6

Even simple program theory evaluations could be used in meta-analysis to provide *benefit.* Several examples of ways to combine the two are explored, including a hypothetical *model* and an actual example of a federal funding program.

Whether and Why? The Potential Benefits of Including Program Theory Evaluation in Meta-Analysis

Anthony *Petrosino*

Determining whether and why a program has worked is risky business. A randomized experiment—if implemented and conducted with full integrity—can provide the least ambiguous answer to the question, Did the program work? But unless additional data, besides the outcome measures, are planned for and analyzed, an experiment can rarely provide an answer to the question, Why did the intervention work?[1] Even if we could successfully design a study that answers both whether and why questions, no small feat in itself, there is another problem. Single, stand-alone evaluations will rarely be definitive. Results from evaluations of the same program often vary across settings because of differences in clients, staff, and so on (Lipsey, 1997).

Two different perspectives, however, have emerged that together may provide better answers to whether and why questions. The first is program theory For at least thirty years, evaluators have written persuasively—and often—about the need to explicitly test program theory in evaluation (for example, Bickman, 1987; Chen and Rossi, 1992; Weiss, 1972). Although the lexicon vanes across

Note: Earlier drafts of this chapter were presented to the Harvard Children's Initiative Evaluation **Task** Force (April 1998) and the American Evaluation Association (November 1998). A Spencer Foundation fellowship at the Harvard Children's Initiative and a Mellon Foundation grant to the Center for Evaluation facilitated this work. I appreciate the comments of Mary Askew, Anne Barten, Leonard Bickman, Iain Chalmers, Jodi Delibertis, Timothy Hacsi, Tracy Huebner, Mark Lipsey, Heather McMillan, Frederick Mosteller, Pamela Perry, Carolyn Petrosino, Patricia Rogers, Sean Riordan, Robert Rosenthal, Carol Weiss, and Stuart Yeh. The thoughts expressed herein, however, are those of the author only.

these writers, all agree that evaluations, whenever possible, should articulate and test the underlying assumptions about why the program should "work." For consistency with other chapters in this volume, I adopt the term *program theory evaluation* (PTE) to describe studies using terminology such as *theory-based* or *theory-driven* to describe similar approaches.[2]

The interest in PTE runs parallel to the advent of a science of reviewing. Although attention to the problem of interpreting separate but similar studies dates back to 1904, the science of reviewing took hold in the 1970s with *meta-analysis* (Hunt, 1997). Because it was recognized that traditional methods for synthesizing research were flawed, meta-analysis surfaced as a rigorous method for summarizing the results of prior research (for example, Lipsey, 1990). In contrast to the use of PTE, the use of meta-analysis has been widespread; Lipsey and Wilson (1993) were able to identify over three hundred meta-analyses of social and educational treatment.

Writers such as Cordray (1992) and Lipsey (1997) suggest that the combination of PTE and meta-analysis could have benefits. For example, Lipsey showed how they could be used mutually to build social intervention theories. In this chapter, I build on these earlier arguments to demonstrate how the accumulation of knowledge from PTEs through meta-analysis could provide beneficial data for social policy and practice decisions. Hypothetical illustrations are relied on, as good examples of synthesizing PTEs in meta-analysis have yet to be reported.

The One-Step Evaluation Model

Many evaluations test only for an intervention's effects on outcome measures—sometimes referred to as one-step models (Weiss, 1997). The problem with the one-step model is that it does not explain why a program should affect the outcome (Chen and Rossi, 1992). It does not address the causal complexity involved in many programs that target outcomes such as criminal behavior. Some social interventions work through indirect processes: a treatment is delivered in one setting and is expected to engage other critical mechanisms in order to affect the outcome in a different setting (Donaldson, forthcoming). For example, school-based drug prevention targets adolescent drug use, presumably by engaging mechanisms such as peer resistance. These mechanisms are supposed to increase resilience to using drugs both inside and outside the schoolyard. The one-step model would ignore the measurement of underlying mechanisms such as peer resistance and would focus only on outcomes such as drug use.

