations, and other issues are also discussed.

Résumé: Les théoriciens et les chercheurs préconisent que l'évaluation met à l'épreuve les hypothèses causales concernant les *raisons* pour lesquelles une intervention devrait donner les résultats escomptés. Même si la terminologie est différente, les auteurs conviennent que la voie empruntée pour arriver aux résultats doit être explicite et soumise à un essai empirique. Le modèle causal, qui vise à fournir des données probantes sur les raisons du succès ou de l'échec d'un programme, est à privilégier lorsque le temps, l'argent, et l'expérience le permettent. Après la définition du modèle causal, un exemple tiré de la documentation sur la prévention du crime sert à mettre cette approche en contraste avec les évaluations traditionnelles. Les avantages, les limites, et d'autres questions sont aussi discutés.

Even when randomized field experiments are conducted — providing a fairly unambiguous answer to the question "did the program work?" — they are often inadequate in helping us to answer *why*. In this article, I discuss an orientation toward evaluation, often referred to as the theory-based or theory-driven approach, that is geared toward producing better answers to the *why* question. In these evaluations, the evaluator creates a model of the micro-

steps or linkages in the causal path from program to ultimate outcome — and empirically tests it (Weiss, 1995). For reasons that become clear later, I opt for the term "causal-model evaluation" to better describe my emphasis in this article. This is similar to the term "causal path" used by Cordray (1992).

Some examples of causal-model evaluation exist. Heflinger and her colleagues (Heflinger, Bickman, Northrup, & Sonnichsen, 1997) report a preliminary evaluation of the Family Empowerment Project. They conducted a randomized experiment to test the effects of a parent-training curriculum designed specifically for caretakers with children in the mental health system. Instead of comparing group success rates on outcomes for experimental and control participants, they articulated a causal model of how the program was assumed to work: (1) parent training would increase the parent's knowledge, self-efficacy, and advocacy skills; (2) parents would then become more involved in their child's mental health system; (3) parents would collaborate more effectively with their child's practitioners; and (4) this would result in improved mental health outcomes for children. Since it was a preliminary report, they only tested the effects of the treatment on knowledge and self-efficacy (finding large treatment effects), but found no useful measures for testing advocacy skills.

Although theorists have written persuasively about the need for evaluations like the Heflinger et al. study over the past three decades (Chen & Rossi, 1992; Schon, 1997; Weiss, 1998), actual practice has not followed suit (Weiss, 1997a). Social program evaluation continues to be criticized for a preoccupation with methodology (Chen & Rossi, 1992; Pawson & Tilley, 1994). Instead of articulating a model of how a program might achieve its effects and using that to guide the evaluation, the traditional study has attempted to address the "did it work?" question with as much precision as possible. In other words, evaluators often start with methods rather than models. Though method is just as critical, the pursuit of rigour should not be to the exclusion of answering the *why* question.

Petrosino's (1997) meta-analysis of 150 crime reduction experiments is instructive in this regard. In 142 of the experiments, the evaluation did not attempt to test a model of how the program would work to reduce subsequent criminality. Crucial mechanisms for change were not identified and tested. As Glaser (1994) noted about offender treatment evaluations in general, prison literacy programs designed to reduce recidivism often provide no data to determine if literacy

itself was impacted. No clues were provided to indicate whether the theory or model underlying the program worked the expected way or whether change occurred through some other yet unidentified pathway. These findings correspond to an earlier survey of the literature by Lipsey (1988), in which he found that less than 10% of social program evaluations involved "integrated theory."

Most of the literature on this approach is composed of wisdom pieces rather than actual examples. Several reasons, however, indicate that this approach will increase in use. First, though still rare, recent conferences attest to the number of evaluations in progress that are testing explicit causal models (American Evaluation Association, 1998). Second, funding agencies are beginning to require such evaluations as a condition for support (e.g., National Institute on Alcohol Abuse and Alcoholism, 1998). Third, articles on the approach are beginning to make their way out of specialized evaluation journals into more discipline-specific outlets, exposing other researchers to this perspective.

In this article, I present an overview of causal-model evaluation. Though the approach has been discussed in relation to planning, implementation, measurement, and other key aspects of evaluation practice, I limit my discussion to outcome or impact studies.

HISTORY AND LEXICON

I stand upon the shoulders of several evaluation giants in writing this article. A few of them have been recommending evaluations that test causal models for a long time, some as long as 30 years, to guide social program evaluation.

Weiss was one of the first theorists to bring attention to the need for testing the underlying assumptions about why a program should work in evaluation. In her article on the use of evaluation results for policy decisions, she stated that "utilization might be increased if the evaluation included such elements as … the explication of the theoretical premises underlying the program and direction of the evaluation to analysis of these premises" (1972a, p. 323).

In her seminal primer on evaluation practice, Weiss talks about creating a model of the program's causal processes and testing it in the evaluation (1972b, pp. 50–51). She also describes two types of variables that should be examined in evaluation. Program operation vari-

ables are those that describe how the intervention operates (e.g., frequency, duration, and quality of contact); bridging variables are those that link the program operation variables to desired outcomes. Weiss (1972b) uses what she calls a "process model" to describe how these variables would be integrated and serve as candidates for testing in the evaluation.

