Evaluation theories: Guidance to evaluating in various circumstances

Melvin M. Mark
Penn State

To appear as Chapter 4 in *The Evaluation Handbook: An Evaluator's Companion*
Leonard Bickman and Debra Rog, Editors

Abstract

The idea of "theory" can be off-putting to some people, perhaps especially in a field such as evaluation. Evaluation is premised on the belief that its practice can make a positive difference in the world. Evaluators who want practical guidance may think that theory is abstract, overly general, and not sufficiently practical. This chapter asserts that, despite some people's skepticism about theory, fluency in evaluation theory can enhance evaluation practice. One major way that fluency in theory can be helpful is by improving judgments about which evaluation approach to employ in different contexts. This chapter: discusses what evaluation theory is; considers multiple reasons why evaluation theory matters; examines two meta-models, or theories about evaluation theories; and reviews a small, select set of classic evaluation theories, including for each theory an example from practice and an update on the theory's influence. The chapter also briefly describes aspects of several more recent evaluation theories, explicates different ways that evaluation theory can guide evaluation practice, and offers suggestions for the future of evaluation theory.

Introduction

For newcomers to the field of evaluation, it does not take long to discover that a vast array of methods are used under the umbrella of evaluation. Scanning through evaluation journals, or reviewing evaluation reports from different sources, or even reading the chapters in a volume such as this *Handbook,* the fledgling evaluator encounters a wide, perhaps even bewildering variety. One article might use a randomized experiment, where prospective program participants are assigned by a flip of the coin either to the program or to a waitlist, and the two groups are compared on outcomes of interest. In another case, the evaluator may conduct analyses of existing data sets to try to answer questions put forward by program managers. Yet another evaluation report may summarize the evaluator's efforts to engage program staff as collaborators, coaching them to collect and interpret data about the program. And another may summarize the use of in-depth interviews and observations to try to understand the lived experience of program participants.

The list of possible methods used by evaluators extends well beyond these few examples. Given such a rich set of alternatives, one of the most challenging tasks facing an evaluator is not *how* to do the evaluation, but *why* to do the evaluation *in a particular way*. In other words, on what basis should the evaluator (or whoever is deciding) choose from among the myriad options that evaluators use?

Evaluation theory is an important possible answer to this question, or at least a part of the answer. This is not a new suggestion. In a classic book on evaluation theory, Shadish, Cook, and Leviton (1991, p. 34) stated, "Evaluation theories are like military strategy and tactics; methods

are like military weapons and logistics. The good commander needs to know strategy and tactics to deploy weapons properly or to organize logistics in different situations. The good evaluator needs theories for the same reasons in choosing and deploying methods." Will Shadish, one of the authors of that book, made a similar point while commenting on evaluation training in the mid-1990s. Shadish said, "It is probably fair to say that most courses in program evaluation emphasize methods. Indeed, courses on theory are often seen as being of secondary relevance to the practical needs of evaluators. In my view nothing could be further from the truth. What the field lacks most is people who know something about when, where, why and how different methods could and should be used in evaluation practice. Theory tells us that" (Shadish, 1996, p. 553).

In this chapter, we (i.e., you, the reader, and me, the author) will explore the role of evaluation theory. We will consider what an evaluation theory is, and why practicing evaluators should be concerned about evaluation theory. We will look at a pair of meta-models, that is, frameworks that are designed to help make sense of a multitude of evaluation theories, and we will briefly consider how these meta-models can guide the study of evaluation theory. Using the two meta-models as a guide, we will compare and contrast a small, select set of (mostly) classic evaluation theories and more selectively review others, including some more recent ones. Then we will turn explicitly to a key, and underexamined, question: How exactly can evaluators draw on evaluation theories as helpful guides to evaluation practice? In that context, we will explore portions of a larger number of evaluation theories. Finally, we will take a look at possible developments in evaluation theory for the future.

What is Evaluation Theory?

Years ago, I was teaching a workshop on evaluation theory to a group of practicing evaluators, some novice and others not. I spent a bit of time early in the workshop talking about why it was important for evaluators to be familiar with evaluation theory, before I was going to turn to what an evaluation theory is. I noticed two people in the back of the room who seemed to grimace and whisper to each other every time I used the phrase "evaluation theory." I tend to spread introductions of workshop participants over time, so had not yet learned that these two were college teachers, from the physical sciences. As I recall, one of them taught physics and the other may have taught chemistry. They had gotten involved at their college with interventions that were intended to improve student success in the sciences, and especially to increase the participation of students from traditionally underrepresented groups. This involvement in turn led to their engagement in evaluation, a field relatively foreign to them. Especially foreign to them was my use of the word theory in the phrase evaluation theory. To them, theories were highly formalized, usually mathematically, with aspirations of wide applicability and generalizability. In contrast, others in the room from fields such as sociology and education were not so obviously bothered by my use of the term.

As I pointed out to those two workshop participants years ago, no one owns the word **theory**, and different people use it in different ways. Shadish, Cook and Leviton (1991, p. 30) said, "*What do we mean by theory?* No single understanding of the term is widely accepted. *Theory* connotes a body of knowledge that organizes, categorizes, describes, predicts, explains, and otherwise aids in understanding and controlling a topic." Shadish, Cook and Leviton went on to

indicate that a good **evaluation theory** would describe the activities to be engaged in while doing an evaluation, the goal(s) that the evaluation is intended to achieve, and the processes that are supposed to link the evaluative activities to the intended goals. Examples to come in this chapter will put some proverbial flesh on the bones of this description.

For the moment, perhaps it will help to note that a good evaluation theory contains more than a simple description of methods to use in doing an evaluation. A good evaluation theory also addresses such questions as what the goal of evaluation should be and why, and whether and how this should vary across situations. An evaluation theory should also clarify what counts as legitimate evidence, as well as where the questions to address in an evaluation should come from and why. A complete evaluation theory also lays out a stance on what the key characteristics are of social programs (or whatever it is that's being evaluated). And a good evaluation theory highlights the implications of these various attributes for how evaluation should be done. An experienced evaluator who is unfamiliar with evaluation theory will nevertheless have views regarding at least some of the issues a good evaluation theory should address. But a good evaluation theory will be more systematic and more explicit, including in the sense of suggesting linkages across its different elements (e.g., indicating how the goal of evaluation emphasized by that theory helps determine whose questions should have priority for an evaluation). Familiarity with evaluation theory, especially with multiple evaluation theories, should enrich evaluators' perspectives on the field, broaden their understanding of options that might be considered, and enhance their ability to match an evaluation to the circumstances.

The two physical science faculty members at my workshop are not the only people who have qualms with the term evaluation theory. Some evaluators also prefer different terminology, or at least are ambivalent about what the best label is. For example, Marv Alkin (2004) said that "In some ways, rather than theories it would be more appropriate to use the term *approaches* or *models*" (p. 4). Still, Alkin settled on using the term evaluation theory in his classic volume that traced the roots of a set of major theorists. Another notable evaluator, Dan Stufflebeam, wrote that his "monograph [which was published as an issue of the quarterly publication, *New Directions for Evaluation*] uses the term *evaluation approach* rather than *evaluation model*" (p. 9). Stufflebeam was concerned that the term "model" might be taken as implying that a particular approach was a good one. (For a touch of irony, check the reference section to see what the title of that issue of *NDE* is). Given that even people in evaluation differ in terms of their preferred terminology, it might be better to refer to "evaluation theories, models, or approaches." But that phrase is cumbersome and would likely grow old over the course of the chapter. Thus, for the most part in the rest of the chapter I refer to evaluation theories, which seems consistent with fairly common usage. And I invite you to substitute your preferred term, if you have one (perhaps keeping in mind that Smith, 2010, has encouraged different meanings for theory, model, and approach).

At the risk of adding confusion, I should note that some evaluators differentiate between different kinds of evaluation theories. We will return to one important distinction of this sort later in the chapter, when we consider the future of evaluation theory.

Why Evaluation Theory Matters

Evaluation is a practice-oriented enterprise. One might ask therefore, why should evaluators care about theory? Theories are often esoteric – and perhaps are viewed as esoteric even when they are not. A popular stereotype is that theorists travel at 30,000 feet, while the real action is taking place on the ground. We have already seen the primary rationale for attending to evaluation theory, which is that theory can serve as a guide to evaluation practice. We will return to this rationale in a subsequent section, which describes different ways of translating from evaluation theory to evaluation practice.

In this section we first attend to some other reasons evaluators should be attentive to evaluation theory. Two additional reasons are related to the idea of evaluation theory as a guide to practice. One is that evaluation theories are a way of consolidating lessons learned, of synthesizing prior experience, and of facilitating future progress. Years ago, Madaus, Scriven & Stufflebeam (1983, p. 4) indicated that evaluators who lack the kind of knowledge that is embedded in evaluation theory "are doomed to repeat past mistakes and, equally debilitating, will fail to sustain and build on past successes." An old expression refers to the inefficiencies of reinventing the wheel. Even more problematic is the possibility of reinventing the square wheel, that is, of adopting an approach that others had already learned does not work. There is also a relevant elaboration of the familiar expression that we should learn from our mistakes. Told in various ways and attributed to sources including Oliver Wendell Holmes, Eleanor Roosevelt, Admiral Hyman Rickover, and Groucho Marx, it goes something like this: "You need to learn from the mistakes of others, because you don't live long enough to make them all yourself." Familiarity with evaluation theory helps us learn the lessons of the past – after all, theorists tend not to advocate practices that in their experience have failed – and in this way evaluation theory can guide us away from repeating others' mistakes, that is, reinventing the square wheel.

A related reason for learning about evaluation theories is that comparing across theories is a way of identifying key debates in the field and, correspondingly, identifying unsettled or disputed practice issues. Newcomers to evaluation sometimes encounter rationales within their organizations of the "but that's the way it's done" variety. Very often, the way it's always been done in one's organization has support in some quarters, but not all. Being able to point to alternative theoretical perspectives that support a different way of doing things can be helpful, at least in facilitating consideration of future change or of fitting alternative evaluation practices to different circumstances. Understanding the points of disagreement between evaluation theories can also be valuable for those who want to contribute to the evaluation literature. These fault lines are often fertile territory for research on evaluation, or for trialing new approaches in an evaluation.

Yet another reason for attending to evaluation theory is that it should be a central part of our identity as evaluators and of our shared community and culture. In the words of Will Shadish, from his presidential address at the American Evaluation Association in the mid-1990s, "Evaluation theory is who we are" (Shadish, 1996, p. 1).