How One-Step Evaluations Affect Meta-Analysis

The choice to conduct one-step evaluations has ramifications for research synthesis, as the original studies provide the data for meta-analysis. It is important for evaluators to see that how they conduct and report their studies because how they do so influences subsequent reviews.

Table 6.1. Hypothetical Meta-Analysis of One-Step Evaluations, Sex Offender Treatment

Program (N)	*Effect on Sex Offense Recidivism*	*BESD*
Cognitive-based (11)	.30	+15.00%
Behavioral (10)	.11	+5.50%
Individualized (12)	.01	+.05%
Group counseling (8)	–.05	–2.50%

Note: Ranked by effect size.

In most meta-analyses, reviewers create an effect size to express the impact of the program on the outcome measure of interest. For example, in an Illinois evaluation, ninety-four juvenile delinquents were randomly assigned to attend a prison tour; sixty-seven youngsters received no contact at all (Greater Egypt Regional Planning & Development Commission, 1979). The program also consisted of an interactive rap session with inmates, who provided the juveniles with realistic stories of prison life, including rape and murder. The "Scared Straight" type program was designed to deter juvenile delinquents from further delinquency. At six months, the evaluators found that 17 percent of the experimental group had been rearrested, compared with 12 percent of the controls, for an effect size of -14.[3] A negative sign is used because the program effect was in the opposite direction anticipated.

In most meta-analyses, effect sizes like these are averaged across all studies (for example, "programs to reduce delinquency had an average effect size of .10"). In some meta-analyses, they are averaged for specific treatments, using broad labels (for example, "vocational programs to reduce delinquency had an average effect size of .05"). Such a label serves two purposes: it provides a title to describe the basic intervention, but it is broad enough to capture more than just a few studies. These categories have implications for meta-analysis. For example, cognitive therapy, behavioral therapy, individualized counseling, and group counseling are all treatments for sex offenders (Laws, 1989). If evaluations of these treatments have been reported, meta-analysis can proceed and effect sizes computed. Most meta-analysts set as a criterion for eligibility that evaluations include either a control or comparison group.

Table 6.1 provides hypothetical results for a meta-analysis of sex offender treatment studies. The table indicates that the effect size for cognitive treatment was .30. Would such a finding be important? Rosenthal and Rubin (1982) provide a way of translating effect sizes into percentage differences, known as the binomial effect size display (BESD). Using BESD, an effect size of .30 translates into a 15 percent improvement for the cognitive group compared with controls. An average 15 percent reduction in sex offender recidivism would be important.[4]

But the broad label of cognitive treatment masks useful information about the programs in that category. There are many different types of cognitive programs for sex offenders, with great variation in how even a single type of program is implemented across sites. Persons charged with responsibility to fund or implement cognitive programs, however, might not find the results in Table 6.1 useful to their decision making. A number of programs are considered cognitive, and the broad label does not indicate which of the cognitive programs they should employ.

How PTEs Could Be Exploited in Meta-Analysis

One way to overcome this problem is to increase the number of PTEs eligible for meta-analysis. The end of this chapter provides one practical suggestion on how to do this. But even simple PTEs that focus on a single key mechanism and the outcome would provide evidence that programs were working through one particular mechanism or another. Meta-analytical findings could then be categorized by the key mechanism tested in original evaluations. If such mechanisms were tested across a number of PTEs, meta-analysis would be better suited to offering clues for effective intervention.

Returning to Table 6.1, the findings showed that cognitive programs were the most effective strategy for reducing recidivism. But no information on why cognitive programs were more effective was provided. Even more problematic is that the average effect size probably includes programs with a wide range of effects; some were likely very effective, but some likely had smaller effects on recidivism than even noncognitive treatments (for example, group counseling).

Instead of the ten evaluations of cognitive treatment reported in Table 6.1, what if a larger number of simple PTEs were included in the meta-analysis? A simple PTE is what Lipsey and Pollard (1989) describe as a two-step model—the measurement and testing of at least one mechanism for change and one outcome. In applying such a model to evaluations of treatment programs for sex offenders, the design would test whether the program first changed something to subsequently affect recidivism (see Table 6.2).