This early work — as well as others that discussed similar approaches to evaluation (Suchman, 1967) — received little attention. Since the discussion was imbedded in a larger evaluation text and the approach was not yet assigned a catchy phrase, it is little wonder that it remained less than influential.

Chen and Rossi (1980, 1983, 1987, 1992) deserve the lion's share of the credit for popularizing the causal-model approach. Beginning in 1980, in a series of articles and books, they describe a similar perspective they call theory-driven evaluation. In their earlier writings, they emphasized the role of the evaluator in forming program theory based on existing social science and cautioned against relying on practitioners and policy-makers' assumptions, since those are often incorrect and primitive (Chen & Rossi, 1992). Chen later seems to acknowledge that the assumptions of program planners and staff can be used — preferably in collaboration with insights from the social sciences — to create a coherent causal theory of how the intervention should work (Chen, 1990a).[1]

Weiss (1995, 1997a) later renamed her approach "theory-based" evaluation and seemed to allow more for testing the common-sense assumptions of staff and policy makers than Chen and Rossi — even if the common-sense assumptions contradict social science (Patton, 1997). Weiss (1998) also places more emphasis on linking specific program activities to specific mechanisms. Similar to her 1972 discrimination between program operation and bridging variables, she distinguishes between implementation and program theory. Implementation theory involves the activities or tasks the program is going to accomplish, while program theory encompasses the mechanisms for change that the activities or tasks will bring about. Both implementation theory and program theory comprise the program's overall "theory of change."

I should mention that other similar approaches are advocated in the evaluation literature. Adding to some confusion, different terms (that do not include the word theory) are used to describe them.[2]

Though the lexicon used by each of these researchers is different, they share a major concern that evaluators articulate and test the causal model and key mechanisms for change that the program puts into action. These include Program Development Evaluation (PDE)[3] developed by Gottfredson (1984), Realist Evaluation advocated by Pawson and Tilley (1994), and Theory of Action Evaluation promoted by Schon (1997).[4] Utilization-Focused Evaluation articulated by Patton (1997) also includes a lengthy discussion of program theory.[5]

WHY CAUSAL-MODEL INSTEAD OF THEORY?

A theory is a plausible explanation about reality that must now be tested (Wilkins, 1964). Following suit, a program theory represents a plausible explanation of how the intervention will work to affect important outcomes (Bickman, 1987). Theory-based or theory-driven evaluation then appear to be suitable terms.

But the use of the word "theory" in the lexicon is misleading. First, all evaluations test some theory (that A will change B), but the problem is that they do not test the important links in the causal pathway (how will A change B). Second, some are calling any evaluation with theoretical implications for larger social science theory "theory-based" or "theory-driven" even though evaluation theorists would not agree (Winfree, Esbensen, & Osgood, 1996). Third, some are stretching what is meant by these terms to the point at which every evaluation that goes beyond the "black box" is labeled theoretically driven (Chen & Rossi, 1992). Fourth, black-box evaluations that test the effects of interventions that have good empirical underpinnings are being labeled theory-based. Fifth, the word theory is used in at least four major approaches (theory-based, theory-driven, theory-of-action in PDE, Schon's [1997] theory-of-action) and what constitutes a theory in each one is described a bit differently.

I opt for the term causal-model evaluation for two reasons. First, it drops theory from the lexicon and sidesteps some of the confusion associated with using that word. Chen and Rossi's (1992) description of causative or causal theory is closest, but it still implies scientific theory (an implication that Chen and Rossi seem to embrace). Second, it accurately describes what most evaluators and theorists mean: *the evaluation is testing the causal model of how the program hopes to achieve its effects*. This model can be research-based or it can simply be a linkage of the common-sense assumptions of the policy maker and practitioner (Petrosino & Petrosino, 1999). The

causal model can include the examination of program components or implementation factors as in Weiss's (1998) theory-based evaluation, but it is not essential to do so to meet the minimum definition.

DEFINITION OF A CAUSAL-MODEL EVALUATION

A causal-model evaluation is one that: (a) establishes an a priori model of how the program will work to affect important outcomes before the evaluation is completed; and (b) empirically tests at least one link in the model apart from implementation or program activity data, along with outcome data.[6] I believe this definition can help eliminate a problem that Chen and Rossi noted about the theory-driven approach, when they stated that "the framework has been stretched to include some practices that we believe are not appropriate to be considered theory-driven" (1992, p. 1).

Though the definition mandates that the evaluation test a specified causal model, it does not imply that the model is rigidly fixed. It can be worked and reworked based on new information from various sources. For example, Bickman (1996) initially created a preliminary model of how a mental health program, The Fort Bragg Project, would work to achieve outcomes. He later modified it after interviews and inspection of program documents. When the evaluation is ready to begin, however, the evaluator must decide on the model that will be tested. This will guide design, data collection, and other decisions for the evaluation.

