Confirmatory Program Evaluation: A Method for Strengthening Causal Inference

Arthur J. Reynolds

American Journal of Evaluation 1998 19: 203
DOI: 10.1177/109821409801900204

Confirmatory Program Evaluation: A Method for Strengthening Causal Inference

ARTHUR J. REYNOLDS

ABSTRACT

This paper discusses current issues in theory-driven evaluation from the perspective of the evaluation practice literature. Applications of theory-driven approaches are not easily found, in part, because there are few procedures for conducting them. I offer confirmatory program evaluation as one way to use theory, in combination with quantitative analytical techniques, to assess the effects of social and educational programs. In contrast to many other approaches, theory-driven evaluation generally emphasizes the explication and testing of *a priori* program theories in determining effectiveness. Confirmatory program evaluation is an impact assessment that examines the pattern of empirical findings against several causal criteria, including temporality, size, gradient (dosage/response), specificity, consistency, and coherence of the program-outcome relationship. A special emphasis is given to identifying causal mechanisms or active ingredients of programs that yield effects. An illustration of confirmatory program evaluation is provided for a child development intervention called the Child Parent Center Program. The limitations of this method are discussed, as well as the conditions under which it is most useful.

Arthur J. Reynolds

The purpose of this article is to describe an approach for conducting theory-driven outcome evaluations that I call confirmatory program evaluation (CPE). Confirmatory program evaluation is designed to clarify the presence or absence of program effects through a systematic process of program analysis. It is appropriate with experimental, quasi-experimental, or non-experimental data. It is most useful when there is extensive longitudinal data available and an established theory of the program. A confirmatory evaluation method facilitates causal inference because it organizes and synthesizes evidence about the size, specificity, consistency, and coherence of the program-outcome relationship, and because it tests the causal mechanisms (i.e., active ingredients) that lead to program outcomes. In that theory-driven

Arthur J. Reynolds • Waisman Center, University of Wisconsin-Madison, 1500 Highland Avenue, Madison, WI 53706; E-Mail: ajreynol@facstaff.wisc.edu.

American Journal of Evaluation, Vol. 19, No. 2, 1998, pp. 203–221.
ISSN: 1098-2140

approaches are relatively unknown and underused in applied social science, the CPE method outlined here may be useful for stimulating greater use of theory in evaluation.

This article is organized as follows. First, theory-driven evaluation is described in the context of more traditional approaches to estimating the effects of social programs. The lack of theory-based outcome evaluations is discussed, followed by a description of CPE. Six criteria are highlighted to organize and interpret findings from this approach. An empirical example from the field of child development is used to illustrate several aspects of conducting such an evaluation and interpreting findings. Finally, the limitations and implications of confirmatory program evaluation are discussed.

INTRODUCTION

Theory-driven evaluations are increasingly recommended as a viable approach to understanding the effects of social programs. In a theory-driven outcome evaluation, the explicit theory of the program is highlighted to establish an *a priori* model of how the program is expected to exert its influence (Bickman, 1987; Chen, 1990; Chen & Rossi, 1983; Worthen, 1996). Causal uncertainty is reduced through an examination of the empirical pattern of findings against the expectations inherent in the program. Contrast this perspective with a purely method-driven approach, in which causal uncertainty is reduced through control exercised during the research design phase of the evaluation, or the statistical modeling approach, whereby control is exercised during the data analysis phase via statistical adjustment. Although these two approaches have their advantages, neither can answer by themselves how and why programs work. This is the contribution of theory-driven evaluation.

Program theory is, not surprisingly, of central importance in theory-driven evaluation. It is typically defined as the "construction of a plausible and sensible model of how a program is supposed to work" (Bickman, 1987, p. 5). Program theory can be based on the application of a social science theory (e.g., labeling, attribution) to a specific program (e.g., delinquency prevention) and target population. (Many programs, however, are developed without the aid of social science theory.) In any event, it is a "small" theory specific to the program and may not generalize across individuals and programs (Lipsey, 1993). Main elements of the program theory include specification of the following: (a) the problem area, or behavior to be addressed by the program and the target population and context conditions; (b) program content, or skills to be acquired that will be sufficient to produce an effect (i.e., active ingredients); and (c) key responses and outcomes of the program by domain. The theory may derive from a variety of sources such as previous research findings, social science theory, program designers, or, if necessary, from the evaluator.

Specification of the program theory is helpful at all stages of evaluation including planning, implementation, and impact assessment. Among its benefits are the increased ability to identify the program and target groups, specify intervening and causal mechanisms, discriminate between program failure in implementation and theory failure, uncover unintended effects, improve the formative use of findings, and contribute to social science knowledge (Bickman, 1987; Chen, 1990; Lipsey, 1993).

Theory-based evaluations can be more confirmatory than other evaluation approaches, given their emphasis on multivariate prediction based on the program concept. Based on the program theory, for example, an evaluator can explicate or test the following: (a) the size of the program effect, (b) the program outcomes that yield the largest as well as the smallest

effects, (c) the consistency of effects across subgroups, models, and analyses, (d) the causal mechanisms or pathways through which the estimated effects are manifested, and (e) the factors that may influence selection into the program and implementation quality. Thus, a major assumption of this approach is that causal inference is strengthened if the empirical patterns of results are consistent with the program theory and hypotheses about the effects. Analyses of theory-driven evaluations are often conducted through traditional analysis of variance and regression procedures, but also include structural modeling as well as pattern matching (Trochim, 1985). Several approaches are summarized in Chen (1990), Bickman (1987), Chen and Rossi (1992), and Reynolds and Walberg (1994). Bickman (1996) illustrates a theory-driven perspective in evaluating mental health services for children and youth.

Lack of Theory-Based Evaluation in Practice

Given that evaluation theorists have discussed the importance of program theory for decades (Suchman, 1967; Weiss, 1972; Wholey, 1979), one might expect theory-driven evaluations to permeate the literature. They do not. Lipsey, et al. (1985) reviewed 119 published evaluation studies and found that two-thirds used program theory at no higher than the subtheoretical level (i.e., vague program descriptions). Only nine percent of the studies reviewed were classified as displaying "integrated" theoretical frameworks.

