1

The historical development *of* program theory evaluation,
current variations in theory and practice, and pressing
issues are discussed.

Program Theory Evaluation: Practice, Promise, and Problems

Patricia]. Rogers, Anthony *Petrosino,* Tracy **A.** *Huebner,*
Timothy A. Hacsi

For over thirty years now, many evaluators have recommended making explicit the underlying assumptions about how programs are expected to work—the program theory—and then using this theory to guide the evaluation. In this chapter, we provide an overview of program theory evaluation (PTE), based on our search to find what was value-added to evaluations that used this approach. We found fewer clear-cut, full-blown examples in practice than expected, but we found many interesting variations of PTE in practice and much to recommend it. And elements of PTE, whether the evaluators use the terminology or not, are being used in a wide range of areas of concern to evaluators. Based on this review, in this chapter we discuss the practice, promise, and problems of PTE.

What Is Program Theory Evaluation?

Because this volume is intended **to** demonstrate the diversity of practice, we have used a broad definition of program theory evaluation. We consider it to have two essential components, one conceptual and one empirical. PTE consists of an explicit theory or model of how the program causes the intended or observed outcomes and an evaluation that is at least partly guided by this model. This definition, though deliberately broad, does exclude some versions of evaluation that have the word theory attached to them. It does not cover all six types of theory-driven evaluation defined by Chen (1990) but only the type he refers to as intervening mechanism evaluation. It does not include evaluations that explicate the theory behind a program but that do not use the theory to guide the evaluation. Nor does it include evaluations in which the program theory is a list of activities, like a "to do" list, rather than a model showing a series of intermediate outcomes, or mechanisms, by which the program activities are understood to lead to the desired ends.

The idea of basing program evaluation on a causal model of the program is not a new one. At least as far back as the 1960s, Suchman suggested that program evaluation might address the achievement of a "chain of objectives" (1967, p. 55) and argued for the benefit of doing this. "The evaluation study tests some hypothesis that activity A will attain objective B because it is able to influence process C which affects the occurrence of this objective, An understanding of all three factors—program, objective and intervening process—is essential to the conduct of evaluative research" (1967, p. 177).

Weiss (1972) went on to explain how an evaluation could identify several possible causal models of a teacher home-visiting program and could determine which was the best as supported by evidence. In the three decades since, many different terms have been used for this type of evaluation, including *outcomes hierarchies* (Bennett, 1975) and *theory-of-action* (Schön, 1997). More commonly, the terms *program theory* (Bickman, 1987, 1990), *theory-based evaluation* (Weiss, 1995, 1997), and *program logic* (Lenne and Cleland, 1987; Funnell, 1997) have been used.

Unfortunately, although there are clear variations in types of PTE, these different labels have not been used consistently to refer to different types and have instead tended to reflect the preferred label in a particular organization or source references. Even though this volume uses the term program theory evaluation in its title, some of the authors use other terms.

Interest in program theory has grown significantly since two previous *New Directions* volumes on the topic (Bickman, 1987, 1990). More agencies and organizations, both in the United States and abroad, are at least paying lip service to program theory. Federal research funders such as the National Institutes of Health now require discussions of program theory in applications submitted for evaluation support. Many not-for-profit agencies have followed the United Way's lead in developing performance mesures based on a generic causal model of inputs-processes-outputs-outcomes (Hatry, van Houten, Plantz, and Greenway, 1996). Arguments for including program theory in evaluation are now appearing not only in evaluation journals but also in discipline-specific journals, such as those in education, criminology, and sociology. The largest-selling evaluation textbook, *Program Evaluation: A Systematic Approach,* has now, in its sixth edition, added a chapter on this approach (Rossi, Freeman, and Lipsey, 1999). Similarly, *Evaluation Models: Evaluation of Educational and Social Programs* (Madaus, Stufflebeam, and Scriven, 1983) has added a chapter on program theory evaluation in its second edition (Rogers, forthcoming).

Practice: Diverse Choices to Meet Diverse Needs

Program theory is know by many different names, created in many different ways, and used for any number of purposes. Here we provide a brief road map to the variety of ways people think about and employ program theory.