My definition here limits causal-model evaluations to those that test explicit models and rules out studies with theoretical implications alone. The Mobilization of Youth study in New York City was a test of an intervention based on Cloward and Ohlin's (1960) differential opportunity theory, but did not include the testing of important links from program operation to desired outcome and cannot be described as a causal-model evaluation (Short, 1975).[7] Likewise, Sherman and Berk's (1984) seminal experiment that tested the effects of arrest on misdemeanour domestic violence offenders was certainly theoretically important (it tested specific deterrence theory), but it did not use the causal-model approach. No attempt to articulate and test the underlying assumptions about why arrest should deter subsequent violence was made.

One of the recommendations in the literature is that interventions be theory-based, that is, developed in accordance with established

research (Fagan, 1996). A research-based model that describes how the program should work is sometimes articulated by investigators, but, unless the evaluation follows suit and tests several of the links in the model, it does not fit the above definition. Without an empirical test of critical links, we cannot know if the data support the assumed model. As Weiss (1997a) has noted, even successful programs can have faulty models; the program may achieve the desired effect, but not for the reasons articulated in the model.

The definition also distinguishes process-outcome studies from causal-model evaluations. Process studies provide data on whether the program was implemented with fidelity, and whether key activities took place as planned. In ambitious studies, the evaluators will examine the relationship of process variables to outcomes. But these process variables represent tasks or activities or components of the program — not the mechanisms for change. Changes occur as people, places, or things respond to the program's activities; it is these mechanisms that causal-model evaluation hopes to test (Weiss, 1999).

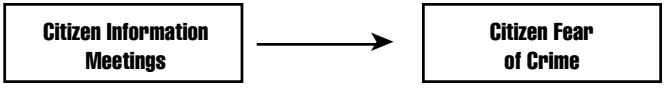
In a similar vein, mechanisms for change and intermediate goals should be distinguished. An intermediate goal is not a mechanism if it is simply a milestone in the path to the ultimate outcome. For example, success at the first follow-up period might be a prerequisite for success at later time intervals, but it is not why participants succeed or fail. But a variable can serve both purposes. For example, in Chandler's (1973) evaluation of a juvenile delinquency program, the mechanism for change (egocentrism) could also be considered an intermediate goal, since it was measured immediately at the end of the treatment period.

THE ONE-STEP VERSUS CAUSAL-MODEL EVALUATION

As Chen and Rossi (1992) have indicated, describing traditional evaluations as "atheoretical" is inaccurate. Every evaluation tests some implicit model, but it is usually a very simplistic one, what we might call a one-step evaluation (Petrosino, 1998). In such studies, the evaluation tests the impact of a program — generally treated as a monolithic whole or black box — on some outcome measure of interest (e.g., recidivism rates).

Figure 1 presents a hypothetical example of the one-step model evaluation, designed to test the impact of a periodic police-community

Figure 1
The One-Step Model



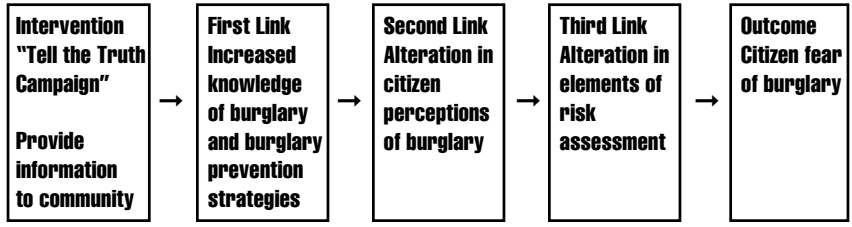
meeting on citizen fear of crime. Evaluation theorists would question whether the one-step study captures the complex set of processes that the intervention would have to work through in order to have an effect on citizen fear. The meetings must do something to citizens who attend them to reduce their anxiety about crime. A causal-model evaluation would make explicit those assumptions that connect attendance at the meetings to a reduction in citizen fear of crime.

Such an evaluation was recently conducted in Denmark (Kuttschreuter & Wiegman, 1997). The evaluators tested a program known as the "To Tell the Truth Campaign." The rationale behind the intervention was that citizens are more afraid of burglary than their actual risk because of excessive media attention. The intervention consisted of holding periodic community meetings, in which police would provide citizens with actual burglary statistics for their neighbourhood and would also empower them with crime prevention information. Through this intervention, citizen fear — now based on actual statistics rather than misperception — would be reduced.

Rather than test the one-step model, the authors articulated a more complex set of linkages in the causal chain from intervention to effect. As Figure 2 shows, the model assumed that the meetings would increase citizen knowledge about burglary and prevention strategies; this in turn would alter their perceptions about burglary in the neighbourhood. Changes in perception of burglary would change the citizenry's risk assessment (lowering it for the most part), leading to a reduced level of fear.

The investigators found, however, that the information campaign did not reduce citizen fear of burglary. In one-step model evaluations, that would be all we would have learned. In this study, we learn more about *why*. The program failed, even though knowledge was increased, because it was ultimately unable to alter the perceptions of citizens about the amount of burglary in the community. As

Figure 2
A Causal-Model Evaluation of the "To Tell the Truth Campaign"



a result, risk assessment was unchanged, as was citizen fear. The model broke down following the first link.

It should be noted that the investigators did not forsake internal validity. Kuttschreuter and Wiegman implemented a quasi-experimental design and collected additional data to rule out plausible internal validity threats. They also collected data to insure that program failure was not due to faulty implementation or process.