Beyond being an important aspect of our professional identity, evaluation theory can have a more tangible benefit, related to the observation that applied social researchers from various disciplines have methods expertise that overlaps with the methods an evaluator would use. What is it that makes an evaluator different than any applied social researcher who happens to conduct

an evaluation? One answer is that it is the knowledge and skill sets embodied in evaluation theory. Skillful use of evaluation theory, as already stated, can help guide judgments about what methods to use, when and why. In addition, familiarity with multiple evaluation theories, and with the different rationales they provide for alternative approaches to evaluation practice, can provide an awareness and appreciation for options that might not be apparent to an evaluator unfamiliar with evaluation theory. For example, familiarity with multiple evaluation theories can sensitize an evaluator to the varied possible sources of value judgments, as illustrated in a later section. Indeed, a case can be made that familiarity with evaluation theory, and the enhanced judgments that result, is the value proposition that makes it advisable for funders to hire an evaluator rather than a skilled research methodologist who is not an evaluator (or who claims the title of evaluator but is unaware to the range of options embedded in alternative evaluation theories).

Meta-Models

This section describes two versions of what can be called a **meta-model** or a theory of evaluation theories. By way of preview, I view one of the meta-models as describing in general terms the issues that a comprehensive evaluation theory should address. The other is an attempt to characterize the lay of the land, in the sense of grouping evaluation theories into something akin to families.

In the first of these meta-models, Shadish et al. (1991) contend that good evaluation theories need to address five general issues. Alternatively phrased, Shadish and colleagues claim that a comprehensive program evaluation theory would include **five components**. These are:

- *Social Programming*, which involves a view of how programs operate, what the external forces are that constrain programs, and the role programs play in social change.
- *Knowledge Construction*, such that the theory takes a stance on how to construct knowledge and justify knowledge claims.
- *Valuing*, in the sense that an evaluation theory indicates how to explicate value issues and select the values to which an evaluation should attend (such as which possible program outcomes to measure, and whether and how to weight each of a set of mixed findings). Foreshadowing some subsequent discussion in this chapter, Shadish and colleagues include matters of social justice in the valuing component of their meta-model.
- *Use*, whereby the theory specifies what kind of evaluation use matters and, if the theory addresses multiple forms of use, under what conditions each is more important, as well as what the evaluator ought to do to facilitate use (under various conditions).
- *Practice*, that is, more specific methods and techniques regarding how to do evaluation, including on-the-ground issues involving evaluation purpose, evaluator role, selection of questions, design, and activities related to use.

Examples to come should help illustrate and clarify these five components. Additional clarification and discussion on various components can also be found in other chapters of this Handbook, such as in Alkin and Vo's discussion of use (Chapter ?), Gates and Schwandt's presentation of valuing (Chapter ??). Archibald et al.'s discussion of evaluative thinking (Chapter ???) is also relevant.

Shadish (1996) used the circles of a Venn diagram to show the idea that these components overlap, with the practice component in the middle and overlapping with and drawing on each of the other four. If we were to apply this kind of graphical representation to individual evaluation theories, using a convention that the size of each circle corresponds to the amount of attention that component receives, we would see that theories differ in terms of their attention to a given component. For example, the utilization-focused evaluation approach of Patton (2008) gives a great deal of attention to use, while the approach associated with Campbell (1969) pays far less attention to use while dealing extensively with knowledge construction.

Shadish, Cook and Leviton used the five-component model to organize discussion of the work of seven evaluation theorists: Michael Scriven, Donald Campbell, Carol Weiss, Joseph Wholey, Robert Stake, Lee Cronbach, and Peter Rossi. This equating of an evaluation theor*y* with an evaluation theor*ist* has been common to date. We return to this matter later in the chapter.

A second meta-model, Alkin and Christie's evaluation theory tree (2004a; Christie & Alkin, 2013) is based on a similar idea, that is, that evaluation theories vary in terms of the *relative degree of emphasis* on three key issues: methods, valuing, and use/users. This is not to say that any evaluation theory addresses only one of these issues. Rather, the idea is that theories tend to emphasize one of these issues *more* than the others.

The evaluation theory tree has been revised several times, including different versions in the three editions of the book *Evaluation Roots* (Alkin, 2004, 2013; Alkin and Christie, in press). The origin of the theory tree lies in the efforts of Alkin and his students to trace the analog of academic genealogies within program evaluation. In other fields such as psychology, people have traced mentor-mentee relationships in ways that resemble family trees (see, e.g., https://academictree.org/psych/). The short history of evaluation as a distinguishable scholarly and practice field does not lend itself to a family tree. Evaluation does not have decades of one faculty member mentoring multiple students, some of whom go on to mentor others. Instead, in a relatively short period of time, many people trained in other fields entered evaluation, and some of them came to be labeled as evaluation theorists. In addition, because evaluation has been multi-disciplinary, some major evaluation figures have trained students who focus on their home discipline. Thus, the family tree model does not work as it does in older disciplines such as psychology (where, e.g., Len Bickman's mentors came from two major lines in social psychology: (1) Stanley Milgram, who was trained by Gordon Allport, who had been trained by Solomon Asch, and (2) Hal Proshansky, who was trained by Morton Deutsch, who had been trained by Kurt Lewin, the latter widely considered the "father" of social psychology). Instead, in classes on evaluation theory, Alkin and his students discussed different representations and ended with the metaphor of a tree. Unlike an academic family tree, the evaluation theory tree is not based on advisor-advisee relationships. And unlike a traditional family tree, the evaluation theory tree does not widen as you move back in time, with one set of parents, two sets of grandparents, and so on. Instead, the theory tree has three main branches, which are defined by *relative focus* and organized in terms of patterns of influence. Like the Shadish et al. (1991) book, the evaluation theory tree and the Alkin (2004; 2013) book has been organized in terms of evaluation theorists. This is explicit in the book's original subtitle: T*racing Theorists' Views and Influences* (Alkin, 2004). This linkage of theory and theorist is evolving, however, as noted in a subsequent section of the chapter on the future of evaluation theory.

INSERT FIGURE 4.1 ABOUT HERE

Figure 4.1 is one version of the **evaluation theory tree**. There have been several versions of the tree, including one or two in each edition of the Evaluation Roots book. Figure 4.1 came from an article that revisited the version of three that had been presented in the first edition of Roots. Later in the chapter, we will see a more recent version, which includes revisions aimed at addressing some of the criticisms of the original tree.

As shown in Figure 4.1, there are **three branches** on the Alkin and Christie evaluation theory tree, each associated with one of three issues: "(a) issues related to the methodology being used; (b) the manner in which the data are to be judged or valued; and (c) the user focus of the evaluation effort" (Alkin, 2004, p. 7). Each theorist is located on one branch, based on which of these three issues they emphasize(d) the most.

In addition, a theorist who tended to influence other theorists is located closer to the trunk of the tree. In the first edition of *Evaluation Roots,* Alkin asked the theorists who contributed chapters to comment on their placement on the evaluation tree and to list the main influences on their work. The structure of the tree was in some sense validated, in that most of the volume contributors were satisfied with their placement, and in that there appeared to be more influence within, rather than across the branches of the tree. For instance, in revisiting the tree in the last chapter of the first edition of *Roots,* Alkin and Christie (2004b, p. 391) note that "On the methods branch, Cook, Weiss, Rossi, and Chen all indicate the strong influence of Donald Campbell. They also note that Cronbach, who was represented in the Roots book by Jennifer Greene, a former student, clearly "was influenced in the way in which he reacted to Campbell's writing" (p. 391).

As illustrated in subsequent sections, the two meta-models can be used to help frame a review of evaluation theories. They also can be used to guide one's own thinking and reading or, if you prefer, to inform one's professional development activities regarding evaluation. In numerous workshops, I have asked people to identify a handful of big picture issues or questions that they would like an evaluation theory to address. Many of the answers translate reasonably well into one or more of the five components of evaluation theory identified by Shadish and his colleagues (usually when an issue involves two components, one of them is practice). However, people often do not generate questions that correspond to social programming. Despite this common omission, attention to the thing that we are evaluating (or as Michael Scriven has labeled it, the "evaluand") can be quite important. Archimedes is supposed to have said something to the effect of, "give me a lever and a place to stand, and I can move the world." To know where there are leverage points, one must know something about the thing being evaluated. Is a yes-or-no, go-or-no-go, thumbs-up-or-thumbs-down kind of decision foreseeable? Or are only smaller, more incremental changes feasible? Are there times when program change is more (or less) likely to take place? An evaluation may not be positioned to make a difference if the evaluator (or someone, such as the commissioner of the evaluation) has not thought about what kind of information is likely to be actionable, and when, given the program that is being evaluated and its environment.

More generally, self-assessment in relation to the five components of the Shadish and colleagues' meta-model can guide one's personal learning agenda for the future. So, for example, if you have no idea what a given component of the model refers to, then it would probably be valuable to do readings, attend talks, or participate in workshops that include a heavy dose of that component. The theory tree can be used in a similar fashion, to guide continuing professional development regarding a branch with which you are unfamiliar or, alternatively, to dig deeper into the writings associated with a branch with which you have an affinity. In short, although they are used in the remainder of this chapter to guide our selection and review of a limited set of evaluation theories, the two meta-models have value in and of themselves.

Evaluation theory is not the same as program theory

To avoid a potential source of confusion, evaluation theory needs to be differentiated from program theory, a concept used widely in evaluation literature and practice. **Program theory** (Bickman, 1989; Chen, 1990) refers to a conceptual model, often presented schematically, that describes the anticipated operations of a program. Many variants exist. Some have boxes that show the components of the program on the left, the long-term anticipated outcomes or goals of the program on the right, intermediate steps or underlying processes in the middle, and arrows representing anticipated cause-and-affect relationships among the components, processes, and outcomes. Developing or uncovering a program theory or one of its cousins, such as a logic model (See Chapter X), is widespread in evaluation practice. It is also a central part of what was originally labeled theory-driven evaluation (Chen, 1990), and which for the sake of clarification has been labeled as program theory-driven evaluation (Donaldson, 2007).

Program theory-driven evaluation can be considered as one example of what we are calling evaluation theory. But "program theory" is a more specific term than evaluation theory. Evaluation theories *can* attend to program theory, but they need not do so. At the risk of repeating words too often in a single sentence: Program theories are central to the evaluation theory called program theory-driven evaluation; on the other hand, there is no requirement for all evaluation theories to attend to what is widely called program theory.

Criticisms and comparisons of the meta-models

The original version of the evaluation theory tree was dominated by theorists who were North American, male and white. Not surprisingly, this has been a cause for criticism. In response, the second edition of the Alkin book includes an expanded list of theorists, as does the final version of the tree in that book. And the third edition, forthcoming at the time this chapter was written, has an even larger and more diverse set of theories and theorists represented (as we shall see in a subsequent figure).