Locating Examples. To try to understand the variety of ways in which program theory evaluation is now being used, we began in early 1998 to comb through available bibliographical databases, citation indexes, and evaluation reports. We also reviewed conference proceedings, dissertations, and articles from a variety of disciplines. In addition, we received many helpful examples in response to an inquiry to the American Evaluation Association's Internet discussion list, EVALTALK. Our efforts turned up examples dating from 1957 to 2000 from the United States, Canada, Australia, New Zealand, and the United Kingdom. We have not included every example that we located in this volume but instead have used examples to identify and illustrate critical challenges in using program theory or ways of addressing them.

Our review showed amazing diversity in theory and practice across two main areas—how program theories are developed and how they are used to guide evaluations.

Developing the Program Theory—Who, When, and What. In some evaluations, the program theory has been developed largely by the evaluator, based on a review of research literature on similar programs or relevant causal mechanisms, through discussions with key informants, through a review of program documentation, or through observation of the program itself (Lipsey and Pollard, 1989). In other evaluations, the program theory has been developed primarily by those associated with the program, often through a group process. Many practitioners advise using a combination of these approaches (Pawson and Tilley, 1995; Patton, 1996; see also Funnell, Chapter Nine).

The program theory can be developed before the program is implemented or after the program is under way. At times, it is used to change program practice as the evaluation is beginning. Most program theories are summarized in a diagram showing a causal chain. Among the many variations, we will highlight just three for now; Rogers discusses other variations in Chapter Five.

At its simplest, a program theory shows a single intermediate outcome by which the program achieves its ultimate outcome. For example, in a program designed to reduce substance abuse, we might test whether or not the program succeeds in changing knowledge about possible dangers and then whether or not this seems important in achieving the desired behavior change. As Petrosino (Chapter Six) points out, for some program areas, articulating this mediating variable and measuring it would be a significant advance on current practice.

More complex program theories show a series of intermediate outcomes, sometimes in multiple strands that combine to cause the ultimate outcomes. So for a substance abuse prevention program, we might theorize that an effective program will generate a positive reaction among participants, change both attitudes and knowledge, and develop participants' skills in resisting peer pressure. Although these more complex program theories may more adequately represent the complexity of programs, it is impossible to design an evaluation that adequately covers all the factors they identify. Weiss (Chapter Four) proposes some ways to select the particular causal links that any one evaluation might study.

The third type of program theory is represented by a series of boxes labeled *inputs, processes, outputs,* and *outcomes,* with arrows connecting them. It is not specified which processes lead to which outputs. Instead the different components of a program theory are simply listed in each box. Although this type of program theory does not show the relationships among different components, these relationships are sometimes explored in the empirical component of the evaluation.

Using the Program Theory to Guide the Evaluation. Program theory has been used in quite different ways to guide evaluation. Examples show diversity in the purpose and audience of the evaluation, the type of research design, and the type of data collected. Within this diversity, it is possible to identify two broad clusters of practice.

In some PTEs, the main purpose of the evaluation is to test the program theory, to identify what it is about the program that causes the outcomes. This sort of PTE is most commonly used in large, well-resourced evaluations focused on such summative questions as, Does this program work? and Should this pilot be extended? These theory-testing PTEs wrestle with the issue of causal attribution—sometimes using experimental or quasi-experimental designs in conjunction with program theory and sometimes using program theory as an alternative to these designs. Such evaluations can be particularly helpful in distinguishing between theory failure and implementation failure (Lipsey, 1993; Weiss, 1997). By identifying and measuring the intermediate steps of program implementation and the initial impacts, we can begin to answer these questions. These intermediate outcomes also provide some interim measure of program success for programs with long-term intended outcomes.

An example of this type of program theory evaluation can be found in the Family Empowerment Project evaluation, in which Bickman and colleagues (1998) conducted an experimental test of the effects of a program that trained parents to be stronger advocates for children in the mental health system. They articulated a model of how the program was assumed to work. First, parent training would increase the parent's knowledge, self-efficacy, and advocacy skills. Second, parents would then become more involved in their child's mental health care. Finally, this collaboration would lead to the child's improved mental health outcomes.

But they did not stop with the articulation of a program theory. They also constructed measures, collected data, and analyzed them to test these underlying assumptions. The program was able to achieve statistically significant effects on parental knowledge and self-efficacy, but no useful measures for testing advocacy skills could be found. Unfortunately, the intervention had no apparent effect on caregiver involvement in treatment or service use and ultimately had no impact on the eventual mental health status of the children.