EXTENSIONS TO THE CAUSAL-MODEL APPROACH

Though any evaluation meeting the two criteria offered earlier is a causal-model study, these are minimum inclusion criteria. Nothing about the definition excludes evaluators from considering other critical program issues.

Investigators may wish to articulate and test more than one model of how the intervention will work. Pawson and Tilley (1994), in their persuasive article on realist evaluation, describe nine potential mechanisms for how closed circuit television (CCTV) might work to reduce auto-related crime in a parking lot. Each might be considered a different model of how the program was supposed to work. In a retrospective study that utilized available administrative data, Pawson and Tilley (1994) demonstrated how some models could be tested to determine their plausibility in interpreting findings. The retrospective nature of their inquiry prevented them from testing all of them — since the data were unavailable — underscoring the importance of prospective causal modeling. Another contribution of the Pawson and Tilley (1994) study was their emphasis on under-

standing the contextual factors that might enhance or decrease the success of an intervention like CCTV.

An evaluation can also go beyond an overall program model to consider the causal mechanisms for each specific component. For example, Wong-Rieger and David (1993) articulated and tested a model of how a workshop for HIV-positive men might work to motivate ongoing support relationships. Using path analysis to test their model, Wong-Rieger and David (1993) examined the strength of particular workshop activities (e.g., safe sex and risk reduction information, encouragement to recognize own needs for support, etc.) on presumed mechanisms for change (e.g., awareness of need) and the ultimate outcome of seeking or providing support.

Evaluators can also go beyond causal models to examine unintended consequences or what Weiss calls "negative program theory" (Weiss, 1998). In other words, what can we expect to happen if things go wrong? Pallone (1990) provides a case of why this is important. A prison program, based on a theory that increased assertiveness would help sex offenders in their relationships with adult women, was introduced into the facility. Though the intervention did improve offender assertiveness, it also resulted in an increase in institutional misconducts and aggression toward prison guards. Along with testing a causal-model of how the program should work, evaluators can include some data checkpoints to insure that the intervention is not backfiring on other dimensions.

DESIGNS, RESEARCH METHODS, AND MEASURES

Since the causal-model approach is an orientation to conducting evaluation rather than a method, it is perfectly compatible with experimental or non-experimental designs. Despite criticisms of the randomized experiment as comprising black-box or atheoretical research, some of the best causal-model evaluations have used controlled trial designs. I think it is wise — given the infrequency of both randomized field experiments and causal-model evaluations — to combine the rigour of the design and the logic of the approach in the same study.

But not all evaluations can have randomized control groups. Can the causal-model approach still be used? Weiss (1995) suggested that such an approach might be used in evaluating Comprehensive Community Initiatives (CCI) — a broad infusion of multiple interven-

tions into a targeted neighbourhood under the direction of a collaborative group of agencies, funders, and local stakeholders — and might compensate for internal validity when there is no adequate comparison area (Weiss, 1995). In other words, if a causal model of a series of micro-steps or links was validated in the evaluation, would that not be strong evidence that the CCI caused the observed effects? This idea is controversial and not without many critics (Cook, 1999), but it drew attention because of the intractable problems of developing good comparisons for evaluating CCIs (Weiss, 1997b).

Causal-model evaluations can be conducted using a wide range of research methodologies and data sources to test the underlying assumptions or links in the model. Generally, naturalistic studies that attempt to make sense of evaluation settings are more likely to be used to generate useful models rather than test them. But Maxwell (1996) has argued persuasively that qualitative methods — and the data they generate — can be very rigorous in examining causal attribution. Likewise, I make no distinction between qualitative or quantitative measures and believe both can be useful in either developing or testing causal models.

WHEN CAUSAL-MODEL EVALUATIONS ARE NOT NEEDED

The more direct the effect of an intervention on its target, the less important are causal-model evaluations. For an obvious example, the effect of handcuffs on restraining arrested offenders is direct; the handcuff does not work through some psychological or attitudinal variable to keep arrestees from using their hands in an attacking manner. Handcuffs will restrain in similar fashion across settings and sites. The effect of jail or prison on incapacitating offenders from committing crimes in the community is also direct; it is the act of confinement that constricts their free movement and prevents their access to vulnerable community members or property.

But most social science interventions are not like handcuffs or incapacitating facilities. Instead, they exert indirect effects (Donaldson, forthcoming). The program, policy or practice does something to persons, places, or things to change them. Usually, the intervention is delivered in the hope that behaviour will change after the program is over. What is a drug prevention program trying to change in its clientele to make them more resilient to drug use? What is a prison treatment program trying to change to help participants adjust to community life? If the desired effect of a program on an outcome

measure of interest is indirect, then the causal-model approach should be considered.

Scriven (1998) has attacked this approach (the theory-driven version) on the grounds of its impracticality and irrelevance in some instances. Let me offer a response. First, I do think causal-model tests should be done when possible, because of our obligation as scientists to society and science to contribute to the knowledge base.[8] In the end, what we learn from a single evaluation may not help much. But what we learn from a number of causal-model tests, in different settings, could help us significantly. I do not mean, however, that causal-model evaluations should be done when there is not enough time, money, or experience to carry them out, nor do I believe that Chen, Rossi, Weiss, or other proponents of this approach would argue that practical matters should be dismissed.