In the past decade, discussions of theory-driven evaluations have become more prevalent in the educational and social science literature (Bickman, 1987; Chen, 1990; Chen & Rossi, 1992; Reynolds & Walberg, 1990, 1994; Worthen, 1996). Despite this fact, the use of theory-driven approaches has been slow to catch on. They remain outside the mainstream of evaluation practice and the literature is largely absent from major journals that serve the social science disciplines.

There are at least three explanations for this state of affairs. First, the field of program evaluation remains significantly associated with research methodology in the tradition of Campbell and Stanley (1966) and Cook and Campbell (1979). Although ethnographic and mixed-method approaches are more frequent today, evaluation in most texts is defined as the application of social science methods to the investigation of social and educational programs. The central question usually is to estimate the main effects of program participation. Explanatory and process evaluation are viewed as supplemental concerns. According to Scriven (1994), "the professional imperative of the evaluator is to evaluate; anything else is icing on the cake" (p. 76). In contrast, theory-driven approaches are more comprehensive in scope and explanation is of primary concern (Cronbach, 1982; Lipsey, 1993).

Second, because program evaluation is an inherently practical activity, it is widely believed that theoretical evaluation is inconsistent with stakeholder and policymaker concerns. Accordingly, theoretical research is typically viewed as the province of basic science, not applied science. The atheoretical status of program evaluation is a principal reason for its marginal status within many social science disciplines (Wang & Walberg, 1983). Even if the relationship between program theory and evaluation practice is accepted, it is not prominent. Of course, existing theory in some program areas is inadequate to inform evaluation practice.

A third explanation for the limited use of theory-driven evaluations is confusion about the meaning and implementation of the approach itself. At a conceptual level, the term "theory-driven" can be mistakenly perceived as normative and pejorative. Both the casual and informed readers of evaluation literature could infer that theory-based evaluations are superior to other approaches, which are, by definition, atheoretical and less desirable scientifically.

Also, the "theory" label suggests that theory plays no role in other evaluation approaches. While theory may not play a central role in these other approaches, relative differences may be lost in the translation.

At an operational level, theory-driven evaluation may be viewed more as a philosophy for conducting evaluations rather than an approach with an explicit methodology or procedure. Yet most proponents describe theory-driven evaluation in practical and operational terms (Bickman, 1987; Chen, 1990; Chen & Rossi, 1992; Lipsey, 1993). To date, there is no consensus about how to conduct a theory-driven evaluation. This state of affairs may have limited its dissemination. Vagueness about implementation is not evident for other method-driven or statistical modeling approaches. Indeed, the explicitness and routinization of these evaluations (e.g., checklists, "how to" guides, software programs) are major strengths of their appeal. CPE is an attempt to delineate one way of conducting theory-driven evaluations.

Nature of Causal Inference in Social Programs

A common belief among several social science disciplines and program areas is that the only way to draw valid causal inferences is through experiments; anything less is insufficient. This belief is understandable if viewed in the historical context of scientific inquiry and the practice literature of the past two decades (Campbell, 1994; Cook & Shadish, 1994). A careful reading of the postpositivist literature on program evaluation, however, indicates that this view is, at best, a narrow interpretation of the nature of causal inference. In a volume devoted to Donald Campbell's four decades of methodological contributions to social science (Overman, 1988), four central themes are evident:

A. All knowledge, however acquired, is fallible. Multiple methodologies are preferred to establish causality.
B. Experiments only probe causal relations and theories; they cannot prove them.
C. The key to causal inference is to rule out *plausible* rival hypotheses. Methodologies only provide a means for doing this. Alternative explanations for program findings are "innocent until proven guilty" of plausibility.
D. While their record is mixed, quasi-experiments can lead to valid causal inferences. The nature of the evidence and pattern of findings are crucial in interpretation.

The confirmatory and theory-based evaluation approaches accept these propositions but support a fifth key theme:

E. The plausibility of an estimated program effect can be enhanced through systematic testing of causal mechanisms and other aspects of the program-outcome relationship.

Although experiments are often the most preferable approach to outcome evaluation, researchers often have to "make do" with designs that are "good enough," given the importance of the program and available resources (Rossi & Freeman, 1993). As discussed in the following sections, there are many things an evaluator can do to enhance causal interpretations regardless of the approach used. Some of these approaches, such as investigating causal mechanisms, are underused in evaluation practice (Cook, Anson, & Walchli, 1993; Cronbach, 1982; Mark, et al., 1992; Rosenbaum, 1984, 1995).

CONFIRMATORY PROGRAM EVALUATION

Confirmatory program evaluation is one method of conducting a theory-driven evaluation in which the objective is to facilitate causal inference about the relationship between program participation and measured outcomes. It is an outcome or impact evaluation in which hypotheses about the program are tested, based on the program theory. Unlike theory-driven evaluation generally, CPE specifically focuses on outcomes by quantitatively estimating program impact. Thus, CPE is distinct from other theory-based approaches such as evaluability assessment (a pre-evaluation) and implementation evaluation, although it does complement them. CPE may be applied to experimental, quasi-experimental, or nonexperimental data, but it enhances causal inference most in quasi-experimental and nonexperimental designs. CPE is primarily designed for investigating effects at the postprogram stage and during postprogram follow-up periods. In many respects, CPE can be viewed as a longitudinal process evaluation.

CPE attempts to strengthen causal inference through systematic investigation of the nature of the relationship between treatment and outcome. Of special interest is testing the causal mechanisms that may lead to longer-term program effects. In CPE, the evaluator investigates the empirical relationships among program, intervening, and outcome variables. A drug abuse prevention program, for example, may be based on the theory that low self-image leads youth to experiment with drugs and to use them frequently. A school-based, social problem-solving program may then be implemented to improve perceived self-competence. If the program alters drug usage by enhancing students' self-image (and not via some other factor), preliminary support for the program would be achieved. To support interpretation of effects, alternative theories of drug abuse prevention could be postulated (e.g., theories based on knowledge acquisition, family functioning), as well as more complex multivariable mechanisms. The theory might also specify which particular outcomes would be most affected by the program (e.g., marijuana use), and which will be least affected (e.g., alcohol use, school achievement). Such systematic testing can be aided by the use of several criteria for interpreting findings.