In Table 6.2, the results from Table 6.1 are compared with a hypothetical meta-analysis of a series of simple PTEs. The advantage of the PTE meta-analysis is that clues about mechanisms for change are provided. Effective programs are more easily identified by the key mechanism they engage. In Table 6.2, for example, cognitive programs that increase the skills of sex offenders to identify and reduce their own high-risk situations are more effective in reducing recidivism. Cognitive programs, if they are to be employed, are even more effective when used in combination with targeting offender empathy. Such findings would be very useful in providing guidance for decision making.

Table 6.2. Comparing Hypothetical Meta-Analysis of One-Step Evaluations with Hypothetical Meta-Analysis of PTEs

Using One-Step Evaluations

Programs (N) *Recidivism*	*Sex Offense Recidivism*
Cognitive-based (11)	.30
Behavioral (10)	.11
Individualized (12)	.01
Group counseling (8)	–.05

Using Simple PTEs

Cognitive-Based Programs (N)	*Effect on Mechanism*	*Sex Offense Recidivism*
Skills in victim empathy (7)	.61	.44
Skills in high-risk situations (14)	.55	.38
Reduction in rationalization (8)	.28	.12
Increase in empathy for victims (12)	.25	.09

Minimum Requirements of PTEs for Meta-Analysis

For PTEs to be exploited in meta-analysis, they should meet three criteria—an explicit causal model on how the program will affect outcomes, the testing of at least one underlying mechanism as an intervening variable along with outcomes, and control-group data reported for both variables. The first criterion requires a prospective and explicit model to be tested in the evaluation. Not only does the explicitness reduce the amount of guessing that the reader has to do about what the program theory was, but the prospective requirement prevents post hoc fitting of models to data.[5]

The second criterion requires that the PTE include at least one *mediator.* In evaluation terms, the mediator is something that the program must affect or change in order to positively influence the main outcome (Lipsey and Pollard, 1989). Some evaluations that have included program theory did not attempt to test any key links or mechanisms in the model (Petrosino, 2000). Even if a program theory is explicit, if only the outcome data are analyzed and reported, such evaluations provide no more information than one-step models; the program's causal theory was not tested. According to the third criterion, the control or comparison data must be reported for both the mediating and outcome variables. If data on the mediator are reported only for the treatment group, the evaluation provides little evidence that the effect on the mediator would have occurred without the program (see Cook, Chapter Three).

Process-outcome evaluations also do not meet the minimum requirement for PTE, as they provide no data on underlying mechanisms. Even when evaluators link the process data to outcomes in their analyses, these reflect the impact of program activities and the degree of fidelity on outcomes—not the underlying theory of change (Weiss, 1997). Something about these activities should engage a critical mechanism. What is that *something?* That is what PTE must articulate and test (see Weiss, Chapter Four).

How Could Meta-Analysis of PTEs Inform Larger Social Theories?

In PTE, theory is an explanation about how the program will cause the intended outcomes. Larger scientific theories offer general explanations of phenomena such as criminality, poor learning, or even how programs are implemented. A meta-analysis of PTEs could potentially inform such larger theories.

For example, in Chandler's (1973) experimental evaluation of a role-modeling intervention with troubled youth, he tested a two-step model: reducing egocentrism (that is, lack of empathy for others) would reduce delinquency (Lipsey and Pollard, 1989). Chandler conducted a two-year follow-up and found that treatment achieved statistically significant reductions in both egocentrism and delinquency.

Instead of just one experiment, Table 6.3 provides a hypothetical example of how fifty studies like Chandler's could inform delinquency theory. In Table 6.3, ten PTEs test the egocentrism model; four other groups of PTEs test different mechanisms for change. The table provides hypothetical effect sizes for each of the five categories. Interventions that targeted egocentrism hypothetically achieved the largest effects on both the mediator and subsequent delinquency. Such a finding suggests that a crucial link in the pathogenesis of delinquency is egocentrism.