HOW COMPLEX SHOULD MODELS BE?

Weiss (1995) has argued that formulating and testing a model of how the program works (in her theory-based evaluation) could help evaluators when conducting studies in which no adequate comparison groups can be found. Her basic premise is that the less rigorous our evaluation is in terms of internal validity, the more links in the model we should test. A non-experimental evaluation should increase a test the number of links in the model.

But Lipsey's (1997) point is well-taken. So few causal-model evaluations are done that even a simple two-step model (involving only one mechanism or link and the final outcome) would represent considerable advancement in certain areas. This would be especially valuable for randomized and quasi-experimental studies, since: (a) rigorous designs usually have stable interventions that are more focused than CCIs and lend themselves to relatively simple causal models; (b) their internal validity rigour is high to permit one to trust the observed effects at each link; and (c) they currently (and almost exclusively) comprise the studies used in systematic reviews and meta-analyses.

BENEFITS OF THE CAUSAL-MODEL APPROACH

The greatest advantage of the causal-model approach is that it provides empirical data on *why* the intervention worked or failed. In most evaluations, we usually work retrospectively from the results

to suggest theories or reasons why a certain effect was observed for the intervention. With no data, our post-hoc theorizing remains untested. The causal-model approach uses the reverse strategy: start with a model or program theory and use the evaluation to test it. Weiss (1995), Chen (1990b), and Bickman (1987) among others have suggested several other potential benefits of using this strategy. I summarize some of these below.

Providing Guidance for Evaluation Decisions

The evaluator is faced with a myriad of decisions at the start of the study. An explicit causal model should provide evaluators with guidance on the important measures to select, the data to collect, and the type of design that would be most appropriate (Weiss, 1998).

Planning

As Weiss (1995) writes, forcing the practitioners, program designers, and other stakeholders to make explicit the underlying causal assumptions is a valuable process in its own right. It can serve a formative purpose, if done early on in the program, helping the staff to improve operations. For example, if evaluators and different stakeholders collaboratively discuss their underlying assumptions, faulty thinking about why the program should work might be exposed. If some aspect of the intervention was hastily built on the assumption, it might still be possible to correct the program before it is up and running full speed.

Reducing Resistance to Evaluation

As Chen (1990a) and Huebner (1998) have noted, the process of sitting down with stakeholders to work out a model of how the program will work has other benefits. Chen writes that the process can "enhance communication between stakeholders and evaluators with respect to their views and consensus in designing a useful evaluation" (1990a, p. 16). Huebner (1998) noted how evaluators were able to overcome barriers of resistance to the study by administrators and teachers by working collaboratively on the program model.

Program Replication and Diffusion

Hacsi (1998) discussed how causal-model evaluations can help policy makers and program planners replicate and diffuse successful in-

novations. Not knowing why a program works does not mean it cannot be used. Understanding the critical mechanisms at work should make diffusing successful programs less chancy. As Bickman (1987) noted, it is not the program that we wish to replicate in other settings (since it will likely be composed of different staff, funding, motivation, and clients), but the underlying causal theory (that this program will work to reduce crime or improve student learning or increase employment).

Suggesting a Future Research Agenda

As the evaluation by Heflinger et al. (1997) showed, articulating a causal model and trying to test it are two different things. They found that a critical link in their model (advocacy skills acquisition) could not be reliably measured given the paucity of instrument development in this area. But pointing out past research deficiencies and suggesting a future agenda is still an important contribution to social science.

Providing Data for Systematic Reviewing

The greatest potential for knowledge contribution may come when systematic reviewing techniques — including meta-analysis when appropriate — can be used to analyze the results of a collection of causal-model studies in a particular intervention area (e.g., juvenile delinquency treatment, school-based violence prevention, etc.). Causal-model evaluations, if reported with full integrity, will provide more data for systematic reviewers to use in their analyses. Meta-analytic techniques, in particular, might give us strong clues about which mechanisms are most important in effective programs (Lipsey, 1997; Petrosino, 1998; Weiss, 1995). Data, however, that are not collected and reported cannot be used in later reviews. By using the causal-model approach, the evaluation is oriented toward collecting, testing, and reporting critical data on mechanisms that can later be examined in systematic reviews. International organizations recently created to prepare, maintain, and make accessible systematic reviews of research on effects of intervention — such as the newly formed Campbell Collaboration for the social sciences and the successful Cochrane Collaboration for health care — could exploit this information wisely.[9]

BARRIERS AND LIMITATIONS

Evaluation sages have long acknowledged the barriers and limitations to conducting causal-model evaluation (Bickman, 1987; Scriven, 1998; Weiss, 1997b). These include some of the following.

Money and Time

Causal-model evaluations increase the amount of data that must be collected and analyzed. As a result, the money spent on evaluation will need to be increased. Evaluators will also have to spend more time in the setting, sometimes working closely with program staff and other stakeholders, to develop a coherent model that can be tested in the study.