Six Criteria for Interpreting Findings in Confirmatory Program Evaluation

Causal inferences about the effects of programs can be facilitated by six empirically verifiable criteria. Adapted from Susser (1973) and Anderson, et al. (1980), satisfaction or affirmation of these criteria in a CPE strengthens the likelihood that the relationship between program participation and outcomes is causal. Although satisfying these conditions enhances the capacity to draw causal inferences, they are not fail-safe. Three qualifications should be considered:

A. Although empirical support for the criteria increases confidence about the relationship between program participation and outcome, lack of support for one or more criteria does not necessarily invalidate a program-outcome relationship.
B. Interpretation of evidence concerning the criteria may be affected by model specification. This is considered a part of the confirmatory approach.
C. The importance of the criteria may differ somewhat by program content and objectives as well as prior research.

In sum, these criteria help build a case for interpreting the effects of a program. Indeed, they are not specific to evaluation, but are relevant to any assessment of causality. Three key assumptions of CPE are that program objectives can be articulated, the program is implemented largely as intended, and that the program theory can be adequately measured.

These cumulative criteria are described in order of least important to most important.

1. Temporality of program exposure. At the most basic level of causal inference, the causal variable (i.e., program participation) must occur prior to the measurement of the response to the program or outcome. Although most evaluation studies satisfy this criterion, studies that are based on secondary analysis of survey data often measure program participation and outcome at the same time, or through retrospective recall. The direction of causality in such studies is not often clear.

2. Strength of association. At the next level of inference, the larger the association between program participation and intended outcome (or size of the estimated program effect), the more likely the association represents a real causal effect. Other factors being equal, a program that yields an effect size of one to two standard deviations, for example, is likely to have a meaningful effect on participants even if treatment and comparison groups are not randomly assigned. The interpretation is that selection bias, testing, or other unforseen circumstances would have to be so severe that they would be beyond the realm of plausibility in most evaluation contexts. The size rule of causal interpretation (Bross, 1967) is one application of the strength of association perspective. Unfortunately, most social programs do not demonstrate effects of this magnitude (Lipsey & Wilson, 1993). Nevertheless, strength of association can play an important role in weighing the evidence about a causal hypothesis. Relative to sample characteristics, program content, intensity, and duration, greater associations between program participation and outcomes can strengthen causal interpretation.

Detection of a strong relationship between program participation and outcome is particularly important if the empirical relationship is consistent with the program theory. Long-lasting programs (e.g., one year or more), intensive programs (e.g., having extensive contact time), or those that provide comprehensive services (e.g., family, educational, and community resources), would be generally expected to have greater effects on behavior than relatively brief or low-intensity programs, or those with a limited array of services. Consequently, an evaluator may be able to postulate the approximate size of the program effect in advance of the data analysis.

3. Gradient effect (dosage/response). A causal inference is more warranted if, other factors being equal, a monotonic relationship exists between program exposure (e.g., number of days or sessions attended, number of contact hours, number of years of participation) and the program outcome. That is, causal inference is strengthened if the outcome condition improves as an increasing function of the amount and duration of program participation. Of course, it is important to control for differences that could lead to different levels of program exposure. Moreover, the absence of a dosage/response relationship does not rule out a causal relationship. Nonlinear relationships and threshold effects may present alternative patterns, and they may be specified as well. The presence of a dosage-response relationship, however, does increase confidence in the capacity to infer causality.[1]

Outside of public health, medicine, and education, gradient effects are rarely investigated, probably because program participation often is not coded as a continuous variable. The gra-

dient (dosage/response) effect was a major criterion for determining that the relationship between cigarette smoking (treatment) and lung cancer is causal (U. S. Department of Health, Education, and Welfare, 1964). Not only do nonsmokers enjoy a lower rate of lung cancer than smokers, but lung cancer rates as well as death rates from lung cancer increases as the number of cigarettes smoked increases.

In education, gradient effects have been consistently found between school achievement and several "treatment" variables such as the amount of time on task, instructional time, hours of homework, and quality of instruction (Walberg, 1986). Thus, in addition to enhancing statistical power in program evaluation, gradient effects also strengthen an investigator's capacity to draw causal inferences. This is especially the case when experiments are not possible, such as in studies of smoking and lung cancer.

4. Specificity. Specificity of association refers to the situation in which the program-outcome relationship is limited to certain domains of behavior or outcome conditions. Causal inference is more straightforward in such cases. In the smoking example, epidemiological research indicated that while risk of cancer increased in smokers for all kinds of cancer, it was highest for lung cancer. Thus, causal inference was strengthened. This pattern of findings was further enhanced by the occurrence of gradient effects and causal mechanisms (i.e., carcinogenic effects of smoking on lung tissue).

In social and educational programs, specificity of effect can be predicted on the basis of the program theory. Findings that are consistent with the program theory and inconsistent with other theories would strengthen causal inference. An estimated effect is more likely to be real if the program affects a response that is consistent with the workings of the program. If a reading program for slow learners is truly effective, then it should affect vocabulary development or comprehension more than quantitative skills or social skills in the classroom. Likewise, a delinquency prevention program based on knowledge acquisition and role playing should impact anti-social behaviors more than academic achievement or occupational expectations.

Evidence that program effects vary by outcome domain not only may support a particular program theory, they may also help refute counterfactuals. In a delinquency prevention program designed as a quasi-experiment, if the program reduced youth antisocial behavior, but did not affect attendance or school achievement, it would be difficult to argue that selection bias explains these differential findings. If "creaming" occurred, then why wouldn't the program group also incur higher rates of attendance and achievement? Trochim's (1989) perspective on concept mapping also shows the value of the specificity criteria.

One consequence of the specificity hypothesis is that it requires an evaluator to collect data on several nonequivalent outcome variables for both treatment and control groups (Cook & Shadish, 1994). Having a program theory and specific objectives makes it considerably easier to identify and measure these outcomes.

5. Consistency. Consistency of association between program exposure and outcome indicates whether the estimated program effect is similar across sample populations and subpopulations, similar at different times and places, under different types of analyses and model specifications, and for similar program theories. The greater the consistency of findings favoring positive effects (or alternatively, absence of effects), the more likely the observed effects are real.

