Credible and Actionable Evidence : A Framework, Overview, and Suggestions for Future Practice and Research

In: Credible and Actionable Evidence: The Foundation for Rigorous and Influential Evaluations

By: Melvin M. Mark

Pub. Date: 2017
Access Date: October 26, 2021
Publishing Company: SAGE Publications, Inc
City: Thousand Oaks
Print ISBN: 9781483306254
Online ISBN: 9781483385839
DOI: https://dx.doi.org/10.4135/9781483385839
Print pages: 275-302

Credible and Actionable Evidence : A Framework, Overview, and Suggestions for Future Practice and Research

Melvin M. Mark

Applied social research, including program and policy evaluation, is typically undertaken with the hope that it will have positive consequences. For instance, an evaluator may hope that her findings will lead to the selection and retention of more effective programs. Or an evaluator may hope that his evaluation, with its recommendations about possible changes, will lead to improved processes and outcomes in an ongoing program. Alternatively, an evaluator might hope to contribute in the long run by improving people's understandings of a social problem and its potential solutions or by increasing capacity within an organization.

In most cases, for an evaluation or for any applied social research to lead to any of these or other positive consequences, somebody has to think that the findings are believable and that they can and should be acted on. In other words, the study and its findings must be taken as credible and actionable. This book focuses on the nature and determinants of credibility and actionability as well as their possible role within broader views about the use of evidence. Choice of research methods, not surprisingly, is one of the potential determinants of credibility and actionability. Accordingly, many of the preceding chapters address the consequences of method choice on credibility and actionability.

Readers of these chapters will have seen that although there are areas of general agreement, complete consensus does not exist about what constitutes credible and actionable evidence. This chapter builds (largely) on ideas presented in previous chapters, with the hope of providing a model that might contribute to a greater sense of consensus where agreement exists and to a clearer understanding of why any disagreements remain. The chapter is not meant to be a comprehensive review or summary of the preceding chapters. Instead, I draw selectively on other chapters in this book in an attempt to sketch out a more comprehensive model. At times, I refer to chapters in the previous edition of this book (Donaldson, Christie, & Mark, 2009) as well as to other work. I begin with a brief discussion of credibility, actionability, and other attributes of evidence. I subsequently consider these attributes in the context of a broader general framework.

Christie and Fleischer (Chapter 2) ground contemporary debates about evaluation methods within different research traditions and paradigms. The current chapter aspires to contribute to moving those debates beyond matters of methods and paradigms. In service of this aspiration, credibility and the other attributes of evidence, along with the broader framework in which they reside, are used here as a lens for considering the reasons that underlie different method and practice preferences. My hope is that if we better understand *why* people disagree about method preferences, we might be able to have more productive discussions in the future. That is, we might move from relatively unproductive debates centered on methods to more productive discussions focusing more on the considerations that underlie diverging preferences for methods. At the end of chapter, some suggestions for practice are given. In addition, Appendix I includes a relatively detailed

discussion of Michael Scriven's views about randomized experiments.

Credibility and Actionability in a General Framework of Evidence Use, Part 1

In addition to credibility, other characteristics of evidence can also affect whether people are inclined to see evaluation and other research findings as a worthy guide to action. Here I draw on, and expand, a framework that Schwandt (Chapter 12) provides. Specifically, in this section, I discuss credibility, inferential potency, relevance, comprehensiveness, and actionability as attributes of evidence. Explicitly or implicitly, people may make judgments about each of these attributes.

I believe that treating these characteristics separately is helpful for those who want to understand the use of evidence. But this formulation, with its five attributes, is not inevitable. (Indeed, many of the chapters in this volume combine characteristics I treat separately here, such as credibility and inferential potency.) Putting this and other possible caveats aside, I turn now to the five attributes.

Credibility refers, in general terms, to whether something is believed or, put differently, taken to be true or correct. In certain research traditions, the concept of validity has received more emphasis. Validity presumably is a determinant of credibility. However, how strongly a technical assessment of validity causes or predicts credibility judgments will vary. When will credibility judgments be based more on considerations other than validity? When the perceiver lacks the expertise in research methods to judge validity and when the perceiver relies more strongly on other possible cues to credibility (see Miller, Chapter 3). Whether credibility judgments are based on validity or not in a particular case, evidence will often need to be seen as credible if it is to be used, as many chapter authors have noted.

Schwandt identifies *probative* (or *inferential*) *force* as another important characteristic of evidence. Drawing on the philosopher Cartwright, high probative force would be evidence that essentially clenches an argument. If an evaluation has high probative force and indicates that a training program increases employment, this should settle the question about that effect. Probative force can be differentiated from credibility, however. In particular, a finding could be credible without having high probative force. That is, the finding is believable but not completely convincing with respect to the issue at hand. It does not clinch the point. Reichardt and Mark (1998) noted a kind of finding from quantitative research that illustrates the distinction: An experiment might be credible (and valid) but have a large confidence interval. Study results like this are believable (i.e., they are credible), but they leave a too-wide range of uncertainty to clinch the argument (i.e., they have lower probative force). I prefer the term *inferential potency* to *probative force* and will use it here.

