American Journal of Evaluation
32(2) 199-226
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1098214010389321
http://aje.sagepub.com
SAGE

A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009

Chris L. S. Coryn[1], Lindsay A. Noakes[1], Carl D. Westine[1], and Daniela C. Schröter[1]

Abstract

Although the general conceptual basis appeared far earlier, theory-driven evaluation came to prominence only a few decades ago with the appearance of Chen's 1990 book *Theory-Driven Evaluations*. Since that time, the approach has attracted many supporters as well as detractors. In this paper, 45 cases of theory-driven evaluations, published over a twenty-year period, are systematically examined to ascertain how closely theory-driven evaluation practices comport with the key tenants of theory-driven evaluation as described and prescribed by prominent theoretical writers. Evidence derived from this review to repudiate or substantiate many of the claims put forth both by critics of and advocates for theory-driven forms of evaluation are presented and an agenda for future research on the approach is recommended.

Keywords

theory-driven evaluation, theory-based evaluation, evaluation theory, evaluation practice, research review

Evaluation theories describe and prescribe what evaluators do or should do when conducting evaluations. They specify such things as evaluation purposes, users, and uses, who participates in the evaluation process and to what extent, general activities or strategies, method choices, and roles and responsibilities of the evaluator, among others (Fournier, 1995; Smith, 1993). Largely, such theories are normative in origin and have been derived from practice rather than theories that are put into practice (Chelimsky, 1998). Stimulated by Miller and Campbell's (2006) review of empowerment evaluation and Christie's (2003) research on the practice–theory relationship in evaluation, the authors of this review sought to replicate certain aspects of Miller and Campbell's study except that the phenomenon under investigation was theory-driven evaluation practice. Therefore, this inquiry also is intended to contribute to the scarcity of systematically derived knowledge about evaluation practice by investigating whether "theoretical prescriptions and real-world practices do or do not align" (Miller & Campbell, 2006, p. 297).

[1] The Evaluation Center, Western Michigan University, Kalamazoo, MI, USA

Corresponding Author:
Chris L. S. Coryn, 1903 West Michigan Avenue, Kalamazoo, MI 49008, USA
Email: chris.coryn@wmich.edu

As Weiss (1997a) remarked more than a decade ago, "The idea of theory-driven evaluation is plausible and cogent, and it promises to bring greater explanatory power to evaluation. However, problems beset its use ... " (p. 501). Consequently, the impetus for this investigation was multifaceted. One of the primary motives was the simple recognition of the continued and growing interest in theory-driven evaluation in the last few decades (Donaldson, 2007). Additionally, specifying program theory recently has been put forth as an essential competency for program evaluators (Stevahn, King, Ghere, & Minnema, 2005). Simultaneously, coupled with the recognition that the approach has not been adequately scrutinized in any meaningful way and in response to Henry and Mark's (2003) agenda for research on evaluation, this investigation also was motivated by the demonstrable need for knowledge regarding the degree to which evaluation theory is enacted in evaluation practice, which can provide valuable insights for improving future practice (Christie, 2003). However, little consensus exists regarding its nomenclature and central features (Donaldson, 2003; Rogers, 2007). A secondary purpose was, therefore, to enumerate a set of fundamental principles for theory-driven evaluation with the intent of describing the approach in a very general, yet consistent and cogent manner (Miller, 2010). Finally, as Stufflebeam and Shinkfield (2007) note, "... if evaluators do not apply evaluation theory ... then it is important to ask why they do not. Perhaps the approaches are not sufficiently articulated for general use, or the practitioners are not competent to carry them out, or the approaches lack convincing evidence that their use produces the needed evaluation results" (p. 62).

An Overview of Theory-Driven Evaluation

Although its origins can be traced to Tyler in the 1930s (with his notion of formulating and testing program theory for evaluative purposes; Garangi, 2003, as cited in Donaldson, 2007), later reappearing in the 1960s and 1970s (e.g., Suchman, 1967; Weiss, 1972) and again in the 1980s (e.g., Bickman, 1987; Chen 1980; Chen & Rossi, 1980, 1983, 1987), it was not until 1990 that theory-driven evaluation resonated more widely in the evaluation community with the publication of Chen's seminal book *Theory-Driven Evaluations.* Since then, conceptual, methodological, and theoretical writings (e.g., Chen & Rossi, 1992; Rogers, 2000, 2008; Rogers, Petrosino, Huebner, & Hacsi, 2000; Weiss, 1995, 1997a, 1997b, 1998, 2004a, 2004b)—and, to a lesser extent, actual case examples (Birckmayer & Weiss, 2000)—on the approach have been commonplace in both the serial and the grey literatures as well as in numerous books, book chapters, and conference presentations and proceedings where theory-driven evaluation is sometimes referred to as program-theory evaluation, theory-based evaluation, theory-guided evaluation, theory-of-action, theory-of-change, program logic, logical frameworks, outcomes hierarchies, realist or realistic evaluation (Mark, Henry, & Julnes, 1998; Pawson & Tilley, 1997), and, more recently, program theory-driven evaluation science (Donaldson, 2007), among many others (Rogers, 2000, 2008; Rogers et al., 2000; Stame, 2004).[1] Even though a common vocabulary, definition, and shared conceptual and operational understanding has largely been elusive, theory-driven forms of evaluation have, nonetheless, increasingly been espoused by numerous evaluation scholars and theorists, practitioners, and other entities as the preferred method for evaluation practice.

In one form or another, such approaches have been widely adopted, including evaluations conducted for and commissioned by the W. K. Kellogg Foundation (1998, 2000) for evaluating their community change initiatives, the United Way of America (1996) for evaluating their health, human service, and youth- and family-serving efforts, and the Centers for Disease Control and Prevention (CDC; Milstein, Wetterhall, & CDC Working Group, 2000) for evaluating public health programs and interventions. In the past few years, theory-driven approaches also have been increasingly promoted in international development settings, including some of the evaluations commissioned and conducted by the Overseas Development Institute (ODI), the International Initiative for Impact

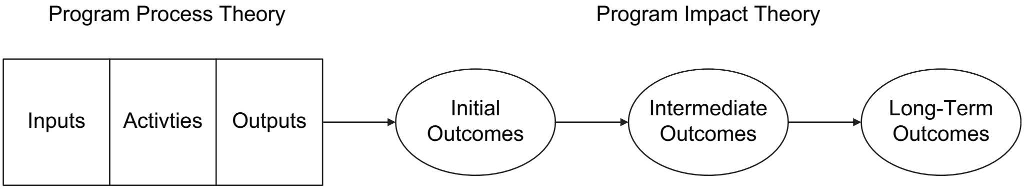

Figure 1. Linear program theory model. Source: Adapted from Donaldson, S. I. (2007). *Program theory-driven evaluation science.* New York, NY: Lawrence Erlbaum, p. 25.

Evaluation (3ie), the United Nations Evaluation Group (UNEG), and the Independent Evaluation Group (IEG) of the World Bank for evaluating humanitarian efforts (Carvalho & White, 2004; Conlin & Stirrat, 2008; White, 2007, 2009; White & Masset, 2007; World Bank, 2003, 2005). More recently, such approaches have been suggested as a means for evaluating military operations in the United States (Williams & Morris, 2009) as well as in a variety of other fields, settings, and contexts (Trochim, Marcus, Masse, Moser, & Weld, 2008; Urban & Trochim, 2009; Weiss, 1997b). Theory-driven forms of evaluation also have been recommended as one possible alternative to randomized controlled trials or randomized experiments—generally patterned after the evidence-based practice model in medicine—for both independent and federally sponsored initiatives charged with identifying efficacious or effective interventions (Government Accountability Office [GAO], 2009).

As defined by Rogers et al. (2000), program theory-driven evaluation is conceptually and operationally premised on " . . . 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" (p. 5). This broad designation is intended to encompass a wide variety of synonyms sometimes used to describe and encapsulate such evaluation approaches including, but not limited to, theory-driven, theory-based, and theory-guided forms of evaluation, but to the exclusion of certain others, and " . . . does not include evaluations that explicate the theory behind a program but that do not use the theory to guide the evaluation" (Rogers et al., 2000, pp. 5–6). Therefore, and although there are many variations and their meaning and usage often differ, the term *theory-driven evaluation* is used throughout this article to denote *any evaluation strategy or approach that explicitly integrates and uses stakeholder, social science, some combination of, or other types of theories in conceptualizing, designing, conducting, interpreting, and applying an evaluation.*

Program Theory in Theory-Driven Evaluation

Program theories are the crux of theory-driven forms of evaluation and are typically represented as graphical diagrams that specify relationships among programmatic actions, outcomes, and other factors, although they also may be expressed in tabular, narrative, or other forms. Such representations vary widely in their complexity and level of detail (Chen, 1990, 2005a, 2005b, 2005c; Donaldson, 2007; Frechtling, 2007; Funnel, 1997; Gugiu & Rodriguez-Campos, 2007; McLaughlin & Jordan, 1999; Patton, 2008; Rogers, 2000, 2008; W. K. Kellogg Foundation, 2000; Wyatt Knowlton & Phillips, 2008). A typical, linear program theory model is shown in Figure 1. Real models are, of course, often much more complex, but the essential message is generally the same.

The elements used to describe or represent a program theory often (but not always) include inputs, activities, and outputs, which in combination loosely form a program process theory, and initial outcomes (sometimes called short-term, proximal, or immediate outcomes), intermediate outcomes (sometimes called medial outcomes), and long-term outcomes (sometimes called distal outcomes or impacts), which are intended to represent a program impact theory, or some variation of these (Donaldson, 2007; Donaldson & Lipsey, 2006; Lipsey, Rossi, & Freeman, 2004; Patton, 2008). Inputs include various types of resources necessary to implement a program (e.g., human,

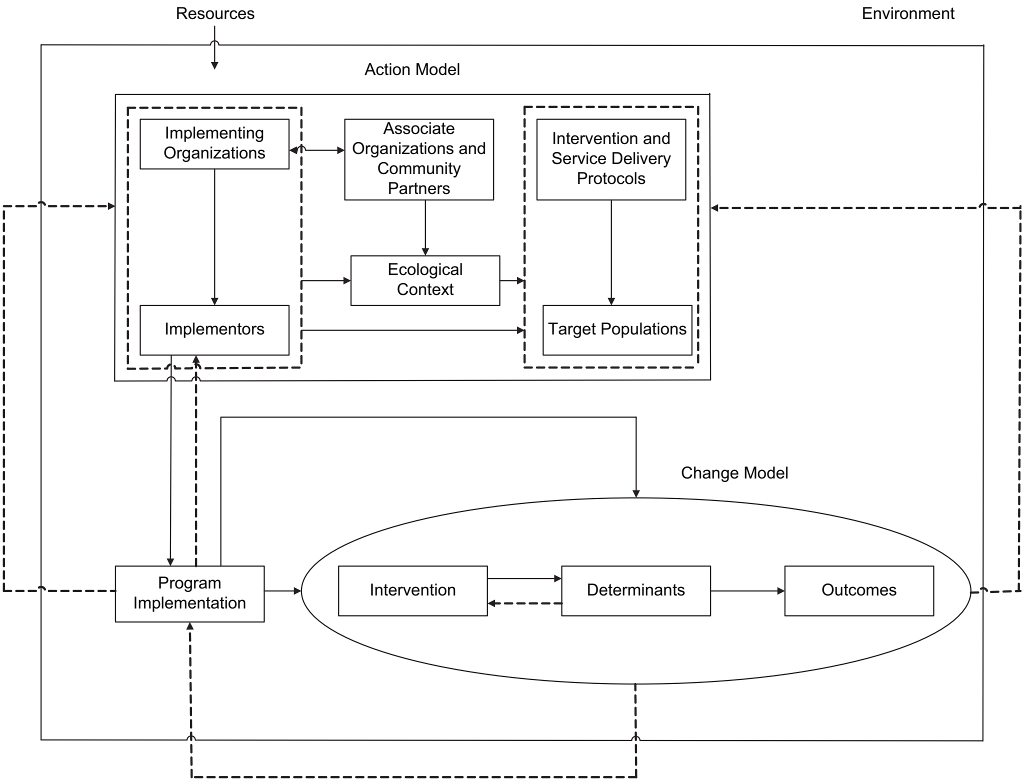

Figure 2. Nonlinear program theory model. Source: Adapted from Chen, H. T. (2005a). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage, p. 31.