Bottom Line Focus

In many instances, evaluators may be the only ones who care about the underlying causal assumptions. Policy makers and other stakeholders may only be interested in final outcome results and unwilling to pay for the evaluator's desire to contribute to the larger social science knowledge base at the client's expense.

Endless "Why?" Questions

As Rosenthal (1998) suggested, like the small inquiring toddler, we can always ask the *why* question of a program. For example, using the Denmark causal-model evaluation results, why did the program fail to alter citizen perceptions of burglary? No matter how detailed our model and results, someone can always ask *why* a certain effect was observed. No evaluation can hope to answer every possible question (Weiss, 1995).

Failure to Specify Effects

Bickman (1998) notes that most program theories or models fail to specify what an acceptable effect is for each link in the model. Using the randomized experiment as an example, models do not specify if and how the treatment group will differ from controls and when that difference will emerge. Sometimes the important effect is not whether a significant difference between the groups at the first model

link is found, but if the experimental group reached a certain level, regardless of how control participants did. This needs to be better specified in our models.

Problems of Mechanism Measurement

As Weiss (1997b) noted, social science has just begun to emphasize and begin testing the role of mechanisms for change. Consequently, measures that operationalize the important concepts still need to be created, tested, and validated. In some areas, adequate measures are not available nor systems set up to collect desired data.

Over-simplicity of Models

Rogers (1998) has pointed out that evaluators using this approach often use simple, linear models. These models presume that variables act and react like dominoes, i.e., link one will change resulting in a change at link two, resulting in a change at link three and so on. The real world, as Rogers (1998) notes, might not work that way at all. Instead, feedback mechanisms, looping, external influences, and other factors not accounted for in the model might also be operating.

CONCLUSION

Given some sizable limitations, the reader might wonder how causal-model evaluations get done at all, let alone how we might increase their number. I suggest one possible remedy. We need to restructure the way that funding for evaluation is carried out at the federal and state level (Petrosino, 1998). As Sherman and his colleagues write about criminal justice (Sherman et al., 1997), the requirement that all grant recipients do an evaluation has resulted in almost none of them getting done. My interviews with evaluation and program managers in government agencies underscore a core theme: *at the end of each fiscal year, despite our millions invested, little has been learned – either about what works or why.*

Instead of asking every grant recipient to do an evaluation on skimpy funds and getting back what we paid for (mostly input data on clients served), we should pool our evaluation monies together. Rigorous, causal-model evaluations should be supported for about 10% of programs (Petrosino, 1998). If done in each state — and at the fed-

eral level — we would amass a significant body of knowledge in a relatively short time. Systematic reviewing techniques — including meta-analysis when appropriate — could then arm us with far more information on programs and their effects than is currently available (Petrosino, 1998).

A second point is the need for social science to keep better track of its evaluations. What good is the extra effort and cost in doing causal-model evaluations if we cannot learn from them, collectively, as they are reported over time?[10] Whether they are published or not, we should by now have a register or centralized catalogue of evaluations, including causal-model studies, that test the effects of some social program, policy, or practice. We can haggle over how well-controlled the evaluations need to be to get into such a register. But the medical field, with its Cochrane Collaboration and electronic library of over 218,000 clinical trials (as well as a database of systematic reviews), has provided us with a nice example in practice. An international group of researchers has initiated such an effort with randomized trials, but these represent a fraction of the evaluative studies conducted (Petrosino, Boruch, Rounding, McDonald, & Chalmers, forthcoming).

Third, the time is ripe for a few studies that empirically test the value added by the causal-model approach. Huebner (1998) has taken a nice first step by describing the experiences of evaluators who conducted several causal-model (she uses the term theory-based) studies of educational interventions; as mentioned earlier, they found that the process helped to break down resistance to the evaluation and helped to build a healthier, non-adversarial relationship between evaluators, administrators, and teachers. I think we also need to examine whether causal-model evaluations are utilized more in subsequent policy, program, or practice decisions than traditional or process-outcome studies.

The causal-model approach could advance social programming more than our current evaluation practice. We would not only learn whether something worked, but also more about *why*. As our knowledge of underlying mechanisms becomes more established, our ability to diffuse successful justice interventions should also be enhanced (Hacsi, 1998; Pawson & Tilley, 1994).

ACKNOWLEDGEMENTS

This article was supported by a Spencer Foundation fellowship in "evaluating programs for children" at the Harvard Children's Initiative, Harvard University. All comments, however, are the responsibility solely of the author and do not represent the Spencer Foundation, Harvard University, or any other institution. The article was sparked and sharpened by many enjoyable, informal discussions with my talented 1997–98 co-fellows: Tim Hacsi, Tracy Huebner, and Patricia Rogers. Tim and Patricia also provided nice feedback on this particular draft. I especially thank Carol Hirschon Weiss for her inspiration in things evaluative and my fellowship mentor and friend, Frederick Mosteller, for the patient guidance.

NOTES

1. Social science is still at the roots of their later discussions. In fact, Chen and Rossi define program theory as "substantive knowledge that is action-oriented, preferably grounded in previous research, concerning a program and the social problem the program attempts to alleviate" (1992, p. 2).