The hypothetical findings also show that interventions generally had smaller effects on measures of self-esteem, job skills, intrafamily functioning, and fear of sanctioning. Such findings could lead delinquency theorists to reexamine the relationship between such factors and subsequent delinquency. Certainly, the effects could be due to poor program implementation or a generally ineffective treatment (if the ten studies were based on one common treatment type). But all things being equal, a generation of PTEs for meta-analysis could provide some guidance, particularly about problems like delinquency.

Minimum Threshold Levels and Cascading Effects. If program theory were well developed for a widespread intervention, PTE data could be exploited in meta-analysis to provide information for decision makers. For example, mediating and outcome variables could permit estimates of minimum threshold levels—the required improvement needed on the mediator to result in improved outcomes. This could be helpful to programs in which the mediators are measured at some point before the outcomes. Failure to achieve a mediating effect could serve as a red flag for decision makers that the program is en route to poor outcomes (Weiss, 1997).

Table 6.3. Hypothetical Effect Sizes for Mediating and Outcome Variables

Mediator (N)	*Mediator Effect*	*Delinquency Effect*
Egocentrism (10)	.64	.34
Self-esteem (10)	.48	−.07
Intrafamily functioning (10)	.36	.17
Job skill enhancement (10)	.22	.02
Fear of legal sanctioning (10)	.12	−.15

Note: Ranked by mediator effect.

Such data could also portray *cascading effects.*[6] With each subsequent mediating variable in certain program theories, smaller effects will likely be reported. For example, in knowledge-attitude-behavior models (Lipsey, 1997), programs that report large effects on knowledge usually report much smaller effects on the later attitudinal and behavioral measures. Data on cascading effects could be used to signal decision makers that a program needs to be retooled in order to achieve larger effects at later stages. Or maybe the model must be revised to account for other intervening variables.

The Value-Added of PTE. Meta-analysis could provide a method for assessing the *value-added* by PTE. Value-added often means something that can be measured mathematically, but here it refers to whether or not PTE provides some benefit beyond other approaches to evaluation. Although the benefits of PTE have long been suggested, they have not been empirically demonstrated. One way to test for value-added is through meta-analysis. For example, Lipton (1995) and his colleagues are conducting a meta-analysis of correctional program evaluations reported since 1968. Their meta-analysis will likely include over a thousand evaluations, some using different approaches such as PTE.

Their data could be used to compare PTE with these other evaluation approaches; the evaluations may not be easily categorized but could be rated along a continuum of how well developed the theory is that is used to guide the evaluation (Lipsey, 1988). The ratings could be analyzed to determine the influence of theory development in PTE on a number of dependent variables, including effect size, program success or failure, and so on. A small sample of the studies could be studied to determine how they were used in subsequent decisions. The data, though suggestive, could provide clues as to the real benefits that PTE provides over process-outcome or one-step model evaluations.

Barriers and Limitations

Few writers would argue against the inclusion of mediators in an evaluation design (see Cook, Chapter Three, for further discussion). But just as there are barriers in conducting a single PTE, there are roadblocks to using PTEs in meta-analysis. Some of them are outlined in the following discussion.

The Low Number of PTEs. The major barrier is the low number of available PTEs reported in the literature. Our own search for good examples of PTE was protracted and painful (see Rogers, Chapter Five, and others in this issue). Even simple PTEs requiring a two-step model are difficult to find; rarely do evaluators explicitly and prospectively articulate a model to be tested.

Emphasis on Experiments and Quasi-Experiments. Most meta-analyses require as an eligibility criterion that original evaluations include a control or comparison group. This is a trade-off, increasing internal validity but excluding potentially useful studies that take different methodological approaches to evaluate programs.

Poor Reporting. Reviewers are universally unhappy about the quality of reporting in original research documents. The combination of PTE and meta-analysis would require more data to be collected, analyzed, and reported by evaluators. Improving the quality of reporting is something that everyone recommends, but finding solutions has been difficult.