physical, and financial). In these types of program theory models, activities are the actions (e.g., training and service delivery) undertaken to bring about a desired end. Outputs are the immediate result of an action (e.g., number of trainings and number of persons trained or who received services). Outcomes are the anticipated changes that occur directly or indirectly as a result of inputs, activities, and outputs. Initial outcomes are usually expressed as changes in knowledge, skills, abilities, and other characteristics (e.g., increased knowledge of safe sexual practices). Intermediate outcomes are often classified as behavioral changes (e.g., increased use of condoms) that are believed to eventually produce changes in long-term outcomes, such as the alleviation, reduction, or prevention of specific social problems or meeting the needs of a program's target population (e.g., reduced incidence of HIV).

In earlier conceptualizations, numerous theorists, including Weiss (1997a, 1997b, 1998) and Wholey (1979), among others, tended to favor linear models to describe program theories. In recent writings, others (e.g., Chen, 2005a, 2005b, 2005c; Rogers, 2008) have advocated for more contextualized, comprehensive, ecological program theory models.[2] As shown in Figure 2, such models diverge considerably from the more simplistic, linear model illustrated in Figure 1. In general, these types of models are intended to integrate systems thinking in postulating program theory, taking contextual and other factors that sometimes influence and operate on program processes and outcomes into account. Even so, these types of theories or models also have been questioned regarding the degree to which they adequately represent complex realities and unpredictable, continuously changing, open and adaptive systems (Patton, 2010).

A crucial aspect of program theory, no matter how it is developed or articulated, is how various components relate to one another (Davidson, 2000, 2005). As such, describing program theory requires an "understanding of how different events, persons, functions, and the other elements represented in the theory are presumed to be related" (Rossi, Freeman, & Lipsey, 1999, pp. 171–172). Most importantly, a program theory should be plausible (i.e., having the outward appearance of truth, reason, or credibility) and stipulate the cause-and-effect sequence through which actions are presumed to produce long-term outcomes or benefits (Donaldson & Lipsey, 2006; Lipsey, 1993).

Donaldson (2001, 2007) has described four potential sources of program theory. These include prior theory and research, implicit theories of those close to the program, observations of the program in operation, and exploratory research to test critical assumptions in regard to a presumed program theory. Patton (2008) favors either deductive (i.e., scholarly theories), inductive (i.e., theories grounded in observation of the program), or user-oriented (i.e., stakeholder-derived theories) approaches to developing program theory for evaluation use. Chen (2005a), however, has principally advocated a stakeholder-oriented approach to program theory formulation, with the evaluator playing the role of facilitator.

Core Principles of Theory-Driven Evaluation

At its core, theory-driven evaluation has two vital components. The first is conceptual, the second empirical (Rogers et al., 2000). Conceptually, theory-driven evaluations should explicate a program theory or model. Empirically, theory-driven evaluations seek to investigate how programs cause intended or observed outcomes. In addition, Chen (2005a, 2005b) distinguished four variants of theory-driven evaluation, depending on which part of the conceptual framework of a program theory the evaluation is focused: theory-driven process evaluation, intervening mechanism evaluation, moderating mechanism evaluation, and integrative process/outcome evaluation. Here, the first three represent options for tailoring theory-driven evaluations so that they are focused only on one aspect, element, or chain of the program theory (Weiss, 2000), rather than comprehensive theory-driven evaluations that are conducted to investigate the whole of the program theory. In earlier writings, Chen (1990) described six types of theory-driven evaluations.[3] Like the former, these too are options for tailored or comprehensive theory-driven evaluations, emphasizing either evaluating the whole of a program theory, particular aspects of it, or evaluating for specific purposes such as to identify variables or factors that moderate or mediate anticipated outcomes or effects.

All in all, the perceived value of theory-driven evaluation is, in part, generating knowledge such as not only knowing whether a program is effective or efficacious (i.e., causal description; *that* a causal relationship exists between A and B) but also explaining a program's underlying causal mechanisms (i.e., causal explanation; *how* A causes B). This knowledge, predominantly aimed at causal generalizations from localized causal knowledge (i.e., interpolation or extrapolation), should provide information useful for decision makers, such as policy formulation (Donaldson, Graham, & Hansen, 1994; Flay et al., 2005; Weiss, 1997a, 1997b, 1998, 2004a). Such evaluations are not always done only for formative or summative purposes (Scriven, 1967) but also for what Chelimsky (1997) and Patton (1997, 2008) refer to as knowledge generation (i.e., "general patterns of effectiveness," Patton, 2008, p. 131). In addition, Rogers (2000) has asserted that the key advantages of theory-driven strategies are that "... at their best, theory-driven evaluations can be analytically and empirically powerful and lead to better evaluation questions, better evaluation answers, and better programs" (p. 209) ... [and they] ... "can lead to better information about a program that is important for replication or for improvement, which is unlikely to be produced by other types of program evaluation." (p. 232).

Shadish, Cook, and Campbell (2002) claim that most theory-driven evaluation approaches share three fundamental characteristics: (a) to explicate the theory of a treatment by detailing the expected relationships among inputs, mediating processes, and short- and long-term outcomes, (b) to measure all of the constructs postulated in the theory, and (c) to analyze the data to assess the extent to which the postulated relationships actually occurred. Nevertheless, "for shorter time periods, the available data may only address the first portion of a postulated causal chain; but over longer periods the complete model could be involved ... [and, therefore] ... the priority is on highly specific substantive theory, high-quality measurement, and valid analysis of multivariate processes ... " (Shadish, Cook, & Campbell, 2002, p. 501). Although some have asserted that the approach requires the application of sophisticated analytic techniques, such as structural equation modeling, to fully evaluate the program theory (Smith, 1994), Donaldson (2007) and Chen (2005a) have stressed that theory-driven strategies and approaches to evaluation are method-neutral, or methodologically pluralistic versus dogmatic, not giving primacy to any particular method, and are therefore equally suited to either quantitative methods, qualitative methods, or both.

Unlike Miller and Campbell's review (2006), which had 10 clearly articulated empowerment evaluation principles against which to assess the degree to which practice adheres to and enacts the underlying principles of empowerment evaluation, the principles used as a framework for this review were derived through a systematic analysis of the approach's central features according to major theoretical writings and writers. These principles were developed both deductively and inductively through an iterative process by carefully cataloguing, analyzing, and reanalyzing the major theoretical descriptions and prescriptions for commonalities as put forth by prominent theorists (e.g., Chen, 1990, 2005a; Donaldson, 2003, 2007; Lipsey et al., 2004; Rogers, 2000, 2007; Rogers et al., 2000; Weiss, 1997a, 1997b, 1998, 2000; 2004a; White, 2009).

Since theory-driven evaluation has no obvious ideological basis, which numerous other forms of evaluation clearly do, and since a wide variety of practitioners would claim to be theory-driven in some capacity (e.g., from those who favor a systems approach to those who believe logic models are the foundation of evaluation), establishing these principles was formidable, given the diffuse development of the approach. Nevertheless, the principles identified and elucidated were vetted by several leading scholars and revised until a reasonable degree of consensus was achieved regarding their organization and content. These principles may neither be exhaustive nor mutually exclusive, and perhaps are an oversimplification, yet they do, however, approximate the most salient features of theory-driven evaluation according to some of the approach's leading theorists, scholars, and practitioners, and provide a limited degree of both conceptual and operational clarity (Miller, 2010; Smith, 1993). The core principles and subprinciples of theory-driven evaluation that arose through this process are illustrated in Table 1. Although some would contest or question the sensitivity and specificity to which some of these principles differentiate theory-driven forms of evaluation from others on the basis that nearly all forms of evaluation engage in similar activities (e.g., formulating evaluation questions), it is how the principles are supposedly enacted that demarcate theory-driven evaluations from others (i.e., in that all are explicitly guided by a program theory). Arguably, specification and application of a program theory is not a necessary condition (e.g., to guide question formulation) for executing many other forms of evaluation (e.g., objectives-oriented evaluation, goal-based evaluation, empowerment evaluation, and responsive evaluation).

Such approaches to or forms of evaluation essentially can be characterized as consisting of five core elements or principles: (a) theory formulation, (b) theory-guided question formulation, (c) theory-guided evaluation design, planning, and execution, (d) theory-guided construct measurement, and (e) causal description and causal explanation, with an emphasis on the latter (i.e., Subprinciples 5.d.i. and 5.d.ii.; causal explanation). At a purely conceptual level, core principles 1, 2, 3, and 4 can be seen as evaluation processes, whereas core principle 5 can be viewed as an evaluation outcome. Even so, some of the specific subprinciples are at the level of general rules of conduct and qualities while others are at the level of methodological action. In addition to using a program theory to guide the execution of an evaluation, it is Principle 5, and its corresponding subprinciples, in

Table 1. Core Principles and Subprinciples of Theory-Driven Evaluation

1. Theory-driven evaluations/evaluators should formulate a plausible program theory
   a. Formulate program theory from existing theory and research (e.g., social science theory)
   b. Formulate program theory from implicit theory (e.g., stakeholder theory)
   c. Formulate program theory from observation of the program in operation/exploratory research (e.g., emergent theory)
   d. Formulate program theory from a combination of any of the above (i.e., mixed/integrated theory)
2. Theory-driven evaluations/evaluators should formulate and prioritize evaluation questions around a program theory
   a. Formulate evaluation questions around program theory
   b. Prioritize evaluation questions
3. Program theory should be used to guide planning, design, and execution of the evaluation under consideration of relevant contingencies
   a. Design, plan, and conduct evaluation around a plausible program theory
   b. Design, plan, and conduct evaluation considering relevant contingencies (e.g., time, budget, and use)
   c. Determine whether evaluation is to be tailored (i.e., only part of the program theory) or comprehensive
4. Theory-driven evaluations/evaluators should measure constructs postulated in program theory
   a. Measure process constructs postulated in program theory
   b. Measure outcome constructs postulated in program theory
   c. Measure contextual constructs postulated in program theory
5. Theory-driven evaluations/evaluators should identify breakdowns, side effects, determine program effectiveness (or efficacy), and explain cause-and-effect associations between theoretical constructs
   a. Identify breakdowns, if they exist (e.g., poor implementation, unsuitable context, and theory failure)
   b. Identify anticipated (and unanticipated), unintended outcomes (both positive and negative) not postulated by program theory
   c. Describe cause-and-effect associations between theoretical constructs (i.e., causal description)
   d. Explain cause-and-effect associations between theoretical constructs (i.e., causal explanation)
      i. Explain differences in direction and/or strength of relationship between program and outcomes attributable to moderating factors/variables
      ii. Explain the extent to which one construct (e.g., intermediate outcome) accounts for/mediates the relationship between other constructs

particular, that distinguishes theory-driven approaches from most other forms of evaluation. The applicability of each of the principles and subprinciples, however, to any given theory-driven evaluation is contingent on a variety of factors such as the nature of the intervention, evaluation purposes, and intended users and uses, for example. Consequently, no claim is made that a high-quality theory-driven evaluation is one in which all or any specific combination of the core principles and subprinciples are applied. These core principles and subprinciples, then, may be viewed as situational rather than absolute criteria.