In a related criticism, Donna Mertens has argued that the evaluation tree should be expanded to include a fourth branch, social justice (e.g., Mertens & Wilson, 2019). There has indeed been a proliferation of evaluation theorists who emphasize social justice, especially since the heyday of the early theorists who were reviewed in Shadish, Cook and Leviton (1991). Mertens and Wilson support the idea of four branches by tying each one to a different philosophical stance, with the social justice branch tied to a transformative paradigm. This is a defensible position, though one

can question whether the relationship between espoused paradigms and evaluation theories is so strong that all the theorists on a branch follow a common paradigm, distinct from paradigms associated with the other branches. An alternative to the Mertens and Wilson approach, as Alkin and Christie suggest, is that evaluation theories which emphasize social justice can be included on the valuing branch. One can argue that this makes sense, given that an emphasis on social justice has implications primarily for the values-bases that are to be imbued within an evaluation. In any case, the most recent version of the Alkin and Christie (in press) theory tree includes social justice oriented theories as a major subbranch of the values branch, as we shall see later. Regardless of whether one supports a three- or a four-branch theory tree, it is an important, positive development to see a larger number of social justice focused theories and a more diverse set of theorists represented on the tree.

The theory tree was also criticized by Shadish (2004), who wrote a chapter representing Campbell for the original *Roots* volume. Shadish argued against the three branch structure of the tree, and in favor of the five component structure he developed with Cook and Leviton. One perspective, adopted here, is that both meta-models are useful. The theory tree gives a comparative map, across multiple evaluation theories, based on relative emphasis and on sources of influence over time. In contrast, the five-component model of Shadish, Cook and Leviton provides an aspirational outline for any individual theory of evaluation that attempts to be comprehensive. (For further discussion related to Shadish's criticism of the evaluation theory tree, and some further thoughts on what the differences between the two meta-models might reveal, see Appendix A).

A Few Evaluation Theories, Reviewed

In a semester long class on evaluation theories, it would be common to review several evaluation theories. In a 15-week semester, a dozen or more theorists might be reviewed. In selecting theorists for review, the evaluation tree can serve as a kind of sampling frame, with one or more theories drawn from each major branch. A more in-depth review than the current chapter could include both older, influential theories near the trunk of the tree, newer theories, toward the tips of each branch, and some theories that fall in-between. For each theory, the five components from Shadish, Cook and Leviton might be used to organize discussion. In this chapter, however, only a few seminal evaluation theories will be summarized in relation to all five components, with a theorist near the trunk from each of the three branches. A few more will subsequently be reviewed more selectively. Readers interested in further exploration of evaluation theories should see the suggested resources at the end of the chapter.

As previously noted, the literature on evaluation theories has often been presented in terms of the work of individual evaluation theorists. Both the Shadish, Cook, and Leviton (1991) and Alkin (2004; 2013) books are organized around evaluation theorists, for example. Mertens and Wilson (2019) also summarize the views of numerous evaluation theorists. I will follow that tradition of organizing this section and the next around evaluation theorists. Descriptions of the approaches associated with Donald Campbell, Joseph Wholey, and Michael Scriven follow, as does a newer approach, culturally responsive evaluation (which is associated with multiple theorists). In the final section of the chapter, I will comment on the tradition of equating evaluation theories with

evaluation theorists, expressing the hope that things will be different in the future and noting a major step in that direction. Now, on to the individual theorists.

Don Campbell was a social psychologist, though he is probably most widely known and remembered for his contributions to research methodology. Much of his writing fits within Shadish and colleagues' knowledge construction component of evaluation theory and within the methods branch of the evaluation theory tree. With various collaborators, Campbell developed and popularized the distinction between internal and external validity, created lists of validity threats, and detailed alternative quasi-experimental designs and examined their susceptibility to various validity threats.

Internal validity, in the context of program evaluation, refers to the accuracy of one's inferences about whether (or to what extent) the program caused a difference in some outcome(s) of interest. **External validity**, in contrast, refers to the accuracy of one's inferences about whether (or to what extent) the findings of an evaluation can be generalized to other settings, to other people, and to other times.

With his list of internal validity threats, Campbell helped sensitize others to the idea of considering and trying to rule out plausible alternative explanations. In the context of a program evaluation, the evaluator of a preschool program would need to rule out the possibility that children score better on outcome measures simply because they are older than when they enrolled; this would be an example of the internal validity threat of **maturation.** Ruling out the threat of maturation might best be done by using a comparative research design, with some children in the preschool program and others in the no-treatment comparison group. But everyday processes, whereby some children end up in preschool and others do not, could lead to qualitatively different groups -- a threat called **selection** -- which would make it harder to tell whether preschool is making a positive difference. Campbell's writings sensitize evaluators to the possibility of such biases, while also indicating which are better ways to try to deal with them. Henry (Chapter ?) discusses some of the design options associated with the Campbellian tradition, as well as more contemporary analysis techniques.

At a more abstract level, Campbell espoused the philosophical stance of **critical realism**. (Bhaskar, 1978; Shadish, Cook & Leviton, 1991, Chapter 4). By critical realism, he meant to suggest that there is indeed a world that exists apart from social construction of it, and that accordingly it was meaningful to talk about such things as the causal effect of an educational program on student achievement. At the same time, Campbell recognized that our understanding of this external reality is imperfect, mediated by human knowledge processes that may be biased. His philosophical view of knowledge construction fit well with his emphasis on preferred research methods as a means of trying to reduce biases in evaluation findings. As Bhaskar (1978, p.43) noted, "to be a fallibilist about knowledge, it is necessary to be a realist about things." In other words, to believe our knowledge may not be perfect, but that better versus worse answers exist, you also have to believe that there is an answer that exists apart from human construction.

Although Campbell's major emphasis was on knowledge construction and methods, his approach also had a particular focus when it came to the social programming component of Shadish et al.'s meta-model. His emphasis was on identifying relatively effective program options. The idea was

that good information about whether a program was effective, or about which of two or more programs was more effective, could help inform judgments by policymakers and others when they make decisions about whether to implement a new program, or about whether to expand a small-scale program, or about whether an existing program should be replaced or continued. Colloquially speaking, Campbell's model of evaluation fits with the decision-making context when there is a fork in the road; for example, when Congress faces reauthorization of major educational or social welfare legislation. In such a circumstance, better choices can be made if sound answers are available to the question: Which, if any, alternatives to the current arrangements lead to better outcomes?

Regarding use in Campbell's approach, the focus was on evaluative information informing decision-makers who made judgments about program adoption, continuation, expansion, or cessation. This form of use illustrates what is sometimes called **direct or instrumental use**, that is, identifiable action or decisions. Campbell did not go very far, however, when it came to the use component. His focus was far more on how to get the best, least biased answer to the question about the relative effects of program or policy alternatives -- not on how to try to ensure that these results were used in actual decision-making. He sometimes referred to evaluators as methodological servants of the experimenting society (Campbell, 1991)

Similarly, Campbell did not go extensively into the issue of valuing. Instead, he seemed to assume that matters such as deciding what indicators to use when assessing the effectiveness of an intervention, or how to weight different criteria when a program does well on some indicators but poorly on others, is best left in the hands of others. Often in practice this appears to translate into adopting official program goals as the criteria to be used in evaluating a program. My own view is that it is a principled, defensible stance to say that valuing should be the responsibility of those who are in legitimate decision-making positions. As we shall see, however, this is far from the only view of valuing within evaluation theories, and other perspectives exist that are also principled and defensible. And more likely to seem appropriate to many contemporary evaluators and stakeholders.

Having provided a summary of one notable evaluation theorist's views, I hasten to add a set of caveats. It is impossible to do justice to decades worth of writings in such a brief summary. A theorist's views may change over time, and such changes or other nuances are hard to capture in a page or two. Nevertheless, even a brief description such as those here can clarify the similarities and differences associated with various evaluation theories. In addition, even a brief overview such as this covers more territory than the typical truncated view of Campbell and other early theorists today. The view that many current evaluators have of Campbell (assuming they have one) is simply of a methodologist who advocated for the use of randomized experiments and strong quasi-experiments in evaluation. One benefit of studying program theorists is to gain a richer view of their perspectives, thereby enriching our own.

BOX STARTS HERE. AND INSERT FIGURE 4.2 ABOUT HERE

Among the quasi-experiments that Campbell advocated is the **"interrupted time series" (ITS) design**. Michieulutte and colleagues (2000) used an ITS to complement a pretest-posttest comparison group design, in their evaluation of an education and support program designed to

increase screening for breast and cervical cancer among women aged 40 and older in low-income housing in Winston-Salem, NC. FIGURE 4.2 shows the ITS portion of the evaluation, with the number of mammograms per 100 patient visits shown over time across 19 two-month intervals (data were aggregated into two-month intervals to avoid instability from smaller numbers and from short-term closures of the mammography unit). Data are reported separately for women 40-49, for whom the official screening recommendations had varied over time, and for women 50 and above, for whom screening recommendations had been consistent. As shown in the solid line, for women aged 40-49, mammogram rates are fairly steady prior to the intervention, with values ranging from 6 to 7 per 100 patient visits. Just after the program started, the rate increased linearly for the next several observations, plateauing at a rate of 8.5 to 8.75 mammograms per 100 patient visits. Statistical analysis showed the increase following the introduction of the program was significant. Note that the time series data give more information about the intervention's effect across time than would a typical pretest-posttest comparison (i.e., that after an initial increase, the rate plateaued). As shown in the dotted line in Figure 4.2, for women 50 and older there was not a similar increase immediately following the intervention. However, mammogram rates among women 50 and older increased later, roughly 2 months after the start of the intervention. The evaluators speculate that this is a delayed effect of the program, with the 40-49 year olds being more immediately susceptible to information in the intervention because of the previously conflicting recommendations about screening in that age group. Uncertainty about the findings for those 50 and older is a good reminder about the benefits of having design adjuncts beyond the simple ITS (Reichardt, 2019), such as (a) implementation assessment, which could indicate whether women of different ages received somewhat different treatment over time, (b) qualitative data from the women, which could have been informative about the processes of change, and (c) time series data from outside the treatment area, which should show no increase in mammogram rates in either age group -- *if* Figure 4.2 is showing an immediate and a delayed treatment effect, respectively, among women 40-49 and 50 and above. END OF BOX

By way of update, Campbell's approach to program evaluation has continuing influence in the field today. Evaluations using experimental and quasi-experimental designs to estimate the effects of a program are fairly commonplace in various programming areas (e.g., Gopalan, 2020). In addition, descendants of Campbell's theory appear on the theory tree and in practice. One of these, generally known as theory-driven evaluation (Chen, 1990, 2005; Coryn et al., 2011, is perhaps better labeled program theory-driven evaluation (Donaldson, 2007) to clarify that it is an evaluation approach that emphasizes program theory. Much of program theory-driven evaluation draws on Campbell, especially with respect to knowledge construction, concern for validity threats, and experimental and quasi-experimental methods. However, in terms of the social program, program theory-driven evaluators emphasize the rationale underlying the program – that is, the program theory – and they use the program theory to guide the evaluation, for instance, in the identification of interim and long-term measures. More generally, Campbell remains influential to many approaches that attempt to identify how and when a program works (Lemire et al., 2020).