There are two dimensions of consistency. Evidence of within-study consistency would be based on the degree to which evaluation data are robust and sensitive (Rosenbaum, 1995;

Rosenbaum & Rubin, 1984; Rubin, 1986). For example, if a program was found to have similar effects for boys and girls and without regard to age and race, a conclusion about program impact would be better justified. If these positive findings remained the same under alternative analytic techniques (i.e., regression, ANCOVA, simultaneous equation modeling) and different model specifications (e.g., covariates), confidence about program impact also would increase.

Evidence for between-study consistency is based on the correspondence of study findings with previous studies using different samples, social contexts, and program variations. If these studies show a consistent and interpretable pattern of results, causal inferences are more likely. Of course, the value of using the existing knowledge base to enhance causal inference about the effects of a particular program depends on the consistency of findings and similarity of conditions in prior studies, as compared to the study under consideration. Meta-analysis is certainly helpful in this regard as well.

For example, the relationship between participation in compensatory preschool education programs and cognitive school readiness has been investigated in hundreds of controlled studies over the past 30 years for different programs, contexts, samples of children, and for different levels of implementation (Haskins, 1989; McKey, et al., 1985; White, 1985). A consistent finding is that program participation enhances children's school readiness or early school performance (as measured by cognitive tests or teacher ratings). Causal interpretation that preschool intervention enhances cognitive readiness is thus likely to be accepted.

In cases where the studies have inconsistent findings, are not in the expected direction, or are not of expected magnitude, causal inferences are more difficult to establish. Investigating the contextual, program, or participant characteristics that may explain differential findings is often warranted. Some program theories, for example, may predict interaction effects (see Mark, Hoffman, & Reichardt, 1992). Nevertheless, a critical issue for CPE is to determine whether inconsistent or unexpected findings are due to theory failure, program implementation failure, or to limitations of the research design or data analysis. Information about the consistency of relationships can help probe these distinctions.

6. Coherence. At the highest level of causal interpretation is the extent to which the evaluation findings show a clear pattern of effects relative to the causes of behaviors the program is attempting to impact, the target population, the program theory, and the program implementation. In other words, do the findings about the effects of a treatment program in a particular study tell a convincing story about the effects of the program? The coherence criterion for enhancing causal inference in outcome evaluation attempts to integrate the five criteria of causal inference discussed above. Given the program theory, target population, and the characteristics of program implementation, do study findings dovetail with the evidence about the temporality, size, gradient, consistency, and specificity of effects? Are there consistencies, for example, between the size of the effects and those predicted by program theory? Moreover, do the causal mechanisms and pathways from program participation to program outcome provide a coherent explanation of the main-effect findings and the theory of the program? If answers to these and other questions are "yes," both coherence and causal inference are strengthened.

Although coherence of explanation about program effectiveness is best judged over several studies or through meta-analyses, CPE can probe relationships among variables and organize findings in a way that tentatively addresses coherence. Ethnographic and qualitative

knowledge about the programs at the local level also help. An example of accummulating evidence of coherence in practice follows.

The Role of Process Variables and Mediating Effects in Assessing Coherence of the Program-Outcome Relationship

One of the most powerful, yet underused techniques of CPE is investigating causal mechanisms as a means to verifying program impact. Once a main effect is demonstrated between program participation and outcome to a satisfactory degree, a critical question becomes the program-related processes that produced the effect. Often viewed as a secondary and independent question, determining the process(es) that mediate the effects of program participation often can reinforce the validity of main-effect findings by providing a plausible causal explanation (Bickman, 1987; Cook, et al., 1993; Lipsey, 1993; Mark, et al., 1992). If the identified causal pathways leading to the desired outcome are consistent with the theory and operation of the program, causal inference is strengthened and the coherence of the program-outcome relationship is supported. Thus, the identification of causal mechanisms is the sine qua non of CPE.

An excellent illustration of the critical importance of investigating causal mechanisms in evaluation is the long-term research on the effects of the High/Scope Perry Preschool Program (Schweinhart, Barnes, & Weikart, 1993). For the evaluation, 125 economically disadvantaged children were randomly assigned to a half-day, structured preschool program, or to a no-treatment control group, over a five-year period. They were then followed through the school-age and early adult years. The program was based on Piaget's theory of cognitive development and implemented a "plan-do-review" daily routine (Schweinhart & Weikart, 1988). The program had substantial effects on participants. Some of the long-term findings were that participants had significantly higher achievement test scores up to age 14, were more likely to graduate from high school (71% vs. 54%), had average or better literacy at age 19 (61% vs. 38%), and achieved higher monthly earnings by age 27 (29% vs. 7% earning $2,000 or more). Moreover, they were less likely than the comparison group to be placed in special education programs, or to incur frequent arrests (7% vs. 35%), or to receive social services (59% vs. 80%).

Although the method-driven, experimental design of the study led to the inference that the program caused these behavioral changes, the most important question was how could a one-or two-year preschool program at age four lead to such pervasive effects on children up to 23 years later? The answer, as demonstrated in Berrueta-Clement, et al. (1984) and Schweinhart, et al. (1993), was that the program elevated children's cognitive development and this cognitive advantage led to a diffusion or cumulation of positive effects during the school-age years. Compared to the control group, program participants were more scholastically motivated, were rated more positively by their teachers during elementary school, had higher school achievement, were less likely to be placed in special education, and had higher educational attainment. This cumulative advantage as shown by their causal model provided a convincing and coherent explanation that the effects of the program were real. More important, this causal explanation was not based on the experimental design of the study. Thus, a causal mechanism approach to program effectiveness, as implemented in confirmatory program evaluation, can directly strengthen the capacity to make causal inferences.

TABLE 1
Key Steps of Confirmatory Program Evaluation

Step	*Evaluation Activity*
1.	Specify program theory and processes that are expected to affect outcomes
2.	Identify and measure outcomes for indexing largest and smallest effects of program participation.
3.	Collect or utilize data on causal mediating factors of the program theory as well as key background factors.
4.	Estimate main effects of program for the total group and any relevant subgroups. Investigate gradient, consistency, and specificity of effects.
5.	If main effects are detected, test causal mechanisms of the program theory to explain outcomes. If not, conduct causal analysis to understand lack of effects.
6.	Interpret the pattern of findings to facilitate generalization and knowledge transfer.
7.	Identify formative uses of findings for program improvement.