Criticisms of and Rationales for Increased Use of Theory-Driven Evaluation

Even before the emergence of the theory-driven evaluation movement, Campbell (1984) was skeptical that social science theories could be used to design and evaluate social programs, given poor substantive theories and existing programs that often are watered down to be politically and administratively acceptable. Nevertheless, since its wider acceptance as a legitimate form of evaluation, theory-driven evaluation has been the subject of both pleas for increased use and sharp criticism. Scriven and Stufflebeam, in particular, have been two of the approach's most vocal critics.

In 1998, Scriven published "Minimalist Theory: The Least Theory that Practice Requires" partly in response to what he viewed as some of Chen's (1980, 1990, 1994) logical flaws in proclaiming that program theory should play a more prominent role in evaluation practice. One of Scriven's major premises in that paper was that the role of evaluators is to determine only whether programs work, not to explain how they work. To attempt explanations, Scriven argues, is beyond the capability of most evaluators (e.g., " . . . as in the classic example of aspirin, one may have no theory of how it works to produce its effects, but nevertheless be able to predict its effects and even its side effects—because we found out what they were from direct experimentation. That does not require a theory" [1998, p. 60]), is not part of the evaluator's primary task of determining merit and worth (1991), and "the extra requirement is possession of the correct (not just the believed) logic or theory of the program, which typically requires more than—and rarely requires less than—state-of-the-art subject-matter expertise" (2007, p. 17). To evaluate, he argues, does not require substantive theory about the object being evaluated and that such theories are:

> . . . a luxury for the evaluator, since they are not even essential for explanations, and are not essential for 99% of all evaluations. It is a gross though frequent blunder to suppose that "one needs a theory of learning to evaluate teaching." One does not need to know anything at all about electronics to evaluate electronic typewriters, even formatively, and having such knowledge often adversely affects summative evaluation. (Scriven, 1991, p. 360)

Chen (1994) countered Scriven's (1994) assertions that the consumer-oriented, product model of evaluation—predominately based on the functional properties of an object (e.g., the functional purpose of a watch is to tell time), although this is an oversimplification (see Fournier, 1995)—is the preferred approach to evaluation by arguing that such approaches do not provide adequate information for improving programs and that they do not provide the explanatory knowledge that decision makers sometimes desire or need. To support this position, Chen (1994) stressed that it is equally important to obtain knowledge of how program goals or objectives are attained, not only whether they are. Chen (1994) illustrates the veracity of this claim by stating:

> . . . if a black box evaluation shows a new drug to be capable of curing a disease without providing information on the underlying mechanisms of that cure, physicians will have difficulty prescribing the new drug because the conditions under which the drug will work and the likelihood of negative side effects will not be known. (p. 18)

Relatedly, Stufflebeam (2001) and more recently, Stufflebeam and Shinkfield (2007), in their analysis of evaluation models and approaches against the Joint Committee's (1994) Program Evaluation Standards, commented that "there is not much to recommend about theory-based program evaluation since doing it right is usually not feasible and failed or misrepresented attempts can be highly counterproductive" (Stufflebeam & Shinkfield, 2007, p. 187). Here the primary criticism is that existing, well-articulated, and validated program theories rarely exist and that by engaging in ad hoc theory development and testing, such forms of evaluation expend valuable resources that otherwise could be used more efficiently (Stufflebeam, 2001). They (Stufflebeam & Shinkfield, 2007) also indirectly assert that these types of evaluation sometimes create an intrinsic conflict of interest in that theory-driven evaluators are essentially evaluating the program theory that they developed or played a major role in developing.

Others (Coryn, 2005, 2007, 2008) simply have questioned whether the priority of theory-driven evaluation is evaluating the program itself or the program's underlying theory, mentioned that questions are often descriptive rather than evaluative, are sometimes only tangentially connected to the postulated program theory, and also have expressed more pragmatic concerns regarding the approach's overly abstract principles and procedures. Much like Scriven's (1973) goal-free evaluation, for example, Coryn (2009, September) also has raised concerns that a majority of writers about theory-driven evaluation approaches tell one what to do, but not how to do it.

In addition to the more general criticisms leveled against the approach, certain aspects of theory-driven evaluation also represent a logical incompatibility from a traditional social science perspective (which often is used to guide many theory-driven forms of evaluation), particularly regarding the identification of unanticipated outcomes and side effects not postulated in a program theory. If one perceives a theory, as many social scientists do, as "a set of interrelated constructs, definitions, and propositions that present a systematic view of phenomena by specifying relations among variables, with the purpose of explaining and predicting phenomena" (Kerlinger, 1986, p. 9), then, by definition, a model is a fixed, testable representation of those relationships. From this perspective, then, one cannot logically identify and test unintended outcomes and side effects because they normally are not considered part of the postulated program theory or model.

Contrary to the critics' assertions, however, the value of theory-driven evaluation, to some, remains not only in ascertaining whether programs work but, more specifically, how they work (Chen, 1990, 2005a, 2005b; Donaldson, 2003, 2007; Donaldson & Lipsey, 2006; Mark, Hoffman, & Reichardt, 1992; Pawson & Tilley, 1997; Rogers, 2000, 2007). This feature is not only a hallmark characteristic of theory-driven forms of evaluation that distinguishes it from others, it also is seen as one of its greatest strengths in that such knowledge claims, if they can be constructed, have the potential to be of vital importance to human affairs and policy making, and therefore social betterment (Donaldson, 2007; Donaldson & Lipsey, 2006; Mark et al., 1998, 2000). So, for example, if a program is effective, such approaches should identify which elements are essential for widespread replication. Conversely, if a program fails to achieve its intended outcomes or is ineffective, a theory-driven evaluation should be able to discover whether such breakdowns can be attributed to implementation failure (e.g., treatment integrity; Cordray & Pion, 2006), whether the context is unsuited to operate the mechanisms by which outcomes are expected to occur (Pawson & Tilley, 1997), or simply theory failure (Rogers, 2000; Suchman, 1967; Weiss, 1997b).

Despite the numerous claims put forth by the approach's advocates and critics alike, and its apparent resonance with practitioners, little, if any, systematically derived evidence to justify or falsify the assertions put forth by either position exist. Accordingly, the authors of this review sought to determine the extent to which such claims are congruent or incongruent with evidence manifest in actual case examples of theory-driven evaluations.

Questions Investigated in the Review

Numerous questions related to the enactment of theory-driven evaluation in actual practice were investigated in this study. Broadly, these included: What do theory-driven evaluators do in practice? How closely does their practice reflect the theory (i.e., core principles) of theory-driven evaluation? Based upon these questions and the principles identified as being essential features of theory-driven evaluation, the following specific questions were investigated:

1. In what kinds of settings (e.g., health and education), of what scale (e.g., small and large) and scope (e.g., local, regional, national, and international), with what populations, and for what purposes (e.g., formative, summative, and knowledge generation) are theory-driven evaluations conducted?
2. Why do evaluators and/or their collaborative partners (e.g., evaluation funders and sponsors) choose theory-driven evaluation as their evaluation strategy?
3. How and to what extent are the core principles of theory-driven evaluation enacted in theory-driven evaluation practice?
   a. How do theory-driven evaluators develop program theory and determine the plausibility of those theories?

b. How do theory-driven evaluators develop and prioritize evaluation questions around program theories?
c. How do theory-driven evaluators use program theories for designing, planning, and conducting theory-driven evaluations?
d. How do theory-driven evaluators use program theory in measuring constructs postulated in a program theory?
e. How do theory-driven evaluators use program theory to identify breakdowns, side effects, determine program effectiveness (or efficacy), and explain cause-effect associations between theoretical constructs?

Whereas Questions #1 and #2 are intended to provide contextual information about the sample of cases included in the review by describing some of the general characteristics of theory-driven evaluation practice as exemplified by the identified cases, Question #3 was devised to address how and the frequency with which the core principles of theory-driven evaluation enumerated are enacted in practice.

Method

Sample

A multistage sampling procedure was used to identify and retrieve case examples of theory-driven evaluations in traditional, mainstream scholarly outlets including journals and books. Other sources, including doctoral dissertations, technical reports, background papers, white papers, and conference presentations and proceedings were excluded from the sampling frame. In the first stage (broad scan), potential theory-driven evaluation cases were identified through systematic searches of databases in the social sciences (e.g., ArticleFirst, International Bibliography of the Social Sciences, PsychINFO, Social Work Abstracts, Sociological Abstracts, WorldCat, and Wilson Select Plus), education (e.g., Education Abstracts and ERIC), and health and medicine (e.g., CINAHL and Medline) published between January 1990 and December 2009. Search terms used to identify potential case examples of theory-driven evaluations included, but were not limited to, *theory*, *program*, *logic*, *model*, *driven*, *based*, *guided*, and *evaluation.* Searches were conducted using simple search methods as well as Boolean operators (Reed & Baxter, 2009). The appearance of these terms was searched for in the abstracts, keywords, and bodies of articles and chapters. Throughout, generally accepted standards for locating studies for use in research reviews and synthesis were followed (see Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, 1998; Higgins & Green, 2008).

In an effort to be comprehensive and reduce the number of false negatives (i.e., relevant sources not identified), the sampling strategy was designed to not limit the search only to evaluation-related journals (see Miller & Campbell, 2006) but also to include substantive journals in disciplinary areas and fields of study where cases of theory-driven evaluations might also be found. Additionally, the American Evaluation Association (AEA) Program Theory and Theory-Driven Evaluation Topical Interest Group (TIG) website "Mechanisms" was searched. Following a preliminary relevance screening, this stage yielded an initial population of 161 potential book chapters and articles that self-identified as being directly related to theory-driven evaluation. Copies of all 161 published book chapters and articles identified from the broad scan search were obtained. From these, the reference lists and works cited in each book chapter and article were searched to identify additional book chapters and articles not previously identified. In this stage (the second phase of the broad scan), 44 additional works were identified, for a broad scan sampling frame of 205. The first two stages of the sampling process were not intended to identify specific case examples, but rather to produce a pool of potential chapters and articles that were directly related to theory-driven evaluation, whether theoretical, conceptual, methodological, or otherwise.