2. I would like to see some standardization of the lexicon. The word "theory" raises too many problems for my comfort, particularly when trying to communicate with persons unfamiliar with evaluation scholarship or jargon. But "model" also seems problematic, given its different uses in social science (Weiss, 1997b). "Causal path" was used by David Cordray of Vanderbilt (1992), and causal model evaluation seems to me to be the best fit, but I remain open to any term with potential to weave together the different jargon used to describe all-too-similar approaches.

3. PDE requires schools and organizations to specify their theories of action on which to base programs and define measurable objectives based on such theories (Gottfredson, 1984).

4. Schon (1997) distinguished his approach from theory-based evaluation in two ways. First, Schon claimed the theory of action evaluation was setting-specific; evaluators should use the approach to learn and adapt practice in the same setting and not try to generalize knowledge learned to other sites. Second, Schon discriminates between three different levels of theory: (a) espoused theory, the explanations that justify the program at the policy level; (b) design theory, how the program is planned before implementation; and (c)

theory in use, how the program is actually carried out.

5. This makes it difficult to track down actual studies that have used a theory-based approach, since some that would fit adopt a different evaluation terminology (e.g., PDE) or do not claim the use of any approach (e.g., Kuttschreuter & Wiegman, 1997).

6. This definition grew out of many conversations with my co-fellows, as well as an engaging seminar with Carol Hirschon Weiss in May, 1998.

7. Short (1975) notes that social science theory may be too general or vague in translating to a social action program. How do we change opportunity structures for minority youth?

8. In examining the code of ethics for leading social science associations, I have not found this obligation mentioned specifically. I believe, nonetheless, that it should be considered an obligation.

9. Information on the Cochrane Collaboration is most easily found at its website (www.cochrane.org). Its sibling organization, the Campbell Collaboration, was recently inaugurated in the UK in July 1999 and will focus on syntheses of research on the effects of social and educational intervention (Davies & Petrosino, 1999).

10. I should also note the difficulty in learning from causal-model evaluations, since the model might be described in one article, the implementation in another, and the actual impact study in still another. Some of the relevant reports may not have been published in accessible outlets, making it logistically more difficult to learn about the evaluation.

REFERENCES

American Evaluation Association. (1998). *Annual meeting program*. Available on-line at <http://www.eval.org>.

Bickman, L. (1987). The functions of program theory. In L. Bickman (Ed.), *Using program theory in evaluation* (pp. 5–18). San Francisco: Jossey-Bass.

Bickman, L. (1996). The evaluation of a children's mental health managed care demonstration. *Journal of Mental Health Administration, 23,* 7–15.

Bickman, L. (1998, November 5). Discussant's comments, Panel on theory-based evaluation, American Evaluation Association annual meeting, Chicago.

Chandler, M.J. (1973). Egocentrism and antisocial behavior: The assessment and training of social perspective-taking skills. *Developmental Psychology, 9*(3), 326–332.

Chen, H.T. (1990a). Issues in constructing program theory. In L. Bickman (Ed.), *Advances in program theory* (pp. 7–17). San Francisco: Jossey-Bass.

Chen, H.T. (1990b). *Theory-driven evaluations*. Newbury Park, CA: Sage.

Chen, H.T., & Rossi, P. (1980). The multi-grade, theory-driven approach to evaluation: A model linking basic and applied social science. *Social Forces, 59*, 106–122.

Chen, H.T., & Rossi, P. (1983). Evaluation with sense: The theory-driven approach. *Evaluation Review, 7*, 283–302.

Chen, H.T., & Rossi, P. (1987). The theory-driven approach to validity. *Evaluation and Program Planning, 10*, 95–103.

Chen, H.T., & Rossi, P. (Eds.). (1992). *Using theory to improve program and policy evaluations*. New York: Greenwood.

Cloward, R.A., & Ohlin, L.E. (1960). *Delinquency and opportunity: A theory of delinquent gangs*. Glencoe, IL: Free Press.

Cook, T.D. (1999, May 12). Considering the major arguments against random assignment: An analysis of the intellectual culture surrounding evaluation in American schools of education. Paper presented to the American Academy of Arts & Sciences, Cambridge, MA.

Cordray, D. (1992). In H.T. Chen & P. Rossi (Eds)., *Using theory to improve program and policy evaluations*. New York: Greenwood.

Davies, P., & Petrosino, A. (1999). *Proceedings of the international meeting on systematic reviews of the effects of social and educational interventions. July 15-16, University College-London*. London: University College-London, School of Public Policy.

Donaldson, S.I. (forthcoming). Mediator and moderator analysis in program development. In *Handbook of program development in health behavior research and practice*.

Fagan, J.O. (1996). *The criminalization of domestic violence: Promises and limits*. NIJ Research Report. Washington, DC: Department of Justice.

Glaser, D.A. (1994). What works, and why it is important: A response to Logan and Gaes. *Justice Quarterly, 11*(4), 711–723.

Gottfredson, G.D. (1984). A theory-ridden approach to program evaluation: A method for stimulating researcher-implementer collaboration. *American Psychologist, 39*(10), 1101–1112.

Hacsi, T. (1998, November 5). Using theory-based evaluation to replicate successful programs. Presentation at the American Evaluation Association annual meeting. Chicago.