Evaluations such as these seem to be at least implicitly based on Weiss's definition of program theory, "[It] refers to the *mechanisms* that mediate between the delivery (and receipt) of the program and the emergence of the outcomes of interest" (1998, p. 57).

The other type of program theory evaluation is often seen in small evaluations done at the project level by or on behalf of project managers and staff. In these cases, program theory is more likely to be used for formative evaluation, to guide their daily actions and decisions, than for summative evaluation. Such PTEs are often not concerned with causal attribution. Although some of these evaluations pay attention to the influence of external factors, there is rarely systematic ruling out of rival explanations for the outcomes. Many of these evaluations have been developed in response to the increasing demands for programs and agencies to report performance information and to demonstrate their use of evaluation to improve their services. In these circumstances, PTE has often been highly regarded because of the benefits it provides to program managers and staff in terms of improved planning and management, in addition to its use as an evaluation tool.

Stewart, Cotton, Duckett, and Meleady (1990) provide an example of this type of PTE in their evaluation of a project that recruited and trained volunteers to provide emotional support for people with AIDS, their lovers, families, and friends. The paper did not provide a diagram of the program theory model nor present any data. Instead Stewart and colleagues reported on the process of developing the model, the types of data that were gathered, and how the data were used. "Performance indicators developed by Ankali [the project] were both for the organisation's own purposes and to meet the requirements of the funding body. Both qualitative and quantitative indicators were selected. . . . Ankali now uses the outcomes hierarchy during orientation of volunteers and report[s] that the process has assisted with improved targeting of volunteers and referral agencies, modification to the training program, supervision of clients and volunteers, and development of proposals for expansion and enhancement of the service" (1990, p. 317).

This type of program theory evaluation appears to be closer to that described by Wholey. "[It] identifies program resources, program activities, and intended program outcomes, and specifies a chain of causal assumptions linking program resources, activities, intermediate outcomes and ultimate program goals" (1987, p. 78).

Despite the apparent popularity of program theory evaluation, we found that the formal evaluation literature still has comparatively few examples. For instance, when we searched the abstracts in six bibliographical databases for the time frame 1995–1999, we found program theory explicitly mentioned in evaluations of children's programs only twice. In addition, many of the evaluations that we found used theory in very limited and specific ways, for example, to help plan an evaluation, but very few used theory as extensively as the most prominent proponents of this approach suggest. But PTEs conducted in small projects or local sites are rarely published in refereed journals or distributed widely, being more likely to be presented as conference papers by practitioners or presented in performance measurement forums. And many of them fail to include what some would consider an essential component of a program theory evaluation—systematic testing of the causal model.

In this volume, we include examples of both types of PTE. Weiss (Chapter Four), Hacsi (Chapter Seven), and Petrosino (Chapter Six) discuss issues associated with *theory-testing PTEs*. Huebner (Chapter Eight) discusses four examples of *action-guiding PTEs*, and Funnell (Chapter Nine) discusses a technique for assisting with this sort of PTE.

Promises and Problems

Program theory has been seen as an answer to many different problems in evaluation. Here we briefly discuss several areas where program theory has been seen as promising.

Understanding Why Programs Do or Do Not Work. Among the promises made for PTE, the most tantalizing is that it provides some clues to answer the question of why programs work or fail to work. Consider the usual practice of trying to understand why a program succeeded or failed. Following reporting of results, evaluators usually work in a post hoc manner to suggest reasons for observed results (Petrosino, forthcoming, 2000). But without data, such post hoc theories are never tested, and given the poor state of replication in the social sciences, they are likely never to be.

In contrast, by creating a model of the microsteps or linkages in the causal path from program to ultimate outcome—and empirically testing it—PTE provides something more about why the program failed or succeeded in reaching the distal goals it had hoped to achieve, as in Bickman and colleagues' evaluation of the family empowerment program (1998). Perhaps the intervention was not able to improve advocacy skills—remember, those could not be measured. Or maybe there was a critical mechanism missing from the model, which the program was not activating or engaging. We learn something more than the program's apparent lack of impact on children's mental health.