In the third sampling stage (initial screening), all 205 articles and chapters were scrutinized to determine if they met inclusion criteria. To be considered for inclusion, the articles and chapters had to (a) explicitly claim that a theory-driven, theory-based, or theory-guided form of evaluation was used and (b) describe a sufficiently detailed case example, including theory formulation, methods, and results. Articles and chapters not meeting these criteria were excluded from the sample. Two reviewers independently screened the 205 articles or chapters to determine whether they met inclusion criteria. Excluded articles and chapters were adjudicated to verify their exclusion, resulting in a total of 90 potential book chapters and articles.

In the fourth stage (final screening), two reviewers worked independently to systematically identify articles or chapters with sufficient or insufficient information for reliable coding. Consensus between both reviewers verified their inclusion or exclusion. The final sample obtained from this procedure yielded 45 articles and chapters describing sufficiently detailed case examples of theory-driven evaluations with satisfactory information for coding.[4] In the 20 years since the appearance of Chen's book *Theory-Driven Evaluations* (1990), an average of 10 ($SD = 5.43$) articles and book chapters directly related to theory-driven evaluation were published per year. Of these, an average of two ($SD = 1.71$) per year were considered adequately detailed and codable case examples for use in this review.

Data Processing and Analysis

An initial coding schema derived from the focal research questions was developed during the final stages of the sampling process. From this procedure, a preliminary structured data abstraction form was created (see Miller & Campbell, 2006). The data abstraction form predominately consisted of fixed items that were binary in nature (i.e., 0 = *absence of the characteristic/trait* or 1 = *presence of the characteristic/trait*). A small number were multiple-selection items (e.g., for theory development there were multiple coding options and coders could code using any combination of codes) and others were open-ended (e.g., populations targeted by programs evaluated, design or method used to support causal inferences).

The data abstraction form predominately consisted of low inference items (i.e., requiring little judgment), whereas others were high inference items (i.e., requiring a greater degree of judgment). Corresponding to the focal research questions, fixed items in the data abstraction form were predominately constructed as they pertained to the specific research questions investigated as well as to what were perceived as observable enactments of the principles and subprinciples enumerated in Table 1. Subprinciple 3.b., for example, was excluded from the data abstraction form, since relevant contingencies related to conducting evaluations (e.g., time and budget) typically are not described in published research reports. To describe how theory-driven principles are enacted in practice (i.e., Research question #3 and its corresponding subquestions), data were coded and analyzed primarily using open coding methods.[5]

In coding fixed items, such as theory formulation (Principle 1), which had four specific elements (i.e., Subprinciples 1.a., 1.b., 1.c., and 1.d.), the text of each article or chapter was searched for information directly pertaining to the particular referent item. In the case of Chen, Weng, and Lin's (1997) evaluation of a garbage reduction program in Taiwan, for instance, the postulated "... program theory comes mainly from hunches and experience of the EPA [Environmental Protection Administration] program designer and the director of the sanitation department of the Nei-fu local government" (p. 29). Here then, "Formulate program theory from existing theory and research" (Subprinciple 1.a.) was coded as 0 (i.e., *absence of the target characteristic/trait*), "Formulate program theory from implicit theory" (Subprinciple 1.b.) was coded as 1 (i.e., *presence of the target* *characteristic/trait*), and "Formulate program theory from observation of the program in operation/exploratory research" (Subprinciple 1.c.) was coded as 0. Subprinciple 1.d., "Formulate program theory from a combination of any of the above," therefore, also was coded as 0, given that a code of 1 on this item required codes of 1 on any combination of two of the other three coding categories (i.e., 1.a. and 1.b., 1.a. and 1.c., or 1.b. and 1.c.).

More complex, open-ended items (i.e., those that required the application of open/substantive coding methods rather than fixed codes and that required a greater degree of inference), while procedurally guided, were fundamentally interpretive and meaning was extracted from text segments that provided information related to the focal research questions (e.g., populations targeted, scale and scope, and research design). In their longitudinal theory-driven evaluation of a regional intervention in Giessen, Germany, which was aimed at changing university students' use of public transportation through a substantially reduced-price bus ticket, Bamberg and Schmidt (1998) obtained multiple, repeated measures of a variety of behavioral and other variables prior to, during, and following two successively introduced and related public transportation interventions between 1994 and 1996. The first intervention (i.e., a reduced-price semester ticket intended to increase students' use of public transportation) was introduced between 1994 and 1995 and the second (i.e., a new circle bus line intended to reduce the time needed when using the public transport system for university purposes) between 1995 and 1996, with multiple pretest and posttest measures. In this case (and using additional information reported in the article), the primary research design was broadly classified as a (short) interrupted time-series design that had two interrupts (i.e., the points at which the first and second interventions were introduced).

Prior to coding of all cases, a calibration procedure, in which each coder worked independently on a small subsample of cases, to familiarize themselves with the coding procedure, was conducted to identify and reduce areas of ambiguity (Wilson, 2009). Following the calibration procedure and modification to the initial data abstraction form after identification of ambiguous items or codes, each chapter and article was randomly assigned to six groups of two coder pairs. Coders first worked independently and later resolved any coding disagreements through a consensus-seeking procedure. Interrater agreement for the independent coding procedure for exact agreement over all fixed items was $p_o = .91$ and accounting for the probability of chance agreements was $\kappa = .87$.

Results

Characteristics of the Sample of Theory-Driven Evaluation Case Examples

Of the 45 cases included in the review, nearly half (47%; $n = 21$) were broadly classified as evaluations of health interventions, with slightly more than one fourth (27%; $n = 12$) being evaluations of educational programs, and the remainder being evaluations of crime and safety, transportation, environmental affairs, and business interventions (27%; $n = 12$) combined. In terms of scale and scope, 31% ($n = 14$) of cases were evaluations of small local programs, and 20% ($n = 9$) were large local programs, 16% ($n = 7$) were small national programs, with the balance being small and large regional and small and large international programs (33%; $n = 15$).[6] Here the distinction between small and large programs (i.e., scope) was one related to the population targeted (i.e., a general population [e.g., all members of a population in a program catchment area] or specific subgroups of a particular population [e.g., injection drug users]), whereas scale was defined by whether a program was local (e.g., covering a single city or county), regional (e.g., covering several counties in a single state or across several states), national (i.e., covering an entire country), or international (i.e., covering more than one country) for a given case.

Populations targeted, whether general or specific, by programs evaluated in the cases were predominately school-age children (31%; $n = 14$). A smaller minority were programs targeted toward college students at 11% ($n = 5$), low income populations and communities at 11% ($n = 5$), adolescents and young adults at 9% ($n = 4$), general populations (i.e., no particular subgroups within a larger population) at 9% ($n = 4$), and other populations (e.g., hospital patients, law enforcement personnel, public health professionals, and small businesses) at 29% ($n = 13$).[7]

Why Evaluators and/or Their Collaborative Partners Select Theory-Driven Evaluation as their Evaluation Strategy

The most frequently occurring motive for selecting a theory-driven evaluation approach was ideological (73%, $n = 33$). In no instance was there sufficient evidence to support conclusions that collaborators (i.e., evaluation funders or sponsors) were directly engaged in the selection of using a theory-driven evaluation approach in the research reports reviewed. Although ideological orientations were the most often occurring rationale, less frequently occurring reasons included guiding program development (13%, $n = 6$), involving stakeholders (11%, $n = 5$), and, in one case, that theory-driven evaluation provided a means for reducing evaluation costs and time relative to other potential approaches (2%, $n = 1$).

In those cases that had an ideological orientation, the evaluators often declared that a theory-driven form of evaluation is more scientifically sound than alternative approaches (e.g., "... contrast this perspective with a purely method-driven approach, in which causal uncertainty is reduced through control exercised during the research design phase, or the statistical modeling approach, whereby control is exercised during the statistical analysis phase via statistical adjustment." Reynolds, 2005, p. 2402). In one form or another, many of the arguments offered to support this position were manifest in claims that theory-based evaluation is one of the only means by which the underlying theoretical propositions of a program or intervention can be systematically evaluated or tested using a scientific method. Another related rationale that emerged, although not explicitly stated, is that theory-driven forms of evaluation are useful for improving internal validity inferences and reducing certain validity threats by permitting tests of more complex causal hypotheses than is typically permissible with many traditional evaluation methods or approaches.

Moreover, as Weiss (1997b) noted, one potential reason, and one that emerged in many of the cases, that evaluators adopt a theory-based approach is that "... the evaluator is also the program developer. A program designer, usually an academic, is engaged in a cycle of program development to deal with a particular problem. He or she develops theory, operationalizes the theory in a set of program activities, tests the program and therefore the underlying theory through evaluation, and revises the intervention." (p. 44). The evidence from this review supports this conclusion in that, across many cases, what is sometimes referred to as theory-driven evaluation are social scientists, rather than practicing evaluators, engaged in testing theoretical propositions and hypotheses derived from their own disciplinary traditions of inquiry as potential solutions to a particular social problem (i.e., "... what I would call applied social psychology." Weiss, 1997b, p. 45).

Enactment of the Core Principles of Theory-Driven Evaluation in Practice

The focal question guiding this review was simply "How and to what extent are the core principles of theory-driven evaluation enacted in practice?" Even so, and due to the very nature of the question, many of the results reported here were derived from categories that were not always discrete/mutually exclusive. As such, any single case potentially could be coded with multiple codes related to a single subquestion. Consequently, the results reported do not always equal, and sometimes exceed, 100%. A summary of the frequency with which theory-driven evaluation principles were enacted in practice corresponding to Table 1 in the cases reviewed is shown in Table 2.

Table 2. Frequency of Enactment of Core Principles and Subprinciples in Theory-Driven Evaluation Practice

Principles and Subprinciples	Number of Cases	Percentage of Cases
1. Theory formulation		
a. Formulate program theory from existing theory and research	41	91%
b. Formulate program theory from implicit theory	22	49%
c. Formulate program theory from observation of the program in operation/exploratory research	6	13%
d. Formulate program theory from a combination of any of the above	19	42%
Subtotal for Principle 1[a]	45	100%
2. Theory-guided question formulation and prioritization		
a. Formulate evaluation questions around program theory	34	76%
b. Prioritize evaluation questions	10	22%
Subtotal for Principle 2[b]	9	20%
3. Theory-guided planning, design, and execution		
a. Design, plan, and conduct evaluation around a plausible program theory	23	51%
b. Design, plan, and conduct evaluation considering relevant contingencies[c]	—	—
c. Determine whether evaluation is to be tailored or comprehensive[d]	26 (19)	58% (42%)
Subtotal for Principle 3[b]	23	51%
4. Theory-guided construct measurement		
a. Measure process constructs postulated in program theory[d]	20 (11)	45% (22%)
b. Measure outcome constructs postulated in program theory[d]	22 (13)	49% (29%)
c. Measure contextual constructs postulated in program theory	16	36%
Subtotal for Principle 4[e]	14	31%
5. Identification of breakdowns and side effects, effectiveness or efficacy, and causal explanation		
a. Identify breakdowns	27	60%
b. Identify outcomes not postulated by program theory	8	18%
c. Describe cause-and-effect associations between theoretical constructs	37	82%
d. Explain cause-and-effect associations between theoretical constructs		
i. Explain differences in direction and/or strength of relationship between program and outcomes	24	53%
ii. Explain the extent to which one construct accounts for/mediates the relationship between other constructs	30	67%
Subtotal for Principle 5[b]	6	13%

Note. [a]Enactment of any form of theory formulation was counted for the subtotal. [b]Subtotal includes only those cases that enacted all of the measured subprinciples. [c]Not measured due to insufficient information provided in most of the research reports. [d]The number of cases and percentage of cases for tailored evaluations (i.e., those that only evaluated a specific part of the program theory) are shown in parentheses. [e]The number of cases and percentage of cases counted for the subtotal includes comprehensive and tailored evaluations that enacted the relevant subprinciples (e.g., comprehensive evaluations were counted only if they enacted all subprinciples).