Compared to Campbell, program theory-driven evaluation is more sensitive to program stage. For example, early in a program's life, the evaluator may focus more on surfacing the program theory and on examining whether the program is operating as the program theory suggests,

including changes in the more proximal indicators. For a more mature program, these evaluators often use a Campbell-like comparative design to test whether the program causes improvements in the long-term outcomes. But going beyond Campbell, this is likely to be accompanied by efforts to see whether changes occur as expected along the various steps in the program theory, often using **mediational analyses** (described in Chapter Z? – Gary's?). Advocates of program theory-driven evaluation say their approach not only shows whether a program is bringing about the desired change, but also why it is (or is not), thereby opening the door to possible improvements. More generally, those following in Campbell's footsteps tend to address a wider array of questions, including about implementation, **mediators**, and **moderators**, assuming resources and practicalities allow.

The continuing influence of Campbell's approach to program evaluation is also the subject of considerable criticism. The criticisms focus largely on overstated claims by advocates of randomized experiments, often referred to as RCTs (for **Randomized Controlled Trials**). RCT advocates, sometimes referred to as Randomistas. are criticized for treating experiments as an unqualified gold standard, indicating they "clinch" a conclusion, and ignoring or at least downplaying the challenges of generalizing such findings elsewhere (Deaton & Cartwright, 2018). The overstated claims of some advocates of RCTs seems ironic, given Campbell developed the concept of external validity and also cataloged quasi-experiments, some of which approach RCTs in terms of internal validity.

For the next evaluation theorist, we consider **Joseph Wholey**. The initial focus will be on one portion of Wholey's approach, specifically on the development of **performance measurement systems**. We turn subsequently to another aspect of Wholey's work, evaluability assessment. Wholey worked for many years at the Urban Institute, where he and his colleagues spent considerable time observing ongoing programs. They noticed that, although program managers played a vital role in the operations of the program, they often lacked quality information about how well things were going. In contrast, in the private sector business managers typically at the very least had relatively good and timely feedback about the proverbial bottom line, that is, about how the business or their unit was doing from a profit-loss perspective. Unlike the private sector, public sector managers typically lacked comparable information about the extent to which program activities were successful in leading to program objectives. Given the shortage of information available to the public sector managers at that time, Wholey and his colleagues became leading advocates for the development of information systems, sometimes called performance monitoring or performance measurement systems.

These data systems typically include information about program activities, such as the extent to which various program services were delivered. They track client outcomes, ideally starting with measures from when the client entered the program and continuing as far downstream as feasible. Such systems should be constructed with managers' information needs in mind. Thus, if managers anticipate needing aggregate data by regions or sites, this should be accommodated. Ideally, information systems provide reports on a regular schedule suited to managers' needs, but can also be queried if desired. They may also have some form of red flag system so that unusual or problematic scores on an indicator result in a report even if one is not scheduled.

In contrast to Campbell and his focus on go/no-go decisions, and on the kind of information policy makers would find useful in choosing the better option, Wholey focused on social programs that are ongoing, and on the kind of information program managers would find useful for improving ongoing program operations. Metaphorically, Campbell was concerned about occasions when there are forks in the road, while Wholey's concern focused on situations in which a program will be moving on down the road for quite some time. Neither is right or wrong. Instead, they focus on different periods in the lifecycle of a social program, and correspondingly on the information needs of different parties (policy makers versus program managers). And importantly, these contrasting views of social programming and evaluation users are associated with very different looking approaches to evaluation.

Wholey had a clear focus with respect to use. He identified a primary intended user group, that is to say, program managers. Information systems were to be developed with their information needs in mind. If the desired use took place, it would mean that managers drew upon the results from the information systems to evaluate how things were going, to implement modifications when and where things were not going well, and to track whether outcomes improved following those modifications. As with Campbell, this is a kind of direct, instrumental use. However, Campbell focused on full scale programs being adopted or not, while Wholey's use would likely involve more minor program modifications, and perhaps many such modifications tried out over time. Using language from Michael Scriven, Campbell's evaluations are **summative** and Wholey's performance measurement-based evaluative judgments are **formative**.

START OF A BOX Sawhill and Williamson (2001) describe the development and use of a performance measurement system that is consistent with Wholey's approach. The Nature Conservancy is a nonprofit organization committed to conserving biodiversity. For much of its existence, it did so primarily by acquiring land thought vital for the survival of at-risk species. It focused on two indicators, "bucks and acres," that is, the total of the donations received, and land acquired. Although the Conservancy was growing annually at double digits, the old bucks and acres approach contributed to some dysfunctions, such as buying land, especially in states rich in donors, rather than dealing with broader ecosystem issues. This is not surprising to those familiar with the mantra, "What gets measured gets managed," an expression that in part serves as a warning about the possible dysfunctional effects of excessive attention to imperfect performance measures. So the Nature Conservancy went to work on a revised system. A first attempt at an improved performance system failed, however, revealing a valuable lesson. In a year-long process, the Conservancy developed a system with 98 measures. It "promptly collapsed under its own weight" (Sawhill & Davidson, p. 374) due to the record-keeping burden and the fact that managers could not tell which portions of the extensive information were most important. Subsequently, drawing on a form of strategic planning, a usable set of performance indicators was developed, with the nine measures falling within three general categories: capacity, activity, and impact. Sawhill and Davidson report that the revised system had effects. For example, it motivated line managers to initiate activities other than land purchases, and guided staff to do more in terms of identifying threats to species in their area and seeking ways to abate them. Notably, evaluative judgments are embedded in such changes. And over time, a performance measurement system like this should enhance managers' ability to adjust program activities based on the impact measures. END OF BOX

The preceding description of Wholey's approach to evaluation actually pulls out one part from the richer landscape of his work. That portion, using performance measurement for what has come to be known as results-oriented management, is what Wholey is best known for today in some circles, such as in public policy and management. At least in part, this is because of the codification of performance measurement into US law, especially in the Government Performance and Results Act of 1993 (GPRA) and the GPRA Modernization act of 2010. The widespread development of and attention to performance measurement systems in the wake of these laws followed in the footsteps of Wholey and his colleagues. Originally, however, Wholey's attention to performance measurement systems was a portion of a multi-step set of options.

Another of Wholey's options was **Evaluability Assessment** (EA) (see, e.g., Wholey, 2004; Trevisan & Walser, 2015). This is another contribution, for which Wholey is best known among some evaluators. An EA involves the evaluator, program managers and staff, and key policymakers if possible. It creates a kind of map of a program, that is, a form of program theory, often in the form of a **logic model**. If agreement exists about program goals, if a plausible set of linkages connects program activities to the desired goals, and if further evaluation could address potential users' information needs, then the EA would conclude that further evaluation activity is warranted. On the other hand, if one or more of these conditions were not met, then the EA would suggest that this problem should be addressed before other evaluation activities take place. For example, if general agreement did not exist about a program's goals -- as might be the case in a multi-component program that evolved willy-nilly over time – this problem would need to be addressed first, perhaps by streamlining the program. EA and the possibility of next steps illustrates what Wholey called the sequential purchase of evaluative information. Colloquially, he encouraged that the right bite-size of evaluation be found.

Another option Wholey endorsed was rapid feedback evaluation. The premise is that oftentimes organizations already have plenty of data with evaluative implications, but they lack individuals with the time and perhaps the ability to make sense of it. In essence, rapid feedback evaluation involves parachuting in a couple of expert evaluators to review and offer recommendations based on observations, interviews, and existing data. The kind of results-oriented management via performance measurement systems that we have already discussed, is another of Wholey's options. So too is what Wholey called intensive evaluation, which he defined less explicitly but might involve an approach such as Campbell's.

Wholey's approach has held considerable sway in government, especially the U.S. federal government, where the development and use of performance measurement systems is a key element of what has come to be called results-based management (Holzer & Ballard, 2011). Though the approach has limits, including the inability to attribute a change in outcomes to the program (DeLancer Julnes, 2006), it has grown, in no small part due to the mandates of GPRA and the GPRA Modernization Act of 2010. Those involved in the development and use of the performance measurement systems stimulated by these laws follow in the footsteps of Wholey and his colleagues. With respect to evaluability assessment, reviews have shown a resurgence of its use early in this century in areas including public health (Leviton et al., 2010) and international development (Davies & Payne, 2015).

Although the writings of a third evaluation theorist, **Michael Scriven**, are plentiful and rich, we will review his work only briefly here. As to knowledge construction, Scriven believes there are real, knowable answers to the question of whether a program (or other evaluand) has merit and worth. But Scriven also believes that a person's perspective can bias their evaluations. As a result, he generally prefers an external (versus an internal) evaluator, especially for summative evaluations. Scriven also endorses **meta-evaluation**, that is, the evaluation of an evaluation. As related to bias control, the idea is that, if evaluators know their judgments will be reviewed by a knowledgeable third party, they will be better at laying out the bases for their judgments and at fairly calling balls and strikes, so to speak.

Scriven has been especially lauded for his approach to valuing, which has been described as involving four general steps (Shadish et al., 1991). The first is to identify criteria of merit, that is, the characteristics by which the evaluand is to be judged. For example, the criteria of merit for a preschool program might include children's language skills, self-regulation, and interpersonal skills. Second, performance standards should be set. Will success be judged relatively (e.g., compared to a control group), or is there an absolute level of performance required for a judgment of success -- and, if so, what is it? Third, performance must be measured. This should be done using whatever measurement techniques are appropriate (and evaluating running shoes calls for different methods than evaluating preschool programs, for example). Fourth, results should be synthesized into an evaluative judgment. How, for example, should a preschool program with moderate positive effects on interpersonal skills and self-regulation but no effect on language skills be rated, relative to with another program with strong positive effects on language skills only?