Implementing a Confirmatory Program Evaluation

Table 1 displays the main steps involved with implementing a CPE. The key to implementation is to identify a program theory and delineate the causal mechanism(s) that contribute to producing a main effect. Ideally, multiple hypotheses or theories should be tested in an alternative models framework, much like confirmatory factor analysis or structural modeling. The key steps are summarized as follows:

1. Specify a program theory and model for testing the theory. Alternative theories are possible and often desirable. Specify the processes through which the program will achieve its short- and long-term objectives. The theory may be based on several sources, including disciplinary knowledge, program documents, or experiential knowledge. Measurement of key constructs of the theory are crucial to model testing.

   Based on the theory, identify the likely magnitude and domain of program effects and whether effects vary by subject characteristics. For example, a delinquency prevention program may be based on the theory that family-youth conflict is the key mediator of delinquency. In CPE, this and other hypotheses can be subjected to empirical tests.
2. Identify and measure outcomes that should produce the largest program effects, given the program theory, as well as the smallest program effects. Although all theory-relevant outcomes should be measured, often only one or two irrelevant outcomes may be available. Sampling on nonequivalent dependent variables helps assess specificity of treatment effects.
3. Collect data on background factors, program implementation, and mediating factors that are expected to transmit the effects of the program over time. This is especially important for quasi-experimental and nonexperimental studies. Measuring the causal mechanisms or active ingredients that promote effectiveness is especially critical.
4. Estimate differences in group performance (main effects) across the outcome variables. Investigate gradient effects by correlating treatment exposure with alternative program outcomes. If there is no comparison group, obtain local, regional, or national data, or compare program variations among exposed groups.

a. If a differential pattern of findings across outcomes emerges and would be predicted by the theory, sensitivity has been established.
b. If selection bias is suspected, conduct sensitivity analysis with different models, analytic techniques, and across subgroups. If similar findings emerge, consistency is established. For between-study consistency, compare findings with previous studies.
c. If the magnitude of estimated effects is large, or if there is evidence for gradient effects, program impact can be more confidently inferred.
d. Absence of main effects may suggest subgroup analyses (if theoretically expected) or reexamination of implementation and outcome measures.

5. If a main effect of treatment exists (based on experimental, quasi-experimental, or nonexperimental design), investigate the presumed causal mechanisms (active ingredients) of the program theory. Use hierarchical regression analysis, path analysis, or structural equation modeling techniques.

    a. If the tested causal mechanisms explain observed group differences, coherence is tentatively established. Notably, a causal mechanism must be significantly associated with program participation and with the program outcome simultaneously.
    b. Confirmatory tests of alternative models are desirable and are more convincing than exploratory analyses of one program theory.
    c. If main effects are not detected, analyses of causal mechanisms may help explain why effects did not occur. This analysis, however, may necessitate different intervening factors, because the factors that promote success may not be the same as those that limit effects. Assessing statistical power, quality of measures, and the implementation context also may be warranted.

6. Use the pattern of findings to establish a tentative interpretation about the program given the evidence on size, consistency, specificity, gradient, and coherence. Although causal mechanisms are one key, any of the criteria can help strengthen inferences, especially if used in combination (e.g., smoking and lung cancer).
7. Indicate implications for use of findings and knowledge generalization in the context of previous research.

    a. If consistent with problem development and other research, both consistency and coherence are enhanced.
    b. What are the implications for designing or modifying programs?

An Example of Confirmatory Program Evaluation From Child Development

A description of a study from the field of early childhood intervention will serve to illustrate the methodology of CPE. Reynolds, Mavrogenes, Bezruczko, and Hagemann (1996) used a CPE approach to investigate the causal mechanisms underlying the effects of a preschool intervention called the Chicago Child-Parent Center (CPC) Program. The program is a compensatory early educational intervention for economically disadvantaged children who are at risk of school failure. Like Head Start, the program includes educational, family support, and health components. A main objective of the program is to promote the acquisi-

tion of basic skills in reading and math, as well as to promote positive socioemotional development. Administered through the Chicago Public Schools, the program provides a half-day, center-based preschool for one or two years, beginning at age three. The key program components are: (a) a structured basic-skills approach to school readiness, (b) a parent program within the center administered by a parent resource teacher, (c) provision of preventive health services, and (d) family outreach services provided by the school-community representative. Staff-to-child ratios are 1 to 8. A diverse set of learning experiences is provided (e.g., through small group and large group activities, and field trips). Implementation studies indicate that the program was delivered as intended (see Chicago Public Schools, 1987; Reynolds, 1995).

The impact evaluation was based on a quasi-experimental design in which 240 children attended the preschool program and 120 did not attend preschool. Both groups enrolled in all-day kindergarten at the six original CPC sites in the fall of 1985 and were active in sixth grade in the spring of 1992. Both groups were mostly African American (95%), attended the same kindergarten schools, lived in the same neighborhoods, and were equally eligible to enroll in Title I funded programs. Groups also were similar on family education, socioeconomic status, and participation in later intervention (see Reynolds, et al., 1996). Data are part of the Chicago Longitudinal Study.

Program theory. The program theory is that children's early scholastic readiness for school entry and beyond will be facilitated through the provision of systematic language learning activities (through center-based early intervention) and opportunities for family support experiences (through parent involvement activities in and outside the center). Consequently, early scholastic development may improve children's longer-term school and social competence. The central theory is embodied in this goal statement for the program: "[CPC is] designed to reach the child and parent early, develop language skills and self-confidence, and to demonstrate that these children, if given a chance, can meet successfully all the demands of today's technological, urban society" (cf. Naisbitt, 1968). The key measures to assess this goal usually include reading and math achievement, grade retention, and special education placement.

Two causal hypotheses were postulated to explain the observed effects of the program on children's scholastic development (i.e., reading and math achievement). In the cognitive advantage hypothesis, the immediate positive effect of preschool on cognitive development at school entry initiates a positive cycle of scholastic development and commitment that culminates in improved school achievement over time (Berrueta-Clement, Schweinhart, Barnett, Epstein, & Weikart, 1984; Schweinhart, Barnes, & Weikart, 1993; Consortium for Longitudinal Studies, 1983). The family support hypothesis states that longer-term effects of interventions will occur to the extent that family functioning has been improved. Because early intervention programs often involve parents, family processes (e.g., parent-child interactions, school involvement) must be impacted to produce longer-term effects on child outcomes (Bronfenbrenner, 1975; Seitz, 1990). Because parent participation in children's schooling is a crucial part of the CPC program theory, parent involvement in school was used as the primary measure of family support.