Relevance is another criterion that Schwandt mentions. A given program evaluation might provide findings that are both credible and inferentially potent but nevertheless are of limited value—they address issues that are not strongly informative for the decision at hand. They clinch a point, but it's not the right point for the decision at hand. For instance, suppose that information is needed about the quality of program

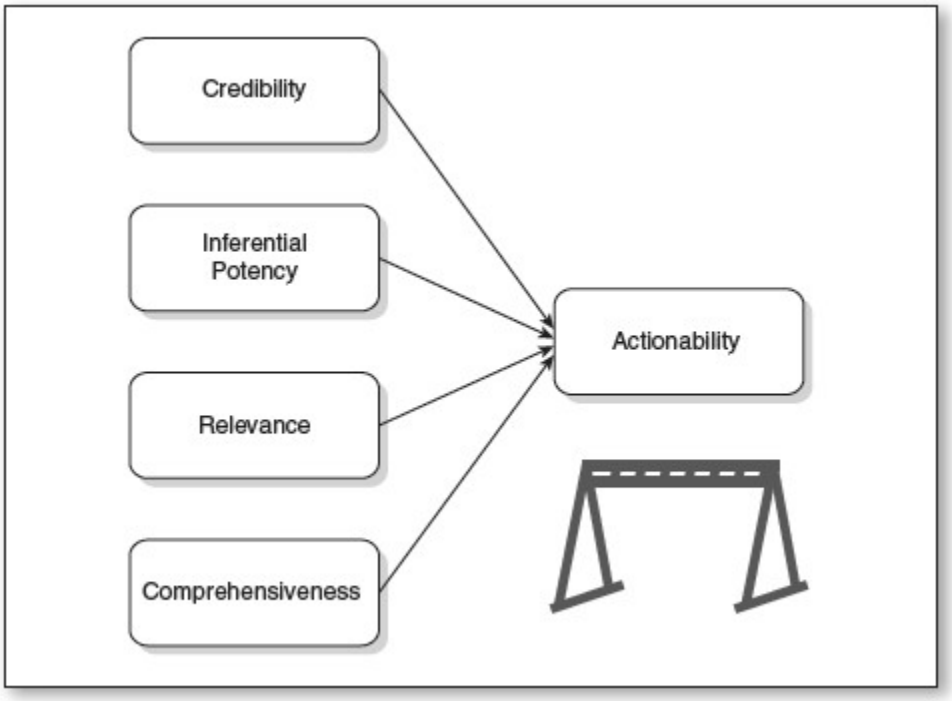
implementation. In that case, an estimate of program effectiveness, such as from a randomized controlled trial (RCT) comparing a treatment and control group, would not be relevant (Rallis, Chapter 7). Relevance is also an issue when one questions whether the findings from an RCT will apply to a new setting (Cartwright & Hardie, 2012). Thus Bickman and Reich's (Chapter 5) and others' discussion of external validity involves the criterion of relevance.

A fourth criterion, *comprehensiveness,* was not explicitly identified by Schwandt. *Comprehensiveness* refers to the extent to which the evidence at hand covers the full range of questions that are of interest. We can further distinguish between *specific comprehensiveness* and *global comprehensiveness.* Consider as an example early evaluations of alternative welfare arrangements, which estimated the effect of then-new welfare practices on employment and income. Those evaluations were lacking in specific comprehensiveness. Decision makers were interested in whether new welfare approaches affected outcomes other than employment and income, such as family configurations (and today, they would also be interested in effects on health and children's well-being). *Specific comprehensiveness*, then, refers to completeness within a particular evaluation task or question, such as, *What are the program's effects?* In contrast, *Global comprehensiveness* refers to completeness across the multiple evaluation questions that are of interest. For instance, imagine an evaluation that gives relatively comprehensive evidence about the relevant range of program effects. This evaluation could nevertheless lack global comprehensiveness if it does not provide information that stakeholders want about the relative cost of implementing the new program and the quality with which it would be implemented at varied locations.

Trade-offs in study planning often involve trade-offs among these four criteria. For example, assuming a given level of funding, efforts to increase comprehensiveness may come at a cost of credibility, inferential potency, or both. If more questions are to be addressed in a study, the quality of the answer to any single question may suffer. Trade-offs in study planning can, of course, also involve trade-offs between one or more of these criteria and the level of resources that are available. An evaluator might be able to increase comprehensiveness without a loss of credibility *if* the study budget and timeframe are increased.