Even if all these issues cannot be adequately addressed in the original evaluation, a PTE can provide an agenda for the next program and evaluation.

For example, a critical link in the Bickman and colleagues study was not tested (advocacy skills acquisition), given the paucity of measurement development in this area. Pointing out this deficiency suggests an agenda to develop an instrument to measure this variable in the next similar study.

Attributing Outcomes to the Program. Another promise sometimes made for PTE is better evidence for causal attribution—to answer the question of whether the program caused the observed outcomes. Program theory has been used by evaluators to develop better evidence for attributing outcomes to a program in circumstances where random assignment is not possible (for example, Homel, 1990, in an evaluation of random breath testing of automobile drivers). In the absence of a counterfactual, support for causal attribution can come from evidence of achievement of intermediate outcomes, investigation of alternative explanations for outcomes, and pattern matching. Support for causal attribution can also come from program stakeholder assessments (for example, Funnell and Mograby, 1995, in their evaluation of the impact of program evaluations in a road and traffic authority) or from data about a range of indicators, including data on external factors likely to influence the theorized causal pathway (for example, Ward, Maine, McCarthy, and Kamara, 1994, in their evaluation of activities to reduce maternal mortality in developing countries). It may be possible to develop testable hypotheses on the basis of the causal model (Pawson and Tilley, 1995), especially if the model includes contingencies or differentiation—expected differences in outcomes depending on differences in context. Causal attribution is also sometimes addressed by combining traditional experimental or quasi-experimental designs with PTE.

Many PTEs do not address attribution at all, simply reporting implementation of activities and achievement of intended outcomes. This approach is particularly common where program theory is used to develop ongoing monitoring and performance information systems. Causal attribution in PTEs is discussed in more detail in the chapters in this volume by Cook (Chapter Three), Davidson (Chapter Two), and Hacsi (Chapter Seven).

Improving the Program. Many of the claims for the benefits of PTE refer to its capacity to improve programs directly and indirectly. Articulating a program theory can expose faulty thinking about why the program should work, which can be corrected before things are up and running at full speed (Weiss, 1995). The process of developing a program theory can itself be a rewarding experience, as staff develop common understanding of their work and identify the most important components. Many accounts of PTE (such as Milne, 1993; and Huebner, Chapter Eight) report that this has been the most positive benefit from conducting PTE. In this way, PTE is very similar to the earlier technique of evaluability assessment.

But PTE is supposed to then use the program theory to guide the evaluation, and it is here that some evaluations falter. Whereas collaboratively building a program theory can be an energizing team activity, exposing this to harsh empirical tests can be less attractive. Practical difficulties abound as well. When PTE is implemented at a small project, staff may not have the time or skills to collect and analyze data in ways that either test the program theory or provide useful information to guide decisions and action. If program theory is used to develop accountability systems, there is a real risk of goal displacement, wherein staff seek to achieve targets and stated objectives at the cost of achieving the ultimate goal or sustainability of the program (Winston, 1991).

Conclusion

In this chapter, we have outlined the range of activity that can be considered program theory evaluation and have identified major issues in its theory and practice. These are discussed in more detail by the other chapters in this volume.

References

Bennett, C. "Up the Hierarchy." *Journal of Extension,* 1975, *13*(2), 7–12.

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

Bickman, L. (ed.). Advances in Program Theory. New Directions for Program Evaluation, no. 47. San Francisco: Jossey-Bass, 1990.

Bickman, L., and others. "Long-Term Outcomes to Family Caregiver Empowerment." Journal of Child and Family Studies, 1998, 7(3), 269–282.

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

Funnell, S. "Program Logic: **An** Adaptable Tool." Evaluation News and Comment, 1997, 6(1), 5–17.

Funnell, S., and Mograby, **A.** "Evaluating Employment Programs Delivered by Community Organisations." Proceedings of the Annual Conference of the Australasian Evaluation Society, 1995, 2, 531–552.

Hatry, H., van Houten, T., Plantz, M. C., and Greenway, M. T. Measuring Program Outcomes: **A** Practical Approach. Alexandria, Va.: United Way of America, 1996.

Homel, R. "Random Breath Testing in New South Wales: The Evaluation **of** a Successful Social Experiment." National Evaluation Conference 1990, Proceedings, vol. 1. Australasian Evaluation Society, 1990.