Core principle 1: Theory formulation. Program theories in the cases studied were principally deductive in origin and derived from existing scientific theory (e.g., "… rooted in several health behavior theories, including the health belief model, social cognitive theory, the transtheoretical model, and the theory of reasoned action." Umble, Cervero, Yang, & Atkinson, 2000, p. 1219). Inductive theories and assumptions held by stakeholders as well as theories derived through program observation were far less common for articulating and specifying program theory as a singular method. Bickman (1996), for example, applied a variety of methods including interviews, document reviews, and focus groups to develop a comprehensive, detailed, and logical theory. Nearly half of the cases reviewed applied some combination of deductive and inductive methods to formulate program theory. In these cases, theories that originated both from existing research and from program stakeholders often were heuristically synthesized to devise a plausible program theory for evaluation use. Generally, the methods applied comport well and share many features with those illustrated by Leeuw (2003) for retrospectively reconstructing program theory, although not all theory formulation was a post hoc activity following implementation of a program, and in a minority of cases preceded implementation.

The plausibility of specified theories most often was ascertained by means of simple validity checks such as face and content validity. Given some of the assertions made by Weiss (1997b) that program theories based only on such assumptions are very often overly simplistic, partial, or even categorically erroneous, several potential complications regarding specification error related both to the stated theory and ensuing evaluation can be raised. Furthermore, alternative theories were rarely considered in the cases reviewed. In the few cases that explored rival theories, these were generally investigated using statistical techniques to identify variables or other factors (e.g., nonsignificant path coefficients, correlated error terms, and overly large residuals) that significantly contributed to model fit or misfit (e.g., R^2 in a regression framework and numerous goodness-of-fit indices in a structural equation modeling context such as X^2, X^2/df ratio, goodness-of-fit index, adjusted goodness-of-fit index, comparative fit index, root mean square error of approximation) in order to make adjustments or modifications to the theoretically specified model (e.g., Bamberg & Schmidt, 2001), rather than true alternatives or competing theories (e.g., stakeholder-derived theories vs. theories arising from prior empirical research).

Core principle 2: Theory-guided question formulation and prioritization. Uniformly, questions investigated in the cases were of a descriptive variety (e.g., "What are individual experiences of the impact of a fungating wound on daily life?" Grocott & Cowley, 2001, p. 534) rather than evaluative questions regarding an intervention's merit or worth (e.g., "So what? What does this tell us about the value of the program?" Davidson, 2007, p. iv). Such questions often appeared in the form of null and alternative hypotheses (e.g., "H1: Participants who receive media literacy training will exhibit a higher level of reflective thinking than participants who did not receive media literacy training." Pinkelton, Austin, Cohen, Miller, & Fitzgerald, 2007, p. 25). Almost universally, coupling of questions to the identified theory was enacted by articulating testable hypotheses derived from the program's underlying logic or theoretical foundations. In terms of prioritizing evaluation questions, a small minority of cases indicated that questions or hypotheses were prioritized due to logistical constraints whereas others prioritized evaluation questions according to predetermined funder or sponsor information needs. Donaldson and Gooler (2003), for example, reported that which questions to answer and how to answer them were determined collaboratively with the evaluation sponsor given resource and other practical constraints. In the majority of cases, however, question prioritization was not explicitly stated.

Core principle 3: Theory-guided planning, design, and execution. In many of the cases reviewed, the explication of a program theory was not perceptibly used in any meaningful way for conceptualizing, designing, or executing the evaluation reported and easily could have been accomplished using an alternative evaluation approach (e.g., goal-based or objectives-oriented). Sato (2005), for example, seemingly expended considerable effort developing a theoretical framework for Japan's foreign student policy toward Thailand and yet essentially evaluated only the degree to which policy objectives were met. In others, however, the specified theory perceptibly was more vital to the planning, design, and execution of the evaluation. For instance, in their evaluation of a gaming simulation as teaching device, Hense, Kriz, and Wolfe (2009) used the underlying theory to guide measurement of constructs specified in the program theory, including exogenous factors, and to design methods for examining the relationships between program processes and outcomes, among others. Even so, as both Rogers (2007) and Weiss (1997b) have observed, and consistent with the cases reviewed, it is not uncommon for evaluators not to use the theory to guide the evaluation. Too often, " . . . the ways program theory are used to guide evaluation are often simplistic . . . [and] . . . consists only of gathering evidence about each of the components in the logic model, and answering the question "Did this happen?" about each one." (Rogers, 2007, p. 65). Additionally, very little information was provided to justify decisions when only particular aspects of a theory were evaluated, such as a single causal chain, rather than the whole of the specified theory (i.e., tailored theory-driven evaluations vs. comprehensive theory-driven evaluations).

Core principle 4: Theory-guided construct measurement. Methods used to measure theory-related or derived processes, outcomes, and exogenous factors (e.g., contextual and environmental) directly associated with the specified program theory or model varied widely (e.g., from secondary or extant data to interviews with program participants to document analysis to large-scale self-report surveys). Very often, samples of theoretical constructs, and their respective targets of generalization, were obtained from poorly devised measures intended to represent broader, more complex, latent constructs. It was not uncommon for single indicator measures, which likely account for little of the variance in the target latent construct, to be the primary means of construct measurement. Mole, Hart, Roper, and Saal (2009), for example, used simple, easily obtained measures of sales and employee growth as proxy indicators of small business productivity as opposed to operationally defining productivity and using the operational definition to support direct theoretical construct measurement. Others, such as Weitzman, Mijanovich, Silver, and Brecher (2009), however, used very refined, sometimes standardized, measures of theoretically derived latent constructs such as neighborhood quality of life, to evaluate a citywide health initiative. Noticeably, few of the cases reviewed reported reliability coefficients and even fewer reported validity coefficients or other information pertaining to the precision and accuracy of information as related to samples of latent or observed constructs or their qualitative counterparts (e.g., trustworthiness, dependability, and confirmability).

Core principle 5: Identification of breakdowns and side effects, determining program effectiveness or efficacy, and causal explanation. More than half of the cases reviewed identified breakdowns in the program theory (e.g., "There were two missing links in the BINP [Bangladesh Integrated Nutrition Project] chain: the first was the relative neglect of some key decision makers regarding nutritional choices . . . and the second the focus on pregnancy weight gain rather than pre-pregnancy nutritional status." White & Masset, 2007, pp. 647–648). Despite the urgings of a majority of theoretical writers, however, fewer searched for unanticipated or unintended outcomes or side effects, whether positive or negative, not specified in the formulated program theory or model. Chen et al. (1997) were one of few exceptions, however, in observing that " . . . as a reaction to the intervention, residents simply saved the same volume of Tuesday garbage and disposed it at the collection sites on Wednesday." (p. 41), as part of their evaluation of a garbage reduction program in Taiwan.

Of all of the characteristics associated with theory-driven evaluation, none provide greater conceptual clarity than subprinciples 5.c. and 5.d. Although advocates of alternative forms of evaluation generally recognize the importance of causal attribution, only those who favor theory-driven forms of evaluation specifically emphasize causal explanation and the mechanisms by which suspected causes produce their effects. By far, mixed-method designs were the most commonly used (42%; $n = 19$) for supporting descriptive causal inferences. These were followed by pretest–posttest designs with nonequivalent control groups at 13% ($n = 6$), randomized controlled trials at 11% ($n = 5$), other types of research designs (e.g., case studies, one-group post-test only designs) at 9% ($n = 4$), one-group pretest–posttest designs at 9% ($n = 4$), interrupted time-series designs at 7% ($n = 3$), qualitative studies at 4% ($n = 2$), and insufficiently/poorly described designs at 4% ($n = 2$).

More importantly, however, and in relation to subprinciples 5.d.i and 5.d.ii, specifically, a large proportion of the cases included in the review investigated either moderators (53%, $n = 24$; e.g., subject characteristics, treatment dosage variations), mediators (67%, $n = 30$; e.g., knowledge or skill acquisition, observable behaviors, and their relationship to other outcomes), or, in nearly half of cases, both (47%, $n = 21$), in an attempt to more fully explicate simple main causal effects. In these cases, causal mechanisms posited in program theories mainly were investigated through direct statistical tests (e.g., "... to test the causal structure postulated ... which contains a chain of mediating causal variables, structural equation modeling ... [was used]" Bamberg & Schmidt, 2001, p. 1308) and less frequently using causal tracing and pattern matching techniques (e.g., "... examining the outcome data in light of what the data on program implementation predicted revealed a mismatch between the expected pattern and the data." Cooksy, Gill, & Kelly, 2001, p. 127). Nonetheless, the coding process did not include a means for examining the warrants or backings (Fournier, 1995) used to support causal inferences in the cases studied, only whether such conclusions were present or absent according to the study's authors. Consequently, and even though a large majority of the cases described and explained cause and effect relationships, no claims are made as to the quality of evidence supporting those conclusions.

Discussion

With the exception of empowerment evaluation (Miller & Campbell, 2006), participatory evaluation (Cousins & Whitmore, 1998; Cullen, Coryn, & Rugh, 2010; Weaver & Cousins, 2004), evaluation standards and their application for metaevaluation (Wingate, Coryn, Gullickson, & Cooksy, 2010), and evaluation use (Brandon & Singh, 2009; Cousins & Leithwood, 1986; Johnson et al., 2009; Shulha & Cousins, 1997), among others, very little empirical evidence exists to buttress the numerous theoretical postulations and prescriptions put forth for most evaluation approaches, including theory-driven forms of evaluation. Yet, for many years, evaluation scholars have urged the evaluation community to carry out empirical studies to scrutinize such assumptions and to test specific hypotheses about evaluation practice (Alkin & Christie, 2005; Christie, 2003; Henry & Mark, 2003; Mark, 2007; Shadish, Cook, & Leviton, 1991; Smith, 1993; Stufflebeam & Shinkfield, 1985, 2007; Worthen, 2001; Worthen & Sanders, 1973).

Although predominately descriptive, this review does provide valuable insight into what has otherwise principally consisted of anecdotal reports regarding one form of evaluation theory and practice. By most accounts, the number of studies on evaluation theories and their enactment in practice is small and such studies have been the exception rather than the norm. In recent years, however, a renewed interest in research on evaluation theories and methods as well as a surge of investigations related to relationships between practice and theory, largely led by Christie, have transpired. These have included, among others, the formation of the AEA Research on Evaluation TIG in 2007, surveys of AEA members (Fleischer & Christie, 2009), a bibliometric analysis of evaluation theorists' published works (Heberger, Christie, & Alkin, 2010), a study of decision-making contingencies related to evaluation design (Tourmen, 2009), research on how evaluation data influences decision makers' actions (Christie, 2007), several systematic reviews and research syntheses (Brandon & Singh, 2009; Chouinard & Cousins, 2009; Johnson et al., 2009; Miller & Campbell, 2006; Trevisan, 2007), and a recent collection of papers on advances in evaluating evaluation theory published in the *American Journal of Evaluation* (Smith, 2010).