As shown on Figure 13.1, credibility, inferential potency, relevance, and comprehensiveness presumably combine to determine judgments of actionability. *Actionability* refers to the perceived degree to which evidence is suitable as a guide to possible action. Little is known about the relative importance of the four criteria as precursors of actionability judgments. To take but one specific example, one might ask how important comprehensiveness is for perceived actionability relative to credibility and the other attributes. (As a speculative response, the importance of comprehensiveness may vary, depending on whether or not multiple evaluation questions are salient.) Nor is a great deal known about the extent to which the four criteria are related or affect each other. For example, if evidence is seen as credible, does that make it seem more relevant? Clarifying the relationships among the criteria in Figure 13.1 would be a potentially valuable research agenda. It could lead to improved understanding of the way people do or do not use research and evaluation evidence. (Note that arrows could be included in Figure 13.1 to represent possible interrelations among the four criteria to the left; these are omitted for the sake of simplicity.)

Figure 13.1 Credibility, Other Characteristics of Evidence, Actionability, and the Threshold for Action



The Threshold for Actionability

In principle, judgments about actionability could be conceptualized on a continuum, with multiple gradations ranging from very low to very high. In practice, however, the decision is more of a dichotomous, go/no-go, act/don't-act kind of choice. The image of a bar or hurdle under actionability in Figure 13.1 is meant to represent the threshold for deciding to act. The height of the bar almost certainly varies across circumstances. Sometimes the threshold is high, sometimes not.

Bickman and Reich (Chapter 5) allude to this idea when they refer to the cost of being wrong about a causal relationship. These costs are not uniform, as Bickman and Reich suggest with their example of drugs that vary in cost. Julnes and Rog (Chapter 11) also address the height of the metaphoric hurdle. They refer to the confidence required for a conclusion. In this regard, Julnes and Rog note that incremental program changes, such as the choice of activities for a one-day teacher in-service program, typically require less confidence than the choice between major program alternatives, such as the selection of an integrated K–12 math curriculum.

Taking Julnes and Rog's discussion a step further, thresholds for action may vary, even across instances of choosing between major program alternatives. When a mandate exists to implement *some* alternative program (e.g., when legislation requires schools to implement a substance abuse program), the only choice is *which* program to implement. Weaker or less-comprehensive evidence about a program may suffice, relative to what would be needed if instead there were no mandate and continuing without a substance abuse

program was therefore an option.

More generally, the hurdle image in Figure 13.1 conveys the idea that the aggregate quality required for evidence to support action is not a constant. Action will be seen as justified in some circumstances by less impressive evidence, while in other instances, more impressive evidence may be required. At times, the height of the hurdle is implicit. At other times, it is the subject of intense, explicit debate. This was the case, for example, when advocates fought for lower evidence requirements for the approval of treatments for HIV/AIDS in the 1980s and 1990s.

Additional Factors in a Broader Framework

The criteria in Figure 13.1 can be seen as one part of a broader framework. Judgments about credibility, comprehensiveness, and the other criteria do not take place in a vacuum. Instead, they occur within, and are influenced by, multiple levels of context. Expanding on this general idea, Julnes and Rog (Chapter 11) provide a framework suggesting how method choices can be guided by thinking about five aspects of context: (a) alternative evaluation questions and tasks that might be addressed, (b) attributes of the phenomenon being investigated, such as whether change is likely to be easy and quick or difficult and slow, (c) characteristics of the program or policy to be investigated, (d) specific circumstances of the evaluation or study, and (e) more general characteristics of the context in which the program or policy operates.

In addition, judgments about credibility and the other criteria can be affected by a variety of processes. Some of these are individual and psychological in nature. Others are interpersonal and involve social interactions. Yet others involve social aggregates and may be political or organizational in nature. As one example of these kinds of processes, Miller (Chapter 3) reviews in considerable detail the individual psychological processes that can affect judgments of credibility and also have implications for the other criteria in Figure 13.1.

I return later to these additional aspects of a broader framework. Prior to that, I consider how the criteria listed in Figure 13.1 can clarify divergent views about method choice in evaluation.

The RCT Debate, Illuminated in Terms of Credibility and Other Criteria

As noted in Chapter 1 and many of the other chapters in this volume, the last dozen years or so have seen considerable debate about the role of randomized experiments in evaluation. One of the tripwires starting these skirmishes was a proposal, subsequently implemented, by the Department of Education through its Institute of Education Sciences (IES) to establish a priority for randomized trials for selected funding programs. In essence, this priority, when applied to a funding program, gives extra points to proposals that use randomized experiments when review panels discuss and score the proposals. Randomized experiments also receive special consideration in the review process at the What Works Clearinghouse (WWC), which was designed to identify educational interventions that have demonstrated effectiveness with credible evidence (Gersten & Hitchcock, 2009). In education, this relatively newfound emphasis on randomized experiments

arose in a context that had led some observers to argue that the field of educational research had been underusing randomized trials relative to their potential benefits (e.g., Cook, 2003). Henry (Chapter 4) reviews the actual funding patterns since the IES priority was implemented—adding a dose of evidence to what has mostly been a data-free debate about evidence.

The move to give preference to randomized experiments (and the closest quasi-experimental approximations) has not been limited to