Although these two hypotheses—cognitive advantage and family support—have been often investigated separately, they are not incompatible and indeed, are intuitively complementary. Thus, a third, more comprehensive hypothesis of the mechanisms through which

preschool intervention affects later school achievement is through both cognitive-advantage and family-support processes. This is the dual mechanism hypothesis.

Findings. The findings presented here illustrate how causal mediation can help enhance inferences about program impact. Thus, coherence was the major criterion for strengthening causal inference. Several design features and analyses of the Chicago Longitudinal Study support other criteria for strengthening causal inference. These include temporality of program exposure and gradient effects through years of participation (Reynolds, 1994, 1998), specificity of the program-outcome relationship to scholastic achievement and competence (Reynolds, 1994, 1995), and consistency of the program-outcome relationship. In Reynolds and Temple (1995), for example, effect sizes of program participation (both unadjusted and adjusted for measured and unmeasured variables) differed by no more than 10-15%. The size of the association between program participation and outcomes was similar to that of many other high-quality programs (Reynolds & Temple, 1995). Of course, each of these criteria could be a focus of program analysis. For brevity, however, they are not discussed.

Table 2 shows that main effects of program participation on grade 6 reading and math achievement as well as cumulative grade retention with and without controls for family and child background factors. Preschool intervention was significantly associated with grade 6 outcomes above and beyond that of the covariates. Preschool participants scored six standard-score points higher (adjusted) than the no-preschool participants in reading and math achievement (about five months of performance). Preschoolers also had a lower rate of grade retention, 19.6% versus 31.7% for the comparison group (a 35% reduction, adjusted). As indicated in the "covariate difference" column, the covariates accounted for 10% to 20% of the size of the estimates of program effects. Program effects were consistent under both "raw" and "adjusted" specifications and also did not vary by gender. Both findings support the consistency of the program-outcome relationship. The main issue is whether the program theory can explain these effects.

Results of the theory-driven, confirmatory model are shown in Figure 1. The model was estimated through latent-variable structural modeling (as implemented in the LISREL statistical program). This is only one of many analytic approaches (e.g., hierarchical regression, path anal-

TABLE 2
Raw and Adjusted Group Means for School Competence Outcomes in Grade 6

	Raw means					
Outcome	*Preschool group (n = 240)*	*No-preschool group (n = 120)*	*Raw diff*	*Adj. diff*	*ES*	*Covar. diff*
Reading Achievement	126.8	119.8	7.0*	6.1*	.37	0.9
Math Achievement	131.9	126.3	5.6*	4.6*	.30	1.0
Grade Retention (%)	19.6	31.7	12.1*	10.6*	−.32	1.5

Notes: Covariates were sex, age, parent education, eligibility for free lunch, and years of primary-grade intervention. Parent education = parents' report of their educational attainment. Full lunch subsidy = child eligible for federal lunch subsidy. Age is measured in months. ES = effect size in standard deviations adjusted for covariates. Within-group standard deviations were 16.6, 15.5, and 0.45, respectively, for reading achievement, math achievement, and grade retention. The latter's effect size was adjusted for the probit method.
*$p < .05$

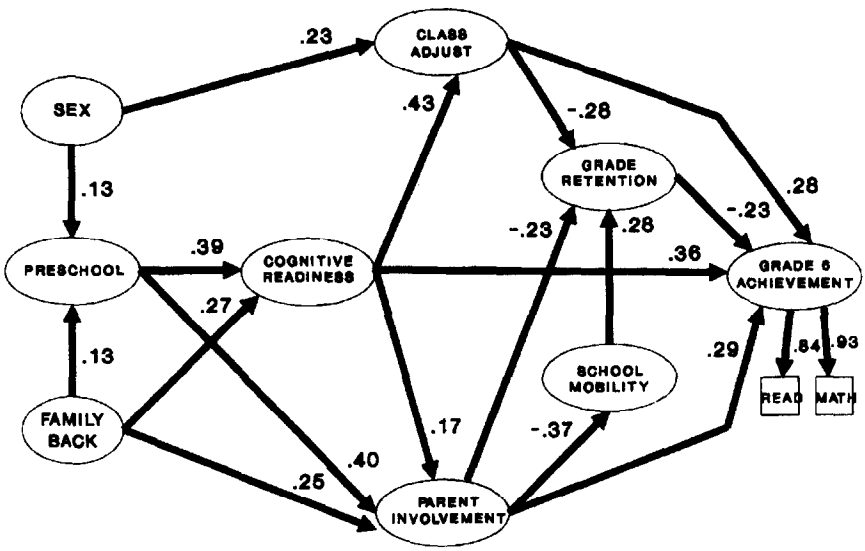

Figure 1. Mediated Effects of Preschool Intervention

ysis) for testing causal mechanisms. Preschool participation was measured in years (0 to 2). Cognitive readiness at age five was measured through the composite early primary battery of the Iowa Tests of Basic Skills. It reliably assessed several key attributes of early school success, including language development, word analysis, listening, and mathematics knowledge (all of which were intended outcomes). Parent involvement in school was a composite of teacher and parent ratings in grades 2 and 4. As a major construct of the program theory, parent involvement is associated with school achievement, especially for low-income children.

As displayed, the effects of preschool intervention on grade 6 achievement were substantially explained by two mechanisms predicted by the program theory—cognitive readiness at school entry and parent involvement in school (a measure of family support). Success from preschool participation also was enhanced by class adjustment (rated by teachers), avoidance of school mobility and grade retention, albeit indirectly. The scholastic benefits appear to be the result, in part, of the cognitive advantage and family support engendered. Notably, the significantly positive estimated effects of preschool participation on cognitive readiness and on parent involvement take into account age, family background (a composite of education & income), sex, and participation in primary-grade intervention.