It is hard to point to an archetypical Scriven-style evaluation, given that he is open to whatever way of measuring performance makes sense for a given evaluand. The commonalities involve the steps of the evaluation, including where the criteria of merit originate. Scriven does *not* see criteria of merit as coming from public officials, managers, or other decisions makers. Indeed, he has advocated for "goal-free evaluation," in which evaluators avoid even familiarizing themselves with official statements about what a program is intended to do, such as would be found in enabling legislation or program documentation. Instead, Scriven suggests that the evaluator should attend to a kind of **needs assessment**, identifying the real needs that a program was intended to address. Scriven assumes a stronger role for the evaluator in identifying needs than most evaluators would. Scriven would approve of the identification of needs being informed by sources such as stakeholders, but would not hand the identification of needs to them. Many other evaluators would instead rely primarily on stakeholder judgments, though various techniques exist that can be used to identify needs (Russ-Eft & Sleezer, 2019).

In any case, the identified needs would then inform valuing, such as selection of the indicators to be measured and the way to integrate across multiple indicators in order to generate a single, bottom-line judgment (regarding the synthesis step, according to Scriven the weighting should reflect the extent to which the indicators are related to basic needs). Many other evaluation theorists do not include in their approaches the synthesis of findings from multiple criteria of merit into a single evaluative judgment. Campbell, for example, would leave this to others, such as the decision makers in a democracy. But for Scriven, evaluating requires getting to a judgment, a grade, a rating – to an evaluative conclusion. For Scriven, the evaluation results are

then made available to the consumer, policymaker, or whoever, with the idea that the evaluation results can help them make better choices if they choose.

In the first edition of the *Roots* book, other theorists on the valuing branch, including Stake, Guba and Lincoln, and House, mentioned the influence of Scriven on their work (Alkin, 2004, p. 391). In a sense, the influence of Scriven today is partially mediated by his influence on these and other influential figures. In other ways, it is difficult to update Scriven's influence, largely because he did not espouse particular research designs, specific stakeholder engagement processes, or the like. His logic is flexible in the sense of being applicable across a wide array of areas and techniques. The work of Davidson (2004, 2012), a past student of Scriven's, is perhaps the most notable example of his direct ongoing influence.

When talking about endeavors such as literature, movies or rock music, a classic is typically one from an earlier era that is of high merit and has stood the test of time. But there is also the idea of an "instant classic," a work that, while more recent, is of such quality that it deserves the accolade "classic" even without the usual passage of time. If we think of the evaluation theories reviewed so far, from Campbell, Wholey, and Scriven, as classics, we can think of a more recent approach, **culturally responsive evaluation (CRE)**, as an instant classic.

CRE, unlike the approaches identified so far, is not affiliated with one single theorist. Instead, this theory is more generally credited to a number of developers, several of whom have collaborated at times. The term culturally responsive evaluation is credited to Stafford Hood (1998) in a presentation and subsequent chapter honoring the work of Robert Stake (1976), who had developed "responsive evaluation" (Hood, Hopson, & Kirkhart, 2015). In short, Stake had argued against pre-ordinate evaluation designs, calling instead for the evaluator to immerse in, and subsequently plan evaluation activities that were responsive to the actual program as implemented and its context. In introducing CRE, Hood (1998) emphasized the importance of responsiveness to cultural features, including race. Since that initial reference to CRE, several key contributions have taken place, including Frierson, Hood and Hughes (2002), Hopson (2009), Frierson, Hood, Hughes and Thomas (2010) and Hood, Hopson and Kirkhart (2015). And it is important to note that, even though I have labeled CRE an instant classic, it has drawn on and acknowledged important precursors and contributions, including scholars of color such as Reid Jackson and Asa Hilliard as highlighted in the "nobody knows my name" project (Hood & Hopson, 2008).

CRE addresses all five components of the Shadish, Cook and Leviton model of evaluation theory. With respect to social programming, CRE is premised on the idea that programs take place within a cultural context and that efforts to evaluate a program without adequate consideration of that context will fall short. Part of accounting for the cultural context, according to CRE, involves recognition of relevant power dynamics. More generally, CRE calls on evaluators to describe the history of the program, including its cultural context.

With respect to the valuing component, a culturally responsive evaluator seeks to include within an evaluation's focus the questions and concerns held by the intended program beneficiaries, along with those of other key stakeholders. Notice the contrast with the previously discussed theories. Also related to valuing, a culturally responsive evaluation will explicitly address

dynamics of power, privilege, and inequity as they arise in the program and its context. This requires having and/or developing an understanding and appreciation of intended beneficiaries' lived experience.

Drawing on social constructivism, CRE emphasizes that knowledge construction takes place in a social and cultural context. Accordingly, who is at the table, or put differently what voices are included in the processes of knowledge construction, matters greatly. Program participants and those from traditionally underserved communities should participate in the shaping of evaluative conclusions, and they should be asked to weigh in on the accuracy of those conclusions. From the moment of the evaluator's initial engagement, the process of building trust is critical, including in terms of the quality of the conclusions that are drawn from an evaluation. Culturally responsive evaluators also tend to eschew deficit-based explanations of the problems a program is designed to address, and instead acknowledge both strengths in the relevant community and the systems factors that contribute to the problem.

Regarding use, a key contention of CRE is that being culturally responsive throughout the evaluation process will result in evaluation findings (and processes) that are both more credible and more useful. Reasons for this include the emphasis on developing trust and the involvement of key stakeholders in question formulation and in the review of preliminary findings. CRE encourages dissemination to a range of stakeholders, certainly including intended program beneficiaries, in culturally appropriate ways, and perhaps with different communication procedures for different groups. Where feasible, feedback to the various stakeholder groups should be done in an ongoing basis.

In terms of practice, CRE calls for bringing a culturally responsive lens to all stages of an evaluation, from preparing to enter the context in which the evaluation will be conducted, to the framing of questions, and to the dissemination of findings and facilitation of use (to mention but a few of the stages evaluation highlighted by Hood et al., 2015). Given its emphasis on the lived experience of program participants, early versions of CRE tended to emphasize qualitative methods. More recent writings about CRE have taken more of a mixed-methods approach.

BOX BEGINS HERE Manswell Butty, Reid and LaPoint (2004) describe a culturally responsive evaluation of school-to-career intervention program implemented in an urban junior high school. The program provided a "career breakfast club," with eight one-hour sessions. Among other things, these sessions included information and discussion on such topics as pathways to college, completing job applications, and the array of available education, training, and career options. Manswell Butty and colleagues illustrated the way CRE informed their evaluation, organized in terms of the eight phases as presented by Frierson, Hood and Hughes (2002). With respect to preparing for the evaluation and engaging stakeholders, the evaluators held multiple initial meetings with various stakeholders, striving to develop genuine collaboration and to freely debate ideas and plans (which was enhanced by the support of key school leaders). In terms of identifying the goals and purpose of the evaluation, the evaluators sought to obtain and respond to stakeholder input, and as a result engaged in both formative and summative evaluation activities with a mixed-methods design. The formative component drew heavily on student feedback after each session to guide improvements with the goal of enhancing participants' engagement. The summative component included qualitative and quantitative measures and a

pretest-posttest design with a non-program comparison group. Question framing and the design of the evaluation were also guided by stakeholder input and, as throughout the process, a desire to be culturally responsive. For example, Manswell Butty et al. report that the students preferred discussions and other interactive activities to completing surveys, which the evaluators took into account in the data collection plan. Considerable attention also went into the selection and development of measures, to ensure they were culturally sensitive. In the case of one instrument that had been developed and normed on a population unlike that of the participating school, supplementary data were collected. As regards data collection, members of the evaluation team, by virtue of their intensive engagement with and sensitivity to the context, strove to be aware of the program's cultural context. The evaluators' cultural awareness (aided by team members' background similarities with key stakeholders) probably also aided them in drawing conclusions from the findings, but also important was obtaining input from stakeholders on how to disaggregate the data and when the findings meant. Findings were reported to the full range of stakeholders, including a reporting out to students in what we are told was a student-friendly manner. END OF BOX

Given its more recent development and status here as an instant classic, talking about CRE's current influence is less relevant than it is for older evaluation theories. However, one notable point of influence is in the development of "equitable evaluation," which provides a framework and support for foundations that fund various initiatives and their evaluation (Center for Evaluation Innovation, 2017). Second, CRE's influence is likely to expand further given the first Executive Order from President Joe Biden, Advancing Racial Equity and Support for Underserved Communities Through the Federal Government (Executive Order 13985). It is also worth noting that CRE has overlap with other contemporary evaluation approaches that Mertens and Wilson (2019) place upon the social justice branch of the expanded evaluation theory tree. These approaches include but are not limited to indigenous evaluation theory, such as that based on evaluation by and for the Maori people (e.g., Cram, 2009), disability rights and deaf rights theory (e.g., Mertens, Sullivan & Stace, 2011), and feminist evaluation theory (e.g., Brisolara, Seigert & SenGupta, 2014).

Ways for Evaluation Theory to Guide Evaluation Practice

In the previous section we examined a small number of evaluation theories, with a classic one from each of the three branches of the original theory tree. We also examined a newer, instant classic evaluation theory from, as you prefer, the social justice branch of the revised tree or the social justice sub-branch of the valuing branch. We looked at each theory from the lens of Shadish and colleagues' five component model. We now turn to a critical question that is less frequently addressed in the literature on evaluation theory. That is, how exactly are these theories supposed to guide practice? How might a practicing evaluator translate from the generalities of an evaluation theory to the specific activities to be implemented for a particular evaluation?

Adhering to a theory

One approach, arguably employed too frequently, is to identify an evaluation theory to use as a guide, and then implement it in most, if not all, of one's evaluations. This is the approach people often take when they identify themselves as a **realist evaluator**, or an **empowerment evaluator**,

or as an adherent to some other theory. This would be fine if the chosen evaluation theory fits well with all the varied circumstances in which evaluations takes place, but this seems unlikely for most evaluation theories. It would also be acceptable if an adherent of theory X did evaluations only when that theory fit well to the circumstances. This sorting could occur either if the evaluator was able to take assignments only in circumstances in which theory X fits well, or if the market effectively selected evaluators only for situations in which their preferred theories fit well. While this is possible, especially for evaluators who work in a narrow niche, it does not seem likely for the most part. Moreover, one of the risks of strong identification with a particular evaluation theory is the possibility that you will assume it fits well with a broader range of circumstances than it does.