Implications

The evidence resulting from this review to repudiate or substantiate many of the claims put forth by critics of and advocates for theory-driven forms of evaluation is, at best, modest, and in some instances conflicting. Support for Scriven's (1991, 1994, 1998) assertions that (stakeholder- or substantively-derived) theory is not a necessary condition for conducting an evaluation, or conversely, the counter arguments put forth by Chen (1990, 1994), is, for example, mixed. In many of the cases reviewed, the explication of a program theory unmistakably was unnecessary, or almost an afterthought in some instances, and was not visibly used in any meaningful way for formulating or prioritizing evaluation questions nor for conceptualizing, designing, conducting, interpreting, or applying the evaluation reported. In these cases, from a methodological perspective, such evaluations very likely would have produced the same results and conclusions even in the absence of articulating or expressing an underlying theory. In other cases, however, the explication of a plausible program theory noticeably was essential to the planning, design, and execution of the evaluation (see Donaldson & Gooler, 2002, 2003). Nevertheless, both Scriven and Chen's positions are fundamentally ideological, and, therefore, cannot logically be tested in any replicable, meaningful way.

As for Stufflebeam's (2001; Stufflebeam & Shinkfield, 2007) criticisms regarding the propriety, utility, feasibility, and accuracy of theory-driven forms of evaluation against the Joint Committee's Program Evaluation Standards (1994), evidence derived from this review is more compelling. In no instance was there dependable confirmation to support contentions that theory-driven evaluations are more or less proper, useful, feasible, or accurate than other forms of evaluation. Many of Stufflebeam's (2001; Stufflebeam & Shinkfield, 2007) other condemnations (e.g., problems associated with engaging in ad hoc theory development, testing, and validation, resource waste) also could not be validated or invalidated satisfactorily due to the nature of the content reported in the majority of studies reviewed. In a small minority of cases, however, what is sometimes referred to as theory-driven evaluation can be more accurately characterized as social scientists, rather than practicing evaluators, testing theoretical propositions and hypotheses derived from their own disciplinary traditions of inquiry (e.g., public health, psychology, and sociology; see Weiss, 1997b) and, therefore, providing some support for Stufflebeam's (2001; Stufflebeam & Shinkfield, 2007) claims regarding potential conflicts of interest. In these instances, such evaluations appeared to be curiosity-driven research endeavors rather than truly evaluative inquiry (Davidson, 2007), and not clearly conducted with the intent to serve any immediate or tangible stakeholder information needs (Patton, 1997, 2008).

Similarly, some of the apprehensions expressed by Coryn (2005, 2007, 2008) regarding the priority of theory-driven evaluation (i.e., evaluating the theory underlying a program rather than the program itself, descriptive questions vs. evaluative questions) could not be falsified or justified definitively. Some of the evidence derived from this review, though, supports this position, depending upon how the overarching function or purpose of evaluation is viewed (e.g., to determine merit or worth, to describe, and to explain).

In terms of methodological implications, the incompatibility of theory-driven evaluation (as regards exogenous and endogenous constructs not specified as part of an a priori model or theory) from the standpoint of many traditional social scientists, largely remains a question for philosophers of science. Like the former assertions, these too are ideological assumptions that are exceptionally difficult to substantiate. However, Donaldson (2003) provides reasonable evidence and an empirically derived rationale for continued use and development of a theory-driven method of evaluation, supported by case examples, in opposition to many of the assertions put forth by some of the approach's antagonists.

Finally, and more generally, the results of this review also suggest that additional exemplars of theory-driven evaluations, including reports of successes and failures, methods and analytic techniques, and evaluation outcomes and consequences, are seriously needed in the published literature. Although several exemplars do exist, the number of case examples clearly documenting and recounting how the approach is enacted, procedures and analytic frameworks, and the subsequent uses of evaluation results is surprisingly low, even spanning two decades, and, despite the apparent recognition and influence of the approach. Unexpectedly, and given the already large, constantly growing, theoretical and methodological literatures, few of the cases reviewed applied more than a nominal number of the prescribed theory-driven evaluation principles in practice.

Limitations

Although theory-driven forms of evaluation are widely discussed in the evaluation literature, at professional development offerings, at meetings of specialized associations and societies, and informally on listservs, actual case examples are sparse. Despite the scarcity of case examples, and even though the number of cases identified and included in the current study substantially exceed those analyzed by Birckmayer and Weiss (2000), the degree to which the sample reviewed is congruent with and representative of the theoretical population of all potential case examples is unknown, given the nature of the sampling design and procedure as well as the search terms used to identify studies. Moreover, the inclusion criteria used for the review were intentionally narrow. Numerous investigations in applied health and health promotion (e.g., those using the health belief model, theory of reasoned action, social leaning theory, and diffusion of innovations theory) and international development, for example, therefore, likely were excluded. Relatedly, and corresponding to one of the caveats reported by Miller and Campbell (2006), the studies included in this review were those that self-identified as being theory-driven evaluations.

In addition to the search terms and inclusion criteria, one major source of potential bias is simply the fact that evaluators often do not have intellectual property rights to the data gathered as part of an evaluation. Many evaluation sponsors can, and often do, prohibit distribution and publication of evaluation findings (Henry, 2009). Another potential source of bias is the exclusion of grey and fugitive literatures and non–English-language book chapters and journal articles as well as unpublished examples such as might be found in doctoral dissertations, technical reports and white papers, and conference presentations and proceedings (Johnson et al., 2009; Reed & Baxter, 2009).

The veracity of the core principles of theory-driven evaluation developed for this study also is a likely source of bias, although of a different type. This bias is particularly evident in terms of the validity of the key tenets intended to represent theory-driven evaluation—that is, the degree to which the core principles reflect or represent both the conceptual and the operational specificity of the construct or phenomena that is believed to characterize or embody theory-driven evaluation (Miller, 2010). Theory-driven evaluation does not have an easily identifiable ideological basis like that of empowerment evaluation (Miller & Campbell, 2006), participatory evaluation (Cousins & Earl, 1992), or utilization-focused evaluation (Patton, 1997, 2008), for example, making distillation of the approach difficult. Nonetheless, a sample of writers about and scholars of theory-driven evaluation confirmed that the core principles established for the study accurately reflect what they perceive as being the key tenets of the approach.

Finally, a large amount of human judgment was involved in this review, and human judgment is fallible (Borenstein et al., 2009). That being said, in an effort to reduce potential biases and systematic errors, as well as increase replicability, the method applied, including the sampling and coding procedures (e.g., multiple sampling stages, clearly defined inclusion criteria, predominantly fixed items [i.e., either a trait or characteristic was evident or it was not] for coding, independent coders, random assignment of case examples to coder pairs, consensus seeking in instances of disagreement), was designed to diminish the likelihood that coders consciously or unconsciously sought confirming or disconfirming evidence in relation to the assertions put forth by either the approach's critics or advocates.

Future Research

As Smith (1993) observed nearly two decades ago, "if evaluation theories cannot be uniquely operationalized, then empirical tests of their utility become increasingly difficult ... [and] ... if alternative theories give rise to similar practices, then theoretical differences may not be practically significant" (p. 240). Certainly then, further research is necessary to determine the extent to which the principles enumerated here are authentic and adequately reflect a common set of core principles that are capable of discriminating between different modes and manifestations of theory-driven evaluation (e.g., logical frameworks, theory-of-change, and outcomes hierarchies) and their resultant implications. Likewise, application of Mark's (2007) framework for research on evaluation and Miller's (2010) standards for empirical investigations of evaluation theory would be of great pragmatic and theoretical benefit in formulating these questions or disputing the results of this review. Such investigations might involve, for example, responding to questions such as "Does constructing a logic model and using the specified model as an operational framework constitute a true theory-driven evaluation?" "What consequences occur, or could possibly occur, as a result of theory misspecification?" "Do decision makers and other stakeholders place greater value on explanations of how a program works than on conclusions only about whether a program works?" or "How is explanatory information used in making decisions about programs?" Such investigations, should they be undertaken, ought to emphasize evaluation consequences rather than simple descriptive questions regarding how theory-driven evaluation is implemented in practice (Henry & Mark, 2003; Mark, 2007).

Acknowledgments

The authors are grateful to E. Brooks Applegate, Katrina L. Bledsoe, Christina A. Christie, J. Bradley Cousins, Lois-Ellin Datta, Stewart I. Donaldson, Jonny A. Morell, Michael Q. Patton, Patricia J. Rogers, James R. Sanders, Michael Scriven, Daniel L. Stufflebeam, Carol H. Weiss, and the four blind peer reviewers for their comments and suggestions on earlier drafts of this article. This review benefited greatly from their individual and collective wisdom.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: Interdisciplinary PhD in Evaluation (IDPE) program at Western Michigan University (WMU).

Notes

1. Although many of these concepts share common characteristics, they also differ in important ways. However, a detailed discussion of these similarities and differences exceeds the scope of this review. Interested readers are referred to Rogers (2007), Rogers et al. (2000), and Weiss (1997a, 1997b). Moreover, the distinctions between program theories (i.e., treatment theories; see Lipsey, 1993) and substantive theories (e.g., more generalized biological or social theories) and their use in evaluation have not been clearly elucidated and are, therefore, debatable and not a central focus of the current study.
2. An explanation of these types of program theory models exceeds the scope of this review. Interested readers are referred to Chen (2005a, 2005b, 2005c).
3. Detailed descriptions of these typologies exceed the scope of this review and interested readers are referred to Chen (1990, 2005a, 2005b) and Donaldson (2007) for a more complete discussion.
4. The studies included in this review can be found in this article's References and are proceeded by an asterisk (*) as is standard practice for systematic reviews. In some instances findings on the same study were reported in multiple publications. In these cases, to avoid duplication in the sample, only one was selected for inclusion.

5. As is sometimes practiced in narrative reviews, relevant scholarly literatures and assumptions are incorporated throughout the presentation of results in order to support and clarify interpretations of data, as well as to lend credibility to those interpretations.
6. Reported percentages do not always total 100% due to rounding error.
7. In many of the cases reviewed, several populations were sometimes targeted (e.g., low income school-age children). However, each case was coded only in a single category reflecting the primary population of interest.

References

Alkin, M. C., & Christie, C. A. (Eds.). (2005). *Theorists' models in action. New Directions for Evaluation*, *No. 106*. San Francisco, CA: Jossey-Bass.

*Bamberg, S. (2006). Is a residential relocation a good opportunity to change people's travel behavior? Results from a theory-driven intervention study. *Environment and Behavior*, *38*, 820-840.

*Bamberg, S., & Schmidt, P. (1998). Changing travel-mode choice as rational choice: Results from a longitudinal intervention study. *Rationality and Society*, *10*, 223-252.

*Bamberg, S., & Schmidt, P. (2001). Theory-driven subgroup-specific evaluation of an intervention to reduce private car use. *Journal of Applied Social Psychology*, *31*, 1300-1329.

Bickman, L. (Ed.). (1987). Using program theory in evaluation. *New Directions for Program Evaluation, No. 33*. San Francisco, CA: Jossey-Bass.

*Bickman, L. (1996). The application of program theory to the evaluation of a managed mental health care system. *Evaluation and Program Planning*, *19*, 111-119.

Birckmayer, J. D., & Weiss, C. H. (2000). Theory-based evaluation practice: What do we learn? *Evaluation Review*, *24*, 407-431.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* West Sussex, UK: John Wiley.

Brandon, P. R., & Singh, J. M. (2009). The strength of the methodological warrants for the findings of research on program evaluation use. *American Journal of Evaluation*, *30*, 123-157.

Campbell, D. T. (1984). Hospital and landsting as continuously monitoring social programs: Advocacy and warning. In B. Cronholm & L. Von Knorring (Eds.), *Evaluation of mental health service programs* (pp. 13-39). Stockholm, Sweden: Forskningsraadet Medicinska.

*Carvalho, S., & White, H. (2004). Theory-based evaluation: The case of social funds. *American Journal of Evaluation*, *25*, 141-160.

*Chang, F., & Munoz, M. A. (2006). School personnel educating the whole child: Impact of character education on teachers' self-assessment and student development. *Journal of Personnel Evaluation in Education*, *19*, 35-49.

Chelimsky, E. (1997). The political environment of evaluation and what it means for the development of the field. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 53-68). Thousand Oaks, CA: Sage.

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. *American Journal of Evaluation*, *19*, 35-55.

Chen, H. T. (1980). The theory-driven perspective [Special issue]. *Evaluation and Program Planning*, *12*, 299-306.

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

Chen, H. T. (1994). Theory-driven evaluation: Needs difficulties and options. *Evaluation Practice*, *15*, 79-82.

Chen, H. T. (2005a). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness.* Thousand Oaks, CA: Sage.

Chen, H. T. (2005b). Theory-driven evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 415-419). Thousand Oaks, CA: Sage.

Chen, H. T. (2005c). Program theory. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 340-342). Thousand Oaks, CA: Sage.

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

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

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

Chen, H. T. & Rossi, P. H. (Eds.). (1992). *Using theory to improve program and policy evaluations.* Santa Barbara, CA: Greenwood Press.

*Chen, H. T., Weng, J. C. S., & Lin, L.-H. (1997). Evaluating the process and outcome of a garbage reduction program in Taiwan. *Evaluation Review*, *21*, 27-42.

*Cho, H., & Witte, K. (2005). Managing fear in public health campaigns: A theory-based formative evaluation process. *Health Promotion Practice*, *6*, 482-490.

Chouinard, J. A., & Cousins, J. B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation*, *30*, 457-494.

Christie, C. A. (2003). What guides evaluation practice? A study of how evaluation practice maps onto evaluation theory. In C. A. Christie (Ed.), *The theory-practice relationship in evaluation* (pp. 7–36). *New Directions for Evaluation, No. 97.* San Francisco, CA: Jossey-Bass.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions. *American Journal of Evaluation*, *28*, 8-25.

*Cole, M. (2003). The health action zone initiative: Lessons from Plymouth. *Local Government Studies*, *29*, 99-117.

Conlin, S., & Stirrat, R. L. (2008). Current challenges in development evaluation. *Evaluation*, *14*, 193-208.

*Conner, R. F., Mishra, S. I., & Lewis, M. A. (2004). Theory-based evaluation of AIDS-related knowledge, attitudes, and behavior changes. *New Directions for Program Evaluation*, *46*, 75-85.

*Cook, T. D., Habib, F., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's school development program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal*, *36*, 543-597.

*Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal*, *37*, 535-597.

*Cooksy, L. J., Gill, P., & Kelly, P. A. (2001). The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning*, *24*, 119-128.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 103-124). Washington, DC: American Psychological Association.

Coryn, C. L. S. (2005). Practical program evaluation: Assessing and improving planning, implementation, and effectiveness [Review of the book *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*, by H. T. Chen]. *American Journal of Evaluation, 26*, 405–407.

Coryn, C. L. S. (2007). *Evaluation of researchers and their research: Toward making the implicit explicit* (Unpublished doctoral dissertation). Western Michigan University, Kalamazoo, MI, USA.

Coryn, C. L. S. (2008). Program theory-driven evaluation science [Review of the book *Program theory-driven evaluation science,* by S. I. Donaldson]. *American Journal of Evaluation, 29*, 215–220.

Coryn, C. L. S. (2009, September). *Contemporary trends and movements in evaluation: Evidence-based, participatory and empowerment, and theory-driven evaluation.* Paper presented at The Evaluation Center's Evaluation Café, Kalamazoo, MI.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, *14*, 397-418.

Cousins, J. B., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research*, *56*, 331-364.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. In E. Whitmore (Ed.), *Understanding and practicing participatory evaluation* (pp. 5–23). *New Directions for Evaluation, No. 80.* San Francisco, CA: Jossey-Bass.

*Crew, R. E. Jr., & Anderson, M. R. (2003). Accountability and performance in charter schools in Florida: A theory-based evaluation. *American Journal of Evaluation*, *24*, 189-212.

Cullen, A., Coryn, C. L. S., & Rugh, J. (2010). *A study of international development evaluators' perceptions of reasons for and the politics and consequences associated with participatory evaluation approaches.* Manuscript submitted for publication.

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

Davidson, E. J. (2005). *Evaluation methodology basics: The nuts and bolts of sound evaluation.* Thousand Oaks, CA: Sage.

Davidson, E. J. (2007). Unlearning some of our social scientists habits. *Journal of MultiDisciplinary Evaluation*, *4*, iii-vi.

Donaldson, S. I. (2001). Mediator and moderator analysis in program development. In S. Sussman (Ed.), *Handbook of program development for health behavior research* (pp. 470-496). Newbury Park, CA: Sage.

Donaldson, S. I. (2003). Theory-driven program evaluation in the new millennium. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 109-141). Mahwah, NJ: Lawrence Erlbaum.

Donaldson, S. I. (2007). *Program theory-driven evaluation science.* New York, NY: Lawrence Erlbaum.

Donaldson, S. I., & Gooler, L. (2002). Theory-driven evaluation of the Work and Health Initiative: A focus on winning new jobs. *American Journal of Evaluation*, *23*, 341-346.

*Donaldson, S. I., & Gooler, L. (2003). Theory-driven evaluation in action: Lessons from a $20 million statewide work and health initiative. *Evaluation and Program Planning*, *26*, 355-366.

Donaldson, S. I., Graham, J. W., & Hansen, W. B. (1994). Testing the generalizability of intervening mechanism theories: Understanding the effects of adolescent drug use prevention interventions. *Journal of Behavioral Medicine*, *17*, 195-216.

Donaldson, S. I., & Lipsey, M. W. (2006). Roles for theory in contemporary evaluation practice: Developing practical knowledge. In I. Shaw, J. C. Greene, & M. M. Mark (Eds.), *The handbook of evaluation: Policies, programs, and practices* (pp. 56-75). London, UK: Sage.

*Farber, M., & Sabatino, C. (2007). A therapeutic summer weekend camp for grieving children: Supporting clinical practice through empirical evaluation. *Child Adolescent Social Work Journal*, *24*, 385-402.

Flay, B. R., Biglan, A., Boruch, R. F., González Castro, F., Gottfredson, D., Kellam, S., ... Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151-175.

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: Results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation*, *30*, 158-175.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps* (pp. 15-32). New Directions in Evaluation, No. 68. San Francisco, CA: Jossey-Bass.

Frechtling, J. A. (2007). *Logic modeling methods in program evaluation.* San Francisco, CA: Jossey-Bass.

*Frosch, D. L., Legare, F., & Mangione, C. M. (2008). Using decision aids in community-based primary care: A theory-driven evaluation with ethnically diverse patients. *Patient Education and Counseling*, *73*, 490-496.

Funnel, S. C. (1997). Program logic: An adaptable tool. *Evaluation News & Comment*, *6*, 5-17.

Government Accountability Office. (2009). *A variety of rigorous methods can help identify effective interventions (GAO Publication No. GAO-10–30).* Washington, DC: U. S. Government Printing Office.

*Grocott, P., & Cowley, S. (2001). The palliative management of fungating malignant wounds—Generalising from multiple-case study data using a system of reasoning. *International Journal of Nursing Studies*, *38*, 533-545.

Gugiu, P. C., & Rodriguez-Campos, L. (2007). Semi-structured interview protocol for constructing logic models. *Evaluation and Program Planning*, *30*, 339-350.

Heberger, A. E., Christie, C. A., & Alkin, M. C. (2010). A bibliometric analysis of the academic influences of and on evaluation theorists' published works. *American Journal of Evaluation*, *31*, 24-44.

*Heflinger, C. (1996). Implementing a system of care: Findings from the Fort Bragg evaluation project. *Journal of Mental Health Administration*, *23*, 16-29.

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

*Hendershott, A., & Norland, S. (1990). Theory-based evaluation: An assessment of the implementation and impact of an adolescent parenting program. *Journal of Applied Sociology*, *7*, 35-48.

Henry, G. T. (2009). When getting it right matters: The case for high-quality policy and program impact evaluations. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 32-50). Thousand Oaks, CA: Sage.

Henry, G. T., & Mark, M. M. (2003). Toward an agenda for research on evaluation. In C. A. Christie (Ed.), *The practice-theory relationship in evaluation* (pp. 69-80). New Directions for Evaluation, No. 97. San Francisco, CA: Jossey-Bass.

*Hense, J., Kriz, W. C., & Wolfe, J. (2009). Putting theory-oriented evaluation into practice: A logic model approach for evaluating SIMGAME. *Simulation & Gaming*, *40*, 110-133.

Higgins, J. P. T., & Green, J. (Eds.). (2008). *Cochrane handbook for systematic reviews of interventions.* West Sussex, UK: John Wiley.

*Janssens, F. J. G., & de Wolf, I. F. (2009). Analyzing the assumptions of a policy program: An ex-ante evaluation of "educational governance" in the Netherlands. *American Journal of Evaluation*, *30*, 411-425.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, *30*, 377-410.

*Johnson, K., Young, L., Foster, J. P., & Shamblen, S. R. (2006). Law enforcement training in Southeast Asia: A theory-driven evaluation. *Police Practice Research*, *7*, 195-215.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed.). New York, NY: Holt, Rinehart and Winston.

*Kidiyala, S., Rawat, R., Roopnaraine, T., Babirye, F., & Ochai, R. (2009). Applying a programme theory framework to improve livelihood interventions integrated with HIV care and treatment programmes. *Journal of Development Effectiveness*, *1*, 470-491.

Leeuw, F. L. (2003). Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation*, *24*, 5-20.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott (Eds.), *Understanding causes and generalizing about them* (pp. 5–38). *New Directions for Program Evaluation, No. 57.* San Francisco, CA: Jossey-Bass.

Lipsey, M. W., Rossi, P. H., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

*Manomaivibool, P. (2008). Network management and environmental effectiveness: The management of end-of-life vehicles in the United Kingdom and in Sweden. *Journal of Cleaner Production*, *16*, 2006-2017.

Mark, M. M. (2007). Building a better evidence base for evaluation theory: Beyond general calls to a framework of types of research on evaluation. In N. L. Smith & P. R. Brandon (Eds.), *Fundamental issues in evaluation* (pp. 111-134). New York, NY: Guilford Press.

Mark, M. M., Henry, G. T., & Julnes, G. (1998). A realist theory of evaluation practice. In G. T. Henry, G. Julnes, & M. M. Mark (Eds.), *Realist evaluation: An emerging theory in support of practice* (pp. 3–31). *New Directions for Evaluation, No. 78.* San Francisco, CA: Jossey-Bass.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrative framework for understanding, guiding, and improving policies and programs.* San Francisco, CA: Jossey-Bass.

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

McLaughlin, J. A., & Jordan, G. B. (1999). Logic models: A tool for telling your program's performance story. *Evaluation and Program Planning*, *22*, 65-72.

*Mercier, C., Piat, M., Peladeau, N., & Dagenais, C. (2000). An application of theory-driven evaluation to a drop-in youth center. *Evaluation Review*, *24*, 73-91.

Miller, R. L. (2010). Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation*, *31*, 390-399.

Miller, R. L., & Campbell, R. (2006). Taking stock of empowerment evaluation: An empirical review. *American Journal of Evaluation*, *27*, 296-319.

Milstein, B., & Wetterhall, S. CDC Working Group. (2000). A framework featuring steps and standards for program evaluation. *Health Promotion Practice*, *1*, 221-228.

*Mole, K. F., Hart, M., Roper, S., & Saal, D. S. (2009). Assessing the effectiveness of business support services in England: Evidence from a theory-based evaluation. *International Small Business Journal*, *27*, 557-582.

*Mulroy, E. A., & Lauber, H. (2004). A user-friendly approach to program evaluation and effective community interventions for families at risk of homelessness. *Social Work*, *49*, 573-586.

*Nesman, T. M., Batsche, C., & Hernandez, M. (2007). Theory-based evaluation of a comprehensive Latino education initiative: An interactive evaluation approach. *Evaluation and Program Planning*, *30*, 267-281.

*O'Day, J., & Quick, H. E. (2009). Assessing instructional reform in San Diego: A theory-based approach. *Journal of Education for Students Placed at Risk*, *14*, 1-16.

*Oroviogoicoechea, C., & Watson, R. (2009). A quantitative analysis of the impact of a computerized information system on nurses' clinical practice using a realistic evaluation framework. *International Journal of Medical Informatics*, *78*, 839-849.

*Palumbo, D. J., & Gregware, P. R. (1992). Evaluating new dawn and Pegasus using the Chen and Rossi multi-goal, theory-driven approach. In H. T. Chen & P. H. Rossi (Eds.), *Using theory to improve program and policy evaluations* (pp. 145-163). Westport, CT: Greenwood Press.

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

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation and use.* New York, NY: Guilford.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London, UK: Sage.

*Pinkleton, B. E., Austin, E. W., Cohen, M., Miller, A., & Fitzgerald, E. (2007). A statewide evaluation of the effectiveness of media literacy training to prevent tobacco use among adolescents. *Health Communication*, *21*, 23-34.

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 73-102). New York, NY: Russell Sage Foundation.

*Reynolds, A. J. (2005). Confirmatory program evaluation: Applications to early childhood interventions. *Teachers College Record*, *107*, 2401-2425.

Rogers, P. J. (2000). Program theory evaluation: Not whether programs work but how they work. In D. L. Stufflebeam, G. F. Madaus, & T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 209-232). Boston, MA: Kluwer.

Rogers, P. J. (2007). Theory-based evaluation: Reflections ten years on. In S. Mathison (Ed.), *Enduring issues in evaluation: The 20th anniversary of the collaboration between NDE and AEA* (pp. 63-67). New Directions for Evaluation, No. 114. San Francisco, CA: Jossey-Bass.

Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation*, *14*, 29-48.

Rogers, P. J., Petrosino, A., Huebner, T. A., & Hacsi, T. A. (2000). Program theory evaluation: Practice, promise, and problems. In P. J. Rogers, T. A. Hacsi, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities* (pp. 5–14). *New Directions for Evaluation, No. 87.* San Francisco, CA: Jossey-Bass.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

*Sandler, I. N., West, S. G., Baca, L., Pillow, D. R., Gersten, J. C., Rogosch, F., & ... Ramirez, R. (1992). Linking empirically based theory and evaluation: The family bereavement program. *American Journal of Community Psychology*, *20*, 491-520.

*Sato, Y. (2005). A case of policy evaluation utilizing a logical framework: Evaluation of Japan's foreign student policy towards Thailand. *Evaluation*, *11*, 351-378.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand-McNally.

Scriven, M. (1973). Goal-free evaluation. In E. R. House (Ed.), *School evaluation: The politics and process* (pp. 319-328). Berkley, CA: McCutchan.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage.

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

Scriven, M. (1998). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation*, *19*, 57-70.

Scriven, M. (2007). Key evaluation checklist (KEC). Retrieved November 21, 2008, from http://www.wmich.edu/evalctr/checklists/kec_feb07.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Ex1perimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Thousand Oaks, CA: Sage.

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *American Journal of Evaluation*, *18*, 195-208.

*Sielbeck-Bowen, K. (2000). Development of local program theory: Using theory-oriented evaluation to make a difference. *Field Methods*, *12*, 129-152.

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice*, *14*, 237-242.

Smith, N. L. (1994). Clarifying and expanding the application of program theory-driven evaluations. *Evaluation Practice*, *15*, 83-87.

Smith, N. L. (2010). Introduction to the forum: Advances in evaluating evaluation theory. *American Journal of Evaluation*.

Stame, N. (2004). Theory-based evaluation and types of complexity. *Evaluation*, *10*, 58-76.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation*, *26*, 43-59.

Stufflebeam, D. L. (2001). Evaluation models. *New Directions for Evaluation, No. 89.* San Francisco, CA: Jossey-Bass.

Stufflebeam, D. L., & Shinkfield, A. J. (1985). *Systematic evaluation.* Norwell, MA: Kluwer-Nijhoff.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, & applications.* San Francisco, CA: Jossey-Bass.

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

*Tilley, N. (2004). Applying theory-driven evaluation to the British Crime Reduction Programme. *Criminal Justice*, *4*, 255-276.

Tourmen, C. (2009). Evaluators' decision making. *American Journal of Evaluation*, *30*, 7-30.

Trevisan, M. S. (2007). Evaluability assessment from 1986 to 2006. *American Journal of Evaluation*, *28*, 290-303.

*Trochim, W. M. K., Marcus, S. E., Masse, L. C., Moser, R. P., & Weld, P. C. (2008). The evaluation of large research initiatives: A participatory integrative mixed-methods approach. *American Journal of Evaluation*, *29*, 8-28.

*Turnbull, B. (2002). Program theory building: A strategy for deriving cumulative evaluation knowledge. *American Journal of Evaluation*, *23*, 275-290.

*Umble, K. E., Cervero, R. M., Yang, B., & Atkinson, W. L. (2000). Effects of traditional classroom and distance continuing education: A theory-driven evaluation of a vaccine-preventable diseases course. *American Journal of Public Health*, *90*, 1218-1224.

United Way of America. (1996). *Measuring program outcomes: A practical approach.* Alexandria, VA: Author.

Urban, J. B., & Trochim, W. (2009). The role of evaluation in research-practice integration: Working toward the "Golden Spike.". *American Journal of Evaluation*, *30*, 538-553.

Weaver, L., & Cousins, J. B. (2004). Unpacking the participatory process. *Journal of MultiDisciplinary Evaluation*, *1*, 19-40.

Weiss, C. H. (1972). *Evaluation.* Englewood Cliffs, NJ: Prentice Hall.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell, A. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New approaches to evaluating community initiatives: Volume 1, concepts, methods, and contexts* (pp. 65-92). New York, NY: Aspen Institute.

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

Weiss, C. H. (1997b). Theory-based evaluation: Past, present and future. In D. J. Rog & D. Fournier (Eds.), *Progress and future directions in evaluation: Perspectives on theory, practice and methods* (pp. 55-41). New Directions for Evaluation, No. 76. San Francisco, CA: Jossey-Bass.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Weiss, C. H. (2000). Which links in which theories shall we evaluate? In P. J. Rogers, T. A. Hasci, A. Petrosino, & T. A. Huebner (Eds.), *Program theory in evaluation: Challenges and opportunities* (pp. 35-45). New Directions for Evaluation, No. 87. San Francisco, CA: Jossey-Bass.

Weiss, C. H. (2004a). On theory-based evaluation: Winning friends and influencing people. *The Evaluation Exchange*, *IX*, 1-5.

Weiss, C. H. (2004b). Rooting for evaluation: A cliff notes version of my work. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 153-168). Thousand Oaks, CA: Sage.

*Weitzman, B. C., Mijanovich, T., Silver, D., & Brecher, C. (2009). Finding the impact in a messy intervention: Using an integrated design to evaluate a comprehensive citywide health initiative. *American Journal of Evaluation*, *30*, 495-514.

*Weitzman, B. C., Silver, D., & Dillman, K. N. (2002). Integrating a comparison group design into a theory of change evaluation: The case of the urban health initiative. *American Journal of Evaluation*, *23*, 371-385.

White, H. (2007). *Evaluating aid impact* (Research Paper No. 2007/75). Brighton, UK: University of Sussex, Institute of Development Studies.

White, H. (2009). Theory-based impact evaluation: Principles and practice. *Journal of Development Effectiveness*, *1*, 271-284.

*White, H., & Masset, E. (2007). Assessing interventions to improve child nutrition: A theory-based impact evaluation of the Bangladesh Integrated Nutrition Project. *Journal of International Development*, *19*, 627-652.

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

Williams, A. P., & Morris, J. C. (2009). The development of theory-driven evaluation in the military. *American Journal of Evaluation*, *30*, 62-79.

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159-176). New York, NY: Russell Sage Foundation.

Wingate, L. A., Coryn, C. L. S., Gullickson, A. R., & Cooksy, L. J. (2010). *The Program Evaluation Standards applied for metaevaluation purposes: Investigating interrater agreement and its implications for use.* Manuscript submitted for publication.

W. K. Kellogg Foundation. (1998). *W. K. Kellogg Foundation evaluation handbook.* Battle Creek, MI: Author.

W. K. Kellogg Foundation. (2000). *Logic model development guide.* Battle Creek, MI: Author.

World Bank. (2003). *Books, buildings and learning outcomes: An impact evaluation of World Bank support to basic education in Ghana.* Washington, DC: Operations Evaluation Department, World Bank.

World Bank. (2005). *Maintaining momentum to 2015? An impact evaluation of interventions to improve maternal and child health and nutrition outcomes in Bangladesh.* Washington, DC: Operations Evaluation Department, World Bank.

Worthen, B. R. (2001). Wither evaluation? That all depends. *American Journal of Evaluation*, *22*, 409-418.

Worthen, B. R., & Sanders, J. R. (1973). Evaluation as disciplined inquiry. In B. R. Worthen & J. R. Sanders (Eds.), *Educational evaluation: Theory and practice* (pp. 10-39). Worthington, OH: Charles A. Jones.

Wyatt Knowlton, L., & Phillips, C. C. (2008). *The logic model guidebook: Better strategies for great results.* Thousand Oaks, CA: Sage.