As shown in Table 3, results indicated that the integrated, dual mechanism model of preschool mediation fit the data better than either the cognitive advantage or family support models. A significant improvement in model fit (chi square change) occurred when cognitive *and* family mediators were estimated together instead of separately. The dual mechanism model was the only one with uniformly acceptable fit statistics. For example, the probability levels for the RMSEA's, a test of close fit to population estimates, were well within the range of acceptability only for Model 4. Also, the alternative explanation of no-preschool mediation (Model 1) was rejected in favor of the mediated models (Models 2 to 4). These findings confirm those of the grade 3 follow-up study (Reynolds, 1992) and provide a coherent explanation of the estimated program effects consistent with the purpose of CPE.

TABLE 3
Comparison of Alternative Structural Models of Preschool Mediation

	Structural models	*df*	χ^2	χ^2 *change*	*AGFI*	*NNFI*	*RMSEA*
1.	Baseline: No preschool mediation	33	176.49	—	.83	.77	.110
2.	Cognitive advantage: Preschool mediation through cognitive readiness	32	113.70	62.79*	.87	.86	.084
3.	Family support: Preschool mediation through parent involvement	32	133.15	43.34*	.86	.83	.094
4.	Integrated: Preschool mediation through cognitive readiness and parent involvement (Figure 1)	30	66.65	66.50*	.92	.94	.058

Notes: AGFI = adjusted goodness of fit index. NNFI = non normed fit index. RMSEA = root mean square error of approximation. For Models 2 and 3, chi-square change = difference calculated from baseline model.
*p < .05

Although other hypotheses could be considered, these two were the most relevant to the theory of the CPC program. Causal hypotheses based on social and motivational advantage, for example, are not well supported in these data or in the literature. Indeed, Figure 1 shows that classroom adjustment—a measure of social development—did not mediate the relationship between preschool participation and grade 6 achievement.

Two implications of these findings are offered. Early childhood interventions are likely to be most effective if they target activities that optimize cognitive readiness and parent participation in the program. Indeed, these two mediators could be a focus of intervention. Moreover, the identified pathways may generalize to other child outcomes of development and to other intervention strategies.

Limitations of Confirmatory Program Evaluation

Confirmatory program evaluation is one approach to conducting an outcome evaluation, especially when causal inference based on method-driven and statistical approaches are not feasible or desirable. This approach complements other approaches to outcome evaluations; it does not replace them. Nevertheless, CPE has three limitations that may restrict its use in some studies. First, a CPE often requires more data collection than other kinds of evaluations. It requires the measurement of intervening causal mechanisms and precise treatment exposure, and benefits from a relatively large number of outcome variables (i.e., more than two), as well as extensive longitudinal follow-up of program participants (see example above). Moreover, data analysis is extensive in a CPE, especially with regard to investigating causal mechanisms.

A second limitation in the use of CPE is that the findings and inferences about treatments are largely dependent on the validity of the program theory and explanatory analysis. CPE works best when the program theories and knowledge base are well established. It could be more problematic in program areas in which it is difficult to specify and measure program theories, or if program effectiveness has been difficult to demonstrate. For example, there is less consensus about the best theories for programs aimed at preventing drug abuse and delinquency than for preventing low school readiness and school failure. In such cases, testing

alternative program theories and hypothesizing why programs do not produce their intended effects are warranted. In the event that findings do not support the program theory, respecification of the model, or further consideration of the program context may be needed through both observational and ethnographic studies.

Finally, confirmatory program evaluations are impact evaluations. In conducting them, an evaluator assumes that the objectives of the program can be accurately articulated, that program implementation has been verified, and that the program theory and associated causal mechanisms can be specified and measured. Certainly, these assumptions are not equally true of all programs. Deviations from these assumptions reduce the capacity to infer causality. In CPE, the explanatory power of the program theory is a key component of study findings. Verification of program theories, as well as identification of new theories, can be further enhanced by qualitative methods such as naturalistic and case-study approaches.

CONCLUSION

Confirmatory program evaluation uses program theory and quantitative analytic techniques to investigate, strengthen, and confirm causal inferences in outcome evaluations. It complements and extends other evaluation approaches. The main advantage of CPE is the specification and testing of a program theory to determine the active ingredients that promote program impact over time. Following an organized process designed to address six criteria of causal interpretation, inferences about program impact can be made more confidently. Analyses of causal mediation are a defining feature and it can be assessed in a variety of ways. An example from a child development intervention illustrated how confirmatory analyses can lead to a coherent understanding of how and why a program works. The validity of CPE depends on the adequacy of the program theory and the extent to which it explains the processes leading to identified outcomes. Confirmatory program evaluation is a systematic, explanatory approach to program evaluation that is consistent with scientific inquiry in the social sciences.

NOTES

Arthur Reynolds is an associate professor of social work, educational psychology, and child & family studies at the University of Wisconsin-Madison. His interests are prevention research, program evaluation, and social policy.

1. Programs in which participation is predominantly self-selective and whose length is open-ended are an exception to this pattern. Because length of treatment indexes motivation or need for intervention, the relationship between program participation and outcome in such cases is not usually a good estimate of program impact. For example, psychotherapeutic treatments, family counseling programs, extracurricular-activity programs, and job training programs often have these characteristics.

REFERENCES

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W., & Weisberg, H. I. (1980). *Statistical methods for comparative studies: Techniques for bias reduction.* New York: Wiley.

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S., & Weikart, D. P. (1984). *Changed lives: The effects of the Perry Preschool Program on youths through age 19*. Ypsilanti, MI: High/Scope.

Bickman, L. (1987). *The functions of program theory*. In L. Bickman, (Ed.), *Using program theory in evaluation: No. 33* (pp. 5-18). San Francisco: Jossey-Bass.

Bickman, L. (1996). A continuum of care: More is not always better. *American Psychologist, 51*, 689-701.

Bross, I. D. J. (1967). Pertinency of an extraneous variable. *Journal of Chronic Disease, 20*, 487-495.

Bronfenbrenner, U. (1975). Is early intervention effective? In M. Guttentag & E. Struening (Eds.), *Handbook of evaluation research* (Vol. 2, pp. 519-603). Beverly Hills: Sage.