Matching theory to situation, based on comparison of theories

An alternative way to translate from evaluation theory to evaluation practice involves using multiple theories to enable a kind of matching function. By thinking about the assumptions and (perhaps implicit) context and purposes embedded in each theory, the evaluator may be able to do a better job selecting an appropriate evaluation approach to the specific case. For example, as noted in the previous section, Campbell and Wholey focused on very different aspects of programs. Campbell emphasized evaluative evidence that could inform the selection (or retention, or expansion) of a program, while Wholey focused on information to help managers better manage existing programs. These different emphases led to very different perspectives on what evaluation would look like. A familiarity with evaluation theories that allows this kind of comparisons can aid in the selection of an evaluation approach method that fits the particular circumstances. Context is an important consideration for evaluation practice (Rog et al., 2012), but evaluators often do not have a good sense of which of the countless contextual attributes should be attended to. Being multilingual with respect to evaluation theories helps point an evaluator to key aspects of context, as illustrated by the juxtaposition of Campbell and Wholey.

Still, a problem can arise from over-adherence to a single evaluation theory, especially when following that theory across situations, but even when applying it after matching theory to the situation. Existing evaluation theories were developed in light of the experiences (and the reading and other influences) of the evaluation theorist, and these may not match all of the specific details of the case at hand. That is, a single theory may provide a general guideline, but probably is not nuanced enough to guide all the decisions for the multi-faceted context of a given evaluation. An evaluation theory should not be treated as a paint by number template or an Ikea-like set of instructions. Admittedly, painting by numbers is easier and takes less judgment, but it is less likely to lead to evaluations that bring benefits, especially if the evaluator only has one paint by number template.

Combining theories I: Fully integrating two theories

An alternative to situational matching is to combine evaluation theories. One version of this involves an attempt to fully integrate two approaches, so that neither predominates. Thomas and Parsons (2017) discussed the integration of two theoretical perspectives, one addressed earlier in the chapter and the other not: CRE and systems-oriented evaluation. Thomas and Parsons also

illustrated an evaluation based on this integration, specifically and evaluation of a science, technology, engineering, and math (STEM) education program.

In addition to the characteristics of CRE sketched earlier in this chapter, the systems-oriented approach brought attention to topics such as: the systems, both formal and informal, that are part of the context within which the project operates; the broader history of STEM instruction and learning relevant to the project; the connections among people, perspectives and the like that influence the power dynamics in and around the project; and the formal and informal systems and their elements that foster or inhibit program operations. Thomas and Parsons did not try to bring one element of Theory A into an evaluation primarily shaped by Theory B. Instead, they sought a thorough combination and integration.

Complete, equal integration of two approaches may be challenging. It probably requires a team with expertise in each approach (which Thomas and Parsons provided), as well as patience, teamwork, and likely more time and resources than if one theory were in the lead. In addition, aspects of some theories appear not to be compatible. To take but one example, Campbell's view of the evaluator as servant to the experimenting society might be hard to mesh seamlessly with Scriven's view that the evaluator needs to synthesize finding into a comprehensive evaluative judgment.

Combining theories II: Mixing and matching

Yet another approach is not to consider evaluation theories in whole, but to think of theories as having parts or modules that might be combined in various ways. For example, various evaluation theories provide widely differing guidance for selecting the values that are embedded in an evaluation. Campbell, as suggested by his description of evaluators as servants of the experimenting society, implied that valuing was the business of others. That is, matters such as what program outcomes to measure, or how to draw a conclusion about a program in the face of mixed results, should be the responsibility of the duly elected or appointed officials who are charged with making decisions about the program. Wholey took a similar stance, relying largely on program managers and perhaps policymakers. In contrast, as indicated in goal-free evaluation, Scriven does not give the same standing to public officials or other decisions makers. Instead, Scriven contends that the evaluator should identify the real needs that a program was intended to address. One could, however, imagine an evaluation generally consistent with any of these three theories, but with a different approach to valuing substituted. For example, several of the practices of CRE could be combined with what otherwise would look like an evaluation inspired by Campbell or Wholey.

A broader review of evaluation theories suggests even more options for mixing and matching. Another evaluation theorist, Ernie House, has advocated over time for two different positions regarding valuing. Indeed, the version of the evaluation theory tree shown in Figure 4.1 (Christie & Alkin, 2008) lists House in two places on the values branch! Earlier in his career, House (1976) was influenced by the philosopher John Rawls' book, *A theory of justice*. A philosophical tome that was as deep as it was long (560 pages), the book addressed the topic of the fair distribution of resources. One of the principles Rawls argued for was that inequalities were fair only if they were to the benefit of the least well off. Drawing on Rawls' argument,

House contended that evaluations should reflect the vantage point of the least well off. From this perspective, knowing that a program brings benefits on average is not sufficient. Instead, one must know whether the program benefits the most disadvantaged, such as by closing performance gaps. One of the notable things about House's Rawlsian perspective is that it brings a values stance to evaluation that comes from outside the programming or decision-making world. House does not suggest that evaluators should apply Rawls because it serves the stated information needs of managers, policy makers, or other stakeholders. He advocates this principle, drawn from philosophy, as applying generally to evaluation. This differs from Campbell's position that valuing is the job of someone, as the phrase goes, at a higher pay grade than evaluators, or from Scriven's view that real client needs are the basis for valuing. Familiarity with multiple evaluation theories can broaden an evaluator's perspectives on the range of options that exist for various tasks.

House, along with his collaborator Ken Howe (1999), is also associated with another evaluation theory. In brief, House and Howe's "deliberative democratic evaluation" specifies a set of procedures for adjudicating value issues in evaluation. In essence, House and Howe provide a set of methods for selecting and engaging representatives from multiple stakeholder groups, and facilitating discussion in order to hash out values issues, such as which possible outcomes to measure and how to draw conclusions in the face of mixed results. House and Howe (1999) fall on the social justice branch as articulated by Mertens and Wilson (2019), along with CRE, indigenous evaluation, and other approaches. These share a general emphasis on inclusion of often-overlooked voices in determining the values and questions that drive an evaluation.

This kind of cross-comparative review of the positions of a number of evaluation theories could be expanded, adding evaluation theories not reviewed in this chapter to our comparison of House and Howe, House's earlier Rawlsian approach, CRE, Campbell, Wholey, and Scriven on where and why the questions that guide an evaluation are supposed to come from. More importantly, such comparisons can be useful in several ways. At the least, considering the array of options can be a kind of mental "stretching exercise," expanding one's views about the available options. It might provide you with support if you want to argue within your organization, or in a proposal, or with a program officer, about doing different than past practice. Or, again, such a comparison across evaluation theories might guide an effort to merge one portion of one evaluation theory onto another. For instance, an evaluator following Campbell's or Wholey's approach in general might add on House and Howe's extensive stakeholder procedures for valuing, or add the lens CRE brings throughout the evaluation. Related to this, an aspect of one theory might serve as an ancillary to another. For instance, an evaluation contract may be written with the (perhaps implicit) assumption that valuing is the responsibility of certain decision makers; but the evaluators might nevertheless be able to engage in a Scriven-based needs assessment to see if other criterial of merit emerge, which the evaluator could suggest be considered.

In short, one approach to translating from evaluation theory to evaluation practice involves eclecticism, or a kind of integrative effort.

Applying a contingency theory

Several evaluation theories emphasize criteria for choosing from among alternative ways of doing evaluation. For some evaluation theories, which can be called **contingency models,** this issue is central. It is at the very core of the theory. At least three different kinds of contingency models exist.

One kind of contingency model is based on the idea that **program stage**, that is, where a program is in its lifecycle, should be a key driver of the kind of evaluation to be done. A relatively simple and early form involves the suggestion that one starts with more program-improvement or formative evaluation, and then shifts towards more thumbs-up/thumbs-down, or summative evaluation later on in the life of a program (Cronbach and Associates, 1980). Wholey also suggested a kind of stage model, as implied by his advocacy of the sequential purchase of evaluation. More detailed forms of program stage models have been proposed by Chen (2014) and by Schreirer (2012). In addition to program stage, Chen considers whether the anticipated use of the evaluation is internal to the program or its broader organization, or for external accountability purposes.

Michael Quinn Patton's popular utilization-focused evaluation illustrates another kind of contingency model based on **intended use.** Central to this kind of contingency model is the idea that the choice of what an evaluation should look like should be driven by what kind of use is sought. The key phrase for Patton is "intended use by intended users." Utilization-focused evaluation can include virtually any kind of evaluation, whether it looks like Campbell, Wholey, Scriven, CRE, any other evaluation theorist's approach, or is an eclectic mix, as long as doing the evaluation like that will facilitate intended use by whoever the intended users may be. Key to the practice of utilization-focused evaluation are techniques for identifying intended users and then considering options with them to identify the intended use.

A third kind of contingency model involves a more **policy analytic** assessment of the potential contribution of different kinds of evaluation. An example of this kind of contingency theory is given by Mark, Henry, and Julnes (2000). Rather than identifying and interacting with intended users, it involves an analysis as to which kinds of decisions about the program are more (or less) feasible in the circumstances at hand. Sometimes, for example, it is clear that a program will not be discontinued for the foreseeable future. The political winds may be too strong, or the timing wrong. But at the same time there may be room to look for opportunities for program improvement, or to test the relative effectiveness of program variants. In contrast, under other circumstances, it may be predictable that consideration will be given to replacing the current program, such as when a program is scheduled for reauthorization or the organization includes sunset clauses. These two different circumstances, according to this kind of contingency model, call out for different forms of evaluation. More generally, this contingency theory calls for a form of situational analysis, of the relative feasibility (and potential payoff) of different forms of actions that might be taken regarding a program and drawing preferences about types of evaluation activities accordingly.

I suggest there are benefits, not only of being familiar with a contingency theory, but also of being "multi-lingual" with respect to such models. The relative fit of the different kinds of contingency models is likely to vary across situations. A slogan might be, "even the contingencies are contingent." For example, it may be easy to engage in discussions with an

intended user in some cases, but quite infeasible in others (e.g., when evaluation is intended to inform the U.S. Congress). To take another example, a desire to engage in extended formative evaluation prior to summative evaluation, as would be encouraged by a program stage-based theory, might be overridden by a need to be responsive to the decision-making timeline specified in legislation. The templates provided by contingency theories of evaluation might fit more situations than simpler evaluation theories, but probably not all. Judgment is still required.

The future of evaluation theory

In this final section of the chapter, a few suggestions are offered for the future development and application of evaluation theory.