Simplistic Program Theories. This chapter has not taken into account complex models. The simple PTEs discussed here are linear and assume a domino effect: a change in one variable will result in a subsequent change in the next measured variable. As Rogers notes (Chapter Five), the world may not operate the way these models suggest. Even in linear theories, models can be lengthy. Weiss (1997) lists seventeen links in her job-training example. As she notes about evaluators doing original studies (Chapter Four), meta-analysts may also be forced to determine which links in which theory to code and examine in reviews.

One Recommendation for Promoting PTEs

As mentioned earlier, the major barrier to this approach is the lack of PTEs. Sherman and his colleagues (1997) suggest a method for increasing high-quality evaluations. In their review of crime prevention studies for the United States Congress, they also examined evaluation requirements set by the federal and state governments when they fund criminal justice–related programs. Although every grant recipient is usually required as a condition of funding to conduct an evaluation, Sherman and his colleagues found that few are reported. One problem is that what passes for evaluation is sometimes no more than the input data or information on clients served. Of those few outcome or impact studies that are conducted, control or comparison group designs are rarely implemented. The end result is that very little is known about what works in crime prevention, despite the billions of dollars spent over the past two decades (Sherman and others, 1997). One contributing factor is the generally inadequate amount of money allocated to evaluation within the program budget. Sherman and colleagues suggest a different approach: instead of requiring evaluation for every funded program, the administrative agency (usually at the federal or

state level) should pool the evaluation monies together to support a small number of high-quality evaluations in just a few sites. Such an approach could help promote an increase in the number of rigorously controlled PTEs.

An example of how this strategy would work in practice is provided by one federal funding program.[7] The Title V Grants for Local Delinquency Prevention Programs from the Office of Juvenile Justice and Delinquency Prevention supported 477 distinct interventions in the United States during fiscal years 1994–1997 (Office of Juvenile Justice and Delinquency Prevention, 1998). Instead of requiring an evaluation at all 477 sites (generally budgeted at $10,000 per site), the $4.7 million earmarked for evaluation could be spent funding PTEs at 20 sites. Each of the PTEs would include random assignment to conditions or reasonably valid comparison groups. Each evaluation would receive $200,000 per site (for a total cost of $4 million). The other monies ($700,000) could be used to build in some limited data collection to monitor non-PTE sites. Following this approach, a meta-analysis of the 20 PTEs could be conducted in a reasonably short time. A systematic and rigorously performed review of 20 PTEs would certainly yield much better information than the 477 low-quality and scattered evaluations that inevitably would be reported.

Conclusion

The short history of meta-analysis indicates that it has the potential to be both informative and influential (Hunt, 1997). Meta-analysis also means that many evaluations will be utilized by reviewers for years—and will have influence beyond the original jurisdiction in which they were conducted (Lipsey, 1997). Given the eternal "shelf life" of some evaluations because of meta-analysis, evaluators could greatly contribute to the knowledge base by conducting even simple PTEs.

Meta-analysis will take on even greater importance in light of a new international organization, known as the Campbell Collaboration, created in 1999 to facilitate the preparation of systematic, updated, and multinational reviews of social program evaluations (Davies, Petrosino, and Chalmers, 1999). Named after Donald Campbell and modeled after its older sibling in health care (the Cochrane Collaboration), the Campbell Collaboration will also explore methods for improving the precision and validity of both original evaluations and subsequent meta-analysis (its Web site is http://campbell.gse.upenn.edu).

One way to improve meta-analysis is to incorporate more PTEs, providing reviewers with data on mechanisms that could be exploited. In this chapter, four potential benefits for the use of PTE data in meta-analysis have been suggested. Such a happy marriage of these two perspectives would produce an offspring of much better evidence to guide decision making.

Notes

1. Some may argue that factorial experiments isolate mechanisms for why a program works.

2. I prefer the term causal model because of the general confusion surrounding "theory" (Petrosino, 2000).

3. Lipsey (1990) provides a conversion formula for effect sizes.

4. Rosenthal and Rubin's (1982) BESD can convert effect sizes into percentage differences, permitting communication of results to nonacademic audiences.

5. Some evaluations collect considerable data on participants, and some information could be conceptualized as mediating variables. Meta-analysis can handle such "kitchen sink" evaluations—if there are enough of them—and can examine correlations for both mediators and outcome measures.