Campbell, D. T. (1994). Retrospective and prospective on program impact assessment. *Evaluation Practice, 15*, 291-298.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

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

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

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

Chicago Public Schools. (1987). *ECIA Chapter I Evaluation of the Child Parent Centers*. Department of Research and Evaluation: Author.

Consortium for Longitudinal Studies (1983). *As the twig is bent...Lasting effects of preschool programs*. Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cook, T. D., Anson, A. R., & Walchli, S. B. (1993). From causal description to causal explanation: Improving three already good evaluations of adolescent health programs. In S. G. Millstein, A. C. Peterson, & E. O. Nightengale (Eds.), *Promoting the health of adolescents: New directions for the twenty-first century* (pp. 339-374). New York: Oxford University Press.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*, 545-580.

Cronbach, L. J. (1982). *Designing evaluations for educational and social programs*. San Francisco: Jossey-Bass.

Haskins, R. (1989). Beyond metaphor: The efficacy of early childhood education. *American Psychologist, 44*, 274-282.

Lipsey, M. W. (1993). Theory as method: Small theories as treatments. In L. Sechrest & A. Scott (Eds.), *Understanding causes and generalizing about them. New directions for program evaluation* (No. 57, pp. 5-38). San Francisco: Jossey-Bass.

Lipsey, M. W., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of science. In D. S. Cordray (Ed.), *Utilizing prior research in evaluation planning. New Directions for Program Evaluation* (No. 27). San Francisco: Jossey-Bass.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48*, 1181-1209.

Mark, M. M., Hoffman, D. A., & Reichardt, C. S. (1992). Testing theories in theory-driven evaluations: (Tests of) moderation in all things. In H. T. Chen & P. H. Rossi (Eds.), *Using theory to improve program and policy evaluations* (pp. 71-84). New York: Greenwood Press.

McKey, R. H., Condelli, L., Ganson, H., Barrett, B. J., McConkey, C., & Plantz, M. C. (1985). *The impact of Head Start on children, families, and communities* (DHHS Publication No. OHDS 85-31193). Washington, DC: U. S. Government Printing Office.

Naisbitt, N. (1968). *Child-Parent Education Centers, ESEA Title I, Activity I*. Unpublished report, Chicago, IL.

Overman, E. S. (Ed.). (1988). *Methodology and epistemology for social science: Selected papers of Donald T. Campbell*. Chicago: University of Chicago Press.

Reynolds, A. J. (1992). Mediated effects of preschool intervention. *Early Education and Development, 3*, 139-164.

Reynolds, A. J. (1994). Effects of a preschool plus follow-on intervention for children at risk. *Developmental Psychology, 30*, 787-804.

Reynolds, A. J. (1995). One year of preschool intervention or two: Does it matter? *Early Childhood Research Quarterly, 10*, 1-31.

Reynolds, A. J. (1998). The Chicago Child-Parent Centers: A longitudinal study of extended early childhood intervention. In J. Crane (Ed.), *Social programs that work*. New York: Russell Sage Foundation.

Reynolds, A. J., Mavrogenes, N. A., Bezruczko, N., & Hagemann, M. (1996). Cognitive and family-support mediators of preschool effectiveness: A confirmatory analysis. *Child Development, 67*, 1119-1140.

Reynolds, A. J., & Temple, J. A. (1995). Quasi-experimental estimates of the effects of a preschool intervention: Psychometric and econometric comparisons. *Evaluation Review, 19*, 347-373.

Reynolds, A. J., & Walberg, H. J. (1990). Program theory in evaluation. In H. Walberg & G. Haertel (Eds.), *International encyclopedia of educational evaluation*. NY: Pergamon.

Reynolds, A. J., & Walberg, H. J. (1994). Theory-based evaluation. In T. Husen & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education* (pp. 6378-6384). New York: Pergamon.

Rosenbaum, P. (1984). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79*, 41-48.

Rosenbaum, P. (1995). *Observational studies*. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies for causal effects. *Journal of Educational Psychology, 66*, 688-701.

Rossi, P. H., & Freeman, H. E. (1993). *Evaluation: A systematic approach* (5th ed.). Beverly Hills: Sage.

Rubin, D. B. (1986). Which ifs have causal answers? *Journal of the American Statistical Association, 81*, 961-962.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: The High/Scope Perry Preschool study through age 27*. Ypsilanti, MI: High/Scope.

Schweinhart, L. J., & Weikart, D. P. (1988). The High/Scope Perry Preschool Program. In R. H. Price, E. L. Cowen, R. P. Lorion, & J. Ramos-McKay (Eds.), *14 ounces of prevention: A casebook for practitioners* (pp. 53-65). Washington, DC: American Psychological Association.

Scriven, M. (1994). The fine line between evaluation and explanation. *Evaluation Practice, 15*, 75-77.

Seitz, V. (1990). Intervention programs for impoverished children: A comparison of educational and family support models. *Annals of Child Development, 7*, 73-103.

Suchman, E. A. (1967). *Evaluation research: Principles and practice in public service and social action programs*. New York: Russell Sage Foundation.

Susser, M. (1973). *Causal thinking in the health sciences, concepts and strategies of epidemology*. New York: Oxford University Press.

Trochim, W. M. K. (1985). Pattern matching, construct validity, and conceptualization in program evaluation. *Evaluation Review, 9*, 575-604.

Trochim, W. M. K. (1989). Concept mapping: Soft science or hard art? *Evaluation and Program Planning, 12*, 87-110.

U. S. Department of Health, Education, and Welfare (1964). *Smoking and health: Report of the advisory committee to the surgeon-general of the Public Health Service*. Washington, DC: Public Health Service Publication No. 1103).

Walberg, H. J. (1986). Synthesis of research on teaching. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 214-229). Washington, DC: American Educational Research Association.

Wang, M. C., & Walberg, H. J. (1983). Evaluating educational programs: An integrative causal-modeling approach. *Educational Evaluation and Policy Analysis*, *5*, 347-366.

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

Wholey, J. S. (1979). *Evaluation: Promise and performance*. Washington, DC: Urban Institute.

White, K. R. (1985). Efficacy of early intervention. *Journal of Special Education*, *19*, 401- 416.

Worthen, B. R. (1996). Editor's note: The origins of theory-based evaluation. *Evaluation Practice*, *17*, 169-171.