Theory-testing, theory-building, and research on evaluation

As noted earlier, some evaluators differentiate evaluation theories into different kinds. For example, according to Alkin, (2004, p. 4), "there are two general types of models: 1) *A prescriptive model*, the most common type, is a set of rules, prescriptions and prohibitions, and the guiding frameworks which specify what a good or proper evaluation is and how evaluation should be done. Such models serve as *exemplars*; 2) *A descriptive model* is a set of statements and generalizations which describes, predicts, or explains evaluation activities. Such a model is designed to offer an *empirical theory*." The evaluation theories that have described in this chapter are predominantly prescriptive models. Shadish, Cook & Leviton (1991, pp. 30-31) referred to something akin to Alkin's (2004) notion of empirical theory. In describing "The ideal (never achievable) evaluation theory" Shadish and colleagues indicated that, among other characteristics, it would "empirically test propositions to identify and address those that conflict with research or other critically appraised knowledge about evaluation." In other words, the developers of both of our meta-models advocate for the increased use of empirical research to test tenets of evaluation theories.

Notable examples of such empirical testing of evaluation theories already exist in the evaluation literature. One focused on empowerment evaluation, a form of collaborative evaluation that was developed by David Fetterman (e.g., Fetterman (1994); also see Chapter X). Miller and Campbell (2006) examined 47 cases described by their authors as empowerment evaluations. Miller and Campbell examined two issues. First was whether the evaluations followed the principles of empowerment evaluation, and thus could be differentiated from other related approaches. Second, Miller and Campbell asked what evidence existed that the empowerment evaluations actually resulted in increased empowerment for those who participated in the evaluation process or for the intended beneficiaries of the program being evaluation. Miller and Campbell reported that collectively the 47 case examples did not do well with respect to either issue. The review by Miller and Campbell addressed a key issue for those who conduct research on evaluation: Do evaluation theories, models or approaches actually result in the benefits they claim will result (Mark, 2008)?

A second notable example comes from a 2012 book by Brad Cousins, an advocate and developer of another collaborative approach to evaluation, practical participatory evaluation, and his ex-student and colleague, Jill Anne Chouinard. Like Fetterman, Cousins falls on the use branch of

the evaluation theory tree. In their 2012 book Cousins and Chouinard synthesize the research by Cousins and others practical participatory evaluation, with a key lesson being that engaging stakeholders can in fact increase evaluation use. Related to the topic of engaging stakeholders, another empirical contribution to a future descriptive theory of evaluation comes from Shulha, Whitmore, Cousins, Gilbert, & al Hudib (2016), who in a multi-phase inquiry, drew on the responses of those experienced in conducting collaborative evaluation to derive and validate a set of principles for collaborative evaluation practice. Cousins (2020) draws on both of the contributions just mentioned in an edited volume. As yet another example, Mark, Allen and Goodwin (2021) showed that those with a stake in an evaluation prefer that multiple stakeholder groups be involved.

Not only should there be increased research that tests the propositions embedded in one or more evaluation theories. In addition, those developing new evaluation theory should emphasize the research (as well as practice) base on which they draw. We should remember that empirical findings from an earlier period of research on evaluation, sometimes called a "golden age" research on evaluation, had strong impact on evaluation theory and concepts. To take one example, Michael Patton found that when direct, instrumental use occurred, there often was an individual in the organization who had served as an advocate for evaluation use. This finding led to the concept of the "personal factor" and served as an underpinning of Patton's influential utilization-focused evaluation (Patton, 1978, 2008). In a second example, the idea of "enlightenment" or conceptual use of research was a research finding (Oral History Project Team, 2006). It arose when Carol Weiss found that, although only a minority of her study participants reported engaging in direct, instrumental use, most also indicated that an evaluation affected the way they thought about the underlying social problem or its potential solutions. This finding helped broaden evaluators' conception of use, was part of Weiss' legacy as an evaluation theorist, and has influenced subsequent approaches such as program-theory driven evaluation.

Evidence suggests that research on evaluation is increasing in frequency (Coryn et al., 2017). However, much remains to be done to tie empirical findings back to evaluation theory and practice. This would be a worthwhile area of contribution that readers of this chapter might make prior to the next edition of this Handbook.

More people shaping evaluation theory, often with less comprehensive scope

Elsewhere I have suggested the value of encouraging more members of the evaluation community to be actively involved in evaluation theory, perhaps with the "slogan, 'Evaluation theory is too important to be left in the hands of a few people labeled as evaluation theorists." (Mark, 2018). It would be a sign of the field maturing, I believe, if we shift away from seeing evaluation theory as the realm of a few people who end up on a theory tree. As in other fields, some individuals might elect to specialize in theory building, but many others might participate in testing and refining a given theory or cluster of theories. And even more are likely to try to apply the lessons in practice.

The latest version of the Evaluation Theory Tree (Christie & Alkin, in press) is a notable step in that direction. As shown in Figure 4.3, the tree now lists theory names. Further progress remains, however, given that almost every chapter is authored by a single theory developer or, in

the case of theorists no longer living, the best available proxy (Another notable change in the tree is the inclusion of additional social justice oriented theories on the valuing branch). Perhaps future editions of the tree will include more contributions from people other than the original theory developer, these being individuals who have continued to modify, update, and test an evaluation theory or set of related theories.

Insert Figure 4.3 about here

Another potentially beneficial step would be to shift, at least partially, away from such extensive focus on relatively comprehensive, five-component evaluation *theories* as detailed by Shadish et al. (1991). Instead, both theory developers and theory testers could increasingly concentrate on one or more theoretical *issues.* Some people might specialize broadly, with an interest such as valuing or one of the other five components of the Shadish, Cook and Leviton model. Others might specialize more narrowly, such as on stakeholder participation or even more specifically on stakeholder participation within, say, empowerment evaluation. If people specialize in specific issues within an evaluation theory, not only is depth of expertise encouraged, but also the cost of entry into theory testing and theory formulation becomes smaller. And it would be valuable to have more people involved in developing and trialing evaluation theories, and in conducting and synthesizing research on evaluation theory. Over time, then, sharable evidence more than assertion would undergird the guidance that comes from evaluation theories.

Evaluation theory and lessons for non-evaluators

This chapter has for the most part side-stepped a very important constraint on the potential application of evaluation theory to evaluation practice. That is, the shape of an evaluation is often specified, not by an evaluator, but by the staff of the agency that houses a program, or the oversight body, or a funder. A potentially important future project might be to try to consolidate lessons from evaluation theory for evaluation practice, and to disseminate them in formats and venues that could inform those, other than the evaluator, who influence the practice of evaluation.

Multi-lingualism, contingency models, and theory as an aid to judgment

Adherence to a single evaluation theory is probably not be the best way forward. True believers in a theoretical approach may bring some benefits, such as pushing a single evaluation approach as far as it can go. But the alternative kinds of contributions that evaluation can bring, and the disparate nature of programs and their broader contexts, are so varied that a single approach to evaluation is not likely to optimize the contributions that the practice of evaluation can bring. Instead, in the ideal, evaluators should be multilingual with respect to evaluation theory. And they should treat evaluation theories as a potential aid to judgment about practice, not as a recipe book that replaces judgment.

BOX STARTS HERE Claire started the morning in a familiar way, checking her work emails with a cup of coffee. She had really been triaging these for more than a week, having been distracted by her stepson's wedding. As she scanned the long queue in her inbox, Claire was pleased to see the increase in requests for proposals and other notices about upcoming

evaluations. As she started digging into the emails and links, she flashed back to the chapter she'd read some time ago about evaluation theory and the online class it has led her to take. One of the first emails Claire opened was from a regional foundation she had done some work for in the past. The email announced a fairly open-ended call for proposals. Claire's first thought was about proposing an evaluability assessment about the foundation's grantmaking as a whole. The foundation's grants had evolved in various directions as program officers came and went. Hmm, another thought was to try to help lay out a variety of evaluation activities the foundation may use, and why, working with the foundation's new Director of Monitoring, Evaluation, and Learning. Claire next turned her attention to an email from a colleague at a large contract research firm. The firm had considerable expertise in randomized experiments and quasi-experiments, and planned to go after a contract for a large RCT. From what her friend said, it seemed to Claire like the circumstances suited such an evaluation. The friend mentioned he knew Claire had never done an RCT, but he and his colleagues liked what she had brought conceptually to the evaluation she had partnered on with them before. The RFP specified the design as well as some outcome measures. But Claire wondered whether there would be value added to propose additional effort to consider other measures, drawing on the values components of several evaluation theories she'd studied. As she reviewed the rest of her emails, and as she started making some notes about a couple of possible proposals, Claire mused occasionally about how much she enjoyed her stepson's marriage. And, she thought as another smile appeared, the marriage of evaluation theory and evaluation practice isn't too bad either. BOX ENDS

Suggested Resources

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

Alkin, M. C. (Ed.) (2013). *Evaluation roots: A wider perspective of theorists' views and influences*. Thousand Oaks, CA: Sage.

Mertens, D. M. & Wilson, A. T. (2019). *Program evaluation theory and practice: A comprehensive guide* (2nd ed.). New York: Guilford.

Additional References

- Alkin, M. C. (2004). Comparing evaluation points of view. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 3-11). Thousand Oaks, CA: Sage.
- Alkin, M. C. & Christie, C. A. (2004a). An evaluation theory tree. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 12-65). Thousand Oaks, CA: Sage.
- Alkin, M. C. & Christie, C. A. (2004a). Evaluation theory tree revisited. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 381-392). Thousand Oaks, CA: Sage.
- Bhaskar, R. A. (1978). *A realist theory of science*. Atlantic Highlands, NJ: Humanities Press.