Heflinger, C.A., Bickman, L., Northrup, D., & Sonnichsen, S. (1997). A theory-driven intervention and evaluation to explore family caregiver empowerment. *Journal of Emotional and Behavioral Disorders, 5*(3), 184–191.

Huebner, T. (1998, November 5). Theory-based evaluation: Gaining a shared understanding between staff and evaluators. Paper presented at the American Evaluation Association annual meeting. Chicago.

Kuttschreuter, M., & Wiegman, O. (1997). Crime communication at information meetings. *British Journal of Criminology, 37*(1), 46–62.

Lipsey, M.W. (1988). Practice and malpractice in evaluation research. *Evaluation Practice, 9*(4), 5–24.

Lipsey, M.W. (1997). What can you build with thousands of bricks? Musings on the cumulation of knowledge in program evaluation. *New Directions for Evaluation, 76*, 7–23.

Maxwell, J. (1996). *Using qualitative research to develop causal explanations*. Cambridge, MA: Harvard Project on Schooling & Children.

National Institute on Alcohol Abuse and Alcoholism. (1998). *Treatment for adolescent alcohol abuse and alcoholism (AA-98-003)/NOA-NIH*. Betheseda, MD: Author.

Pallone, N.J. (1990). *Rehabilitating criminal sexual psychopaths*. New Brunswick, NJ: Transaction.

Patton, M.Q. (1997). *Utilization-focused evaluation* (3rd ed.). Thousand Oaks, CA: Sage.

Pawson, R., & Tilley, N. (1994). What works in evaluation research? *British Journal of Criminology, 34*(3), 291–309.

Petrosino, A. (1997). *What works? Revisited again: A meta-analysis of randomized experiments in rehabilitation, deterrence and prevention*. PhD dissertation. Ann Arbor, MI: University Microfilms.

Petrosino, A. (1998, November 5). Building better information: The potential mutual benefits of theory-based evaluation and meta-analysis. Paper presented at the American Evaluation Association annual meeting. Chicago.

Petrosino, A., Boruch, R.F., Rounding, C., McDonald, S., & Chalmers, I. (forthcoming, July 2000). Assembling a social, psychological, educational and criminological trials register. *Evaluation Research in Education*.

Petrosino, A., & Petrosino, C. (1999). The public safety potential of Megan's Law in Massachusetts: An assessment using a sample of criminal sexual psychopaths. *Crime and Delinquency, 40*(1), 140–158.

Rogers, P.J. (1998, November 5). Alternative causal models in program theory evaluation and monitoring. Paper presented at the American Evaluation Association annual meeting. Chicago.

Rosenthal, R. (1998). Personal communication.

Schon, D. (1997, April). Notes for a theory-of-action approach to evaluation. Paper presented to the Harvard Project on Schooling & Children Evaluation Task Force.

Scriven, M. (1998, November 5). Comments, Presidential Strand Debate, Evaluation in opposition: Black box versus theory-driven evaluation. American Evaluation Association annual meeting. Chicago.

Sherman, L.W., Gottfredson, D., MacKenzie, D., Eck, J., Reuter, P., & Bushway, S. (1997). *Preventing crime: What works, what doesn't,*

what's promising. A report to the United States Congress. College Park, MD: University of Maryland, Department of Criminology and Criminal Justice.

Sherman, L.W., & Berk, R.A. (1984). The deterrent effects of arrest for domestic assault. *American Sociological Review, 49*, 261–272.

Short, J.F. (1975). The natural history of an applied theory: Differential opportunity and "Mobilization for Youth." In N.J. Demerath, O. Larsen, & K. Schuessler (Eds.), *Social policy and sociology* (pp. 193–210). New York: Academic Press.

Suchman, E. (1967). *Evaluative research*. New York: Russell Sage Foundation.

Weiss, C.H. (1972a). Utilization of evaluation: Toward comparative study. In C.H. Weiss (Ed.), *Evaluating action programs: Readings in social action and education* (pp. 318–326). Boston: Allyn & Bacon.

Weiss, C.H. (1972b). *Evaluation research: Methods of assessing program effectiveness*. Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C.H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J.P. Connell, A.C. Kubisch, L.B. Schorr, & C.H. Weiss (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts* (pp. 65–92). Aspen, CO: Aspen Foundation.

Weiss, C.H. (1997a). How can theory-based evaluation make greater headway? *Evaluation Review, 21*, 501–524.

Weiss, C.H. (1997b). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation, 76,* 41–55.

Weiss, C.H. (1998). *Evaluation*. Englewood Cliffs, NJ: Prentice-Hall.

Weiss, C.H. (1999, May 12). *What to do until the random assigner comes*. Paper presented to the American Academy of Arts & Sciences. Cambridge, MA.

Wilkins, L.T. (1964). *Social deviance: Social policy, action, and research*. London: Tavistock.

Winfree, L.T., Esbensen, F., & Osgood, D.W. (1996). Evaluating a school-based gang-prevention program: A theoretical perspective. *Evaluation Review, 20*(2), 181.

Wong-Rieger, D., & David, L. (1993). Causal evaluation of impact of support workshop for HIV+ men. *Canadian Journal of Public Health, 84* (Supplement 1), 66–70.