Lenne, B., and Cleland, H. "Describing Program Logic." Program Evaluation Bulletin 1987, no. 2. Public Service Board of New South Wales, 1987.

Lipsey, M. W. "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.

Lipsey, M. W., and Pollard, J. "Driving Toward Theory in Program Evaluation: More Models to Choose From." Evaluation *and* Program Planning, 1989, 12, 317–328.

Madaus, G., Stufflebeam, D., and Scriven, M. Evaluation Models: Evaluation of Educational and Social Programs. Norwell, Mass.: Kluwer, 1983.

Milne, C. "Outcomes Hierarchies and Program Logic as Conceptual Tools: Five Case Studies." Paper presented at the international conference of the Australasian Evaluation Society, Brisbane, 1993.

Patton, M. Q. *Utilization-Focused* Evaluation. (3rd. ed.) Thousand Oaks, Calif.: Sage, 1996.

Pawson, R., and Tilley, N. *Realistic* Evaluation. London: Sage, 1995.

Petrosino, A. J. "Answering the Why Question in Evaluation: The Causal-Model Approach." Canadian Journal of Program Evaluation, 2000, *15*(1), 1–24.

Rogers, P. J. "Program Theory Evaluation: Not Whether Programs Work But Why." In G. Madaus, D. Stufflebeam, and T. Kelleher (eds.), Evaluation Models: Evaluation of *Educational* and Social Programs. Norwell, Mass.: Kluwer, forthcoming.

Rossi, P. H., Freeman, H., and Lipsey, M. W. Program Evaluation: A Systematic Approach. Thousand Oaks, Calif.: Sage, 1999.

Schon, D. A. "Theory-of-Action Evaluation." Paper presented to the Harvard Evaluation **Task** Force, Apr. 1997.

Stewart, K., Cotton, R., Duckett, M., and Meleady, K. "The New South Wales Program Logic Model: The Experience of the AIDS Bureau, New South Wales Department of Health." Proceedings *of* the Annual Conference of the Australasian Evaluation Society, 1990, *2*, 315–322.

Suchman, E. **A.** Evaluative Research: Principles **and** Practice in Public Service *and* Social *Action* Programs. New York: Russell Sage Foundation, 1967.

Ward, V. M., Maine, D., McCarthy, J., and Kamara, **A.** **"A** Strategy for the Evaluation ***of*** Activities to Reduce Mortality in Developing Countries." Evaluation Review, 1994, 18, 438–457.

Weiss, C. H. Evaluation Research: Methods *of* Assessing Program *Effectiveness.* Englewood Cliffs, N.J.: Prentice Hall, 1972.

Weiss, C. H. "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, and C. H. Weiss (eds.), New Approaches to Evaluating Community Initiatives: Concepts, *Methods* and *Contexts.* Washington, D.C.: Aspen Institute, 1995.

Weiss, C. H. "How Can Theory-Based Evaluation **Make** Greater Headway?" Evaluation Review, 1997, *21*, 501–524.

Weiss, C. H. Evaluation: Methodsfor Studying Programs and Policies. (2nd ed.) Englewood Cliffs, N.J.: Prentice Hall, 1998.

Wholey, J. S. "Evaluability Assessment: Developing Program Theory." In **L.** Bickman (ed.), Using Program Theory in *Evaluation.* New Directions for Program Evaluation, no. ***33.*** San Francisco: Jossey-Bass, 1987.

Winston, J. **A.** "Linking Evaluation and Performance Management.'' Paper presented at the annual conference of the Australasian Evaluation Society, Adelaide, 1991.

PATRICIA J. ROGERS *is* director of the Program for Public Sector Evaluation *in the* Faculty *of Applied Science,* Royal Melbourne Institute of Technology, *Australia.*

ANTHONY PETROSINO is research *fellow* at *the Center for* Evaluation, *Initiatives* for Children Program, American Academy of **Arts** *and* Sciences, and research associate at the *Harvard* Graduate School of Education.

*TRACY **A.*** HUEBNER is coordinator for comprehensive school *reform* at *WestEd.*

TIMOTHY **A.** *HACSI* is research fellow at **the** *Harvard* Children's *Initiative* and teaches history at the Harvard Extension School.