6. Mark Lipsey suggested the term cascading effects (personal communication with author, April 1998).

7. Agencies usually administer grants under different funding "streams." For example, the U.S. Department of Justice has many streams of funding (such as the Violence Against Women Act). Block grants are made to states for each stream, and states then make subgrant awards (Sherman and others, 1997).

References

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

Chandler, M. J. "Egocentrism and Antisocial Behavior: The Assessment and Training of Social Perspective-Taking Skills." *Developmental* Psychology, 1973, *9*(3), **326–332.**

Chen, H. T., and Rossi, P. H. (eds.). Using *Theory* to Improve Program and Policy *Evaluations.* Westport, Conn.: Greenwood Press, 1992.

Cordray, D. S. "Theory-Driven Meta-Analysis: Practices and Prospects." In H. T. Chen and P. H. Rossi (eds.), Using Theory to Improve Program and Policy Evaluations. Westport, Conn.: Greenwood Press, 1992.

Davies, P., Petrosino, A., and Chalmers, I. "The Effects of Social and Educational Interventions: Developing an Infrastructure for International Collaboration to Prepare, Maintain, and Promote the Accessibility of Systematic Reviews of Relevant Research." In Proceedings *of the* international meeting on Systematic Reviews of the Effects of *Social* and Educational Interventions, London, July 15–14, 1999. Donaldson, S. I. "Mediator and Moderator Analysis in Program Development." In S. Sussman (ed.), Handbook of Program Development *in* Health Behavior Research *and Practice,* forthcoming. Thousand Oaks, Calif.: Sage.

Greater Egypt Regional Planning & Development Commission. *Menard* Correctional Center. Juvenile Tours Impact Study. Carbondale, Ill.: Greater Egypt Regional Planning & Development Commission, 1979.

Hunt, M. The *Story of Meta-Analysis.* New York Russell Sage Foundation, 1997.

Laws, R. E. Relapse Prevention with *Sex* Offenders. New York: Guilford Press, 1989.

Lipsey, M. W. "Practice and Malpractice in Evaluation Research." *Evaluation* Practice, 1988, *9*(4), 5–24.

Lipsey, M. W. Design Sensitivity: Statistical Power for Experimental Research. Thousand **Oaks,** Calif.: Sage, 1990.

Lipsey, M. W. "What Can You Build with Thousands Of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation." In D. J. Rog and D. Fournier (eds.), Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods. New Directions for Evaluation, no 76. San Francisco: Jossey-Bass, 1997.

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

Lipsey, M. W., and Wilson, D. "The Efficacy of Psychological, Educational and Behavioral Treatment: Confirmation from Meta-Analysis." American Psychologist, 1993, 48(12), 1181–1209.

Lipton, D. S. "Correctional Drug Abuse Treatment Effectiveness: Updating the Effectiveness of Correctional Treatment 25 Years Later." Journal of Offender Rehabilitation, 1995, 22, 1–20.

Office of Juvenile Justice and Delinquency Prevention. 1998 Report to Congress: Title V Incentive Grants for Local Delinquency Prevention Programs. Washington, D.C.: Office of Juvenile Justice and Delinquency Prevention, 1998.

Petrosino, A. "Answering the Why Question in Evaluation: The Causal-Model Approach." Canadian Journal of Program Evaluation, 2000, 12(1), 1–25.

Rosenthal, R., and Rubin, D. "A Simple, General Purpose Display of Magnitude of Experimental Effect." Journal of Educational Psychology, 1982, 74, 166–169.

Sherman, L. W., and others. Preventing Crime: what Works, What Doesn't, What's Promising: A Report to the United States Congress. College Park: Department of Criminology and Criminal Justice, University of Maryland, 1997.

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

Weiss, C. H. "Theory-Based Evaluation: Past, Present, and Future." In D. J. Rog and D. Fournier (eds.), Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods. New Directions for Evaluation, no. 76. San Francisco: Jossey-Bass, 1997.

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.