- Bickman, L. (1987), The importance of program theory. In L. Bickman (Ed.), *Using program theory in evaluation*. Jossey Bass, San Francisco (1987), pp. 5-18.
- Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, 24, 409-429.
- Campbell, D. T. (1991). Methods for the experimenting society. *Evaluation Practice*, *12*(3), 223–260.
- Center for Evaluation Innovation, Institute for Foundation and Donor Learning, Dorothy A Johnson Center For Philanthropy, Luminare Group (2017, July). *Equitable Evaluation Framework (EEF) Framing Paper*. Equitable Evaluation Initiative, www.equitableeval.org
- Chen, H. T. (1990). *Theory-driven evaluations*. Newbury Park, CA: Sage.
- Chen, H. T. (2014). *Practical program evaluation: Theory-driven evaluation and the integrated evaluation perspective*. Sage Publications.
- Christie, C. A. & Alkin, M. C. (2004). An evaluation theory tree. In M. C. Alkin (Ed.), *Evaluation roots: A wider perspective of theorists' views and influences* (pp. 11-57). Thousand Oaks, CA: Sage.
- Christie, C. A. & Alkin, M. C. (2008). Evaluation theory tree re-examined. Studies in Educational Evaluation, 34(3), 131-135.
- Christie, C. A. & Alkin, M. C. (in press). An evaluation theory tree. In M. C. Alkin & C.A. Christie (Eds.), *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage. UPDATE REF
- Coryn, C. L., Noakes, L. A., Westine, C. D., & Schröter, D. C. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American journal of Evaluation*, *32*(2), 199-226.
- Cousins, J. B. (2020). *Collaborative Approaches to Evaluation: Principles in Use*. Sage.
- Cousins, J. B., & Chouinard, J. A. (2012) *Participatory evaluation up close: A review and integration of the research base*. Charlotte, NC: Information Age Press.
- Cram, F. (2009). Maintaining indigenous voices. *The handbook of social research ethics*, 308-322.
- Cronbach, L. J. and Associates (1985). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.
- Davidson, E. J. (2004). *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. Sage.
- Davidson, E. J. (201x) *Actionable evaluation basics: Getting succinct answers to the most important questions*. Real Evaluation,
- Davies, R., & Payne, L. (2015). Evaluability Assessments: Reflections on a review of the literature. *Evaluation*, *21*(2), 216–231.
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2-21.
- De Lancer Julnes, P. (2006). Performance Measurement: An Effective Tool for Government Accountability? The Debate Goes On. *Evaluation*, *12*(2), 219–235.
- Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. New York, NY: Lawrence Erlbaum.
- Fetterman, D. M. (1994). Empowerment Evaluation. *Evaluation Practice*, *15*(1), 1–15.
- Frierson, H., Hood, S., & Hughes, G. (2002). A guide to conducting culturally responsive evaluation. In *The 2002 User-Friendly Handbook for Project Evaluation* (pp. 63-73). National Science Foundation.

- Frierson, H., Hood, S., Hughes, G. & Thomas, V. (2010). A guide to conducting culturally responsive evaluation. In J. Frechtling (Ed), *The 2010 User-Friendly Handbook for Project Evaluation*. (pp. 75-96). National Science Foundation.
- Holzer, M., & Ballard, A. (Eds.). (2021). *The Public Productivity and Performance Handbook*. Routledge.
- Hood, S. (1998). Responsive evaluation Amistad style: Perspectives of one African American evaluator. In *Proceedings of the Stake symposium on educational evaluation* (pp. 101-112). Champaign: University of Illinois at Urbana–Champaign.
- Hood, S., & Hopson, R. K. (2008). Evaluation Roots Reconsidered: Asa Hilliard, a Fallen Hero in the "Nobody Knows My Name" Project, and African Educational Excellence. *Review of Educational Research, 78*(3), 410–426.
- Hood, S., Hopson, R. K., & Kirkhart, K. E. (2015). Culturally responsive evaluation. *Handbook of practical program evaluation*, *281*-317.
- Hopson, R. K. (2009). Reclaiming knowledge at the margins: Culturally responsive evaluation in the current evaluation moment. *The Sage international handbook of educational evaluation*, 429-446.
- House, E. R. (1976). Justice in evaluation. In G. V. Glass (Ed.), *Evaluation Studies Review Annual*, Vol. 1, Beverly Hills, CA: Sage.
- House, E. R. (2003). Evaluation theory. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation*. Dordrecht/Boston/London: Kluwer Academic.
- House, E. R. & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.
- Lemire, S., Peck, L.R. & Porowski, A. (2020), The growth of the evaluation tree in the policy analysis forest: Recent developments in evaluation. *Policy Studies Journal, 48,* S47-S70.
- Leviton, L. C., Khan, L. K., Rog, D., Dawkins, N. & Cotton, C. (2010). Evaluability assessment to improve public health policies, programs, and practices. *Annual Review of Public Health, 31(1),* 213-233.
- Madous, G. F., Scriven, M., & Stufflebeam, D. L. (1983). *Evaluation models*. Boston: Kluwer-Nijhoff.
- Manswell-Butty, J.-A.L., Reid, M.D. and LaPoint, V. (2004), A culturally responsive evaluation approach applied to the talent development school-to-career intervention program. *New Directions for Evaluation*, 37-47.
- Manswell Butty, J.-A. L., Wakiaga, L. A., McKie, B. K., Thomas, V. G., Green, R. D., Avasthi, N., & Swierzbin, C. L. (2015). Going Full Circle With Teacher Feedback: Conducting Responsive Evaluations in Urban Pre-K Classrooms. *SAGE Open*
- Mark, M. M. (2008). Building a better evidence-base for evaluation theory. In P. R. Brandon & N. L. Smith (eds). *Fundamental issues in evaluation* (111-134). New York: Guilford.
- Mark, M. M. (2018). Strengthening links between evaluation theory and practice, and …: Comments inspired by George Grob's 2017 Eleanor Chelimsky Forum presentation. *American Journal of Evaluation*, 39(1), 133-139.
- Mark, M. M., Allen, J. & Goodwin, J. (2021). Does stakeholder involvement affect perceptions of an evaluation? And which stakeholder groups do people think should participate? *Evaluation Review*

- Mark, M. M. & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation, 10(1),* 35-57.
- Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: and integrated framework for understanding, guiding, and improving policies and programs*. San Francisco, CA: Jossey-Bass.
- Mark, M. M. & Mills, J. (2007). The use of experiments and quasi-experiments in decision making. In G. Morcöl (Ed.), Handbook of Decision Making. (459-482) New York: Marcel Dekker.
- Michielutte, R., Shelton, B., Paskett, E. D., Tatum, C. M. & Valez, R. (2000). Use of an interrupted time-series design to evaluate a cancer screening program. *Health Education Research, 18,* 615-623.
- Miller, R. L., & Campbell, R. (2006). Taking Stock of Empowerment Evaluation: An Empirical Review. *American Journal of Evaluation*, *27*(3), 296–319.
- Oral History Project Team (2006). The oral history of evaluation, part 4: The professional evolution of Carol H. Weiss. *American Journal of Evaluation*, *27*(4), 475–484.
- Patton, M.Q. (2008). *Utilization-focused evaluation,* 4th edition. Thousand Oaks, CA: Sage.
- Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford Publications.
- Rog, D. J., Fitzpatrick, J. L. & Conner, R. F. (2012). *Context: A Framework for its Influence on Evaluation Practice,* New Directions for Evaluation, no. 135. Jossey-Bass.
- Russ-Eft, D. & Sleezer, C. (2019). *Case Studies in Needs Assessment*. Sage.
- Sawhill, J. C. & Williamson, D. (2001). Mission impossible? Measuring success in nonprofit organizations. *Nonprofit Management & Leadership, 11*(3), 371-386.
- Scheirer, M. A. (2012). Planning Evaluation Through the Program Life Cycle. *American Journal of Evaluation, 33*(2), 263–294.
- Scriven, M. (1974). Pros and cons about goal-free *evaluations. Evaluation in Education: Current Applications:* 34–67
- Shadish, W. R. (1996, June). Teaching evaluation theory. *Evaluation News and Comment, 5 (1),* 553.
- Shadish, W. R. (1998). Evaluation theory is who we are. American Journal of Evaluation,
- 19(1), 1-19.
- Shulha, L. M., Whitmore, E., Cousins, J. B., Gilbert, N., & al Hudib, H. (2016). Introducing Evidence-Based Principles to Guide Collaborative Approaches to Evaluation: Results of an Empirical Process. *American Journal of Evaluation*, *37*(2), 193–215.
- Smith, N. L. (2010). Characterizing the Evaluand in Evaluating Theory. *American Journal of Evaluation, 31*(3), 383–389.
- Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation*.
- Stufflebean, D. L. (2001). *Evaluation Models* (New Directions for Evaluation, Number 89). San Francisco: Jossey-Bass.
- Thomas, V. G., & Parsons, B. A. (2017). Culturally Responsive Evaluation Meets Systems-Oriented Evaluation. *American Journal of Evaluation*, *38*(1), 7–28.
- Trevisan, M., & Walser, T. (2015). *Evaluability assessment*. SAGE Publications, Inc.
- White House (2021). *"Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government". January 21, 2021.*

- Wholey, J. S. (2004). Evaluability assessment. *Handbook of practical program evaluation, 2*, 33-62.

Appendix A

As noted elsewhere in this chapter, in a chapter in the *Roots* book on behalf of Campbell (who had died years earlier), Will Shadish argued against the theory tree with its three branches, arguing instead for the five component Shadish, Cook, and Leviton (2001) model. The two meta-models overlap a great deal, of course. Both include use and valuing. In addition, the evaluation theory tree includes a methods branch, which corresponds generally to the knowledge construction component of Shadish, Cook and Leviton. That is, the methods proclivities of an evaluation theorist are for the most part the more specific, practice-oriented techniques of knowledge construction. (An evaluation theory can include methods that align with other components, such as methods for valuing or for facilitating use, but it is methods related to knowledge construction that are emphasized in the theory tree).

Regarding the practice component, perhaps it is not surprising that Alkin and Christie did not include this as a branch on the tree. After all, evaluation theories almost inevitably include a strong emphasis on evaluation practice. At the risk of taking the tree metaphor too far, perhaps one can think of practice as the acorns that fall to the ground from the tree.

On reflection, it also may not be surprising that social programming is not a separate branch on the evaluation theory tree. After all, the theorists represented in the *Roots* book are all involved in program evaluation (although Scriven aspires to be a more general theorist of the evaluation of anything). On the other hand, *if* Alkin had included chapters by theorists of personnel evaluation, and theorists of policy evaluation, and theorists of product evaluation, then the subject of the evaluation, or as Scriven has called it, the evaluand, might have appeared as part of the graphical representation. It seems likely that, rather than adding another branch to the current program evaluation tree, this would have involved drawing multiple trees, one for each kind of evaluand (program, policy, personnel), with each tree with having branches (containing theorists who might appear on only that tree).

Alternatively, one could imagine the development of alternative approaches to program evaluation, with each premised on a different view of social programs, their role and operation. For example, one evaluation theory view might assume a competitive market of privately-owned programs and center that feature, while another evaluation theory might focus on government-run programs that rarely are discontinued and give priority to the characteristics of such programs in laying out its approach to evaluation. Or an evaluation theory could give major focus to the way programs vary in their complexity, such as in the number of distinct components and the expected or allowed adaptability (Mark, in press), laying out alternative evaluation approaches based on the degree of complexity of the program. In such an alternative world, there may have been need for a social program branch on the evaluation theory tree. But

given the absence of different schools of thought about social programs as underpinnings of evaluation theories, Alkin and Christie's theory tree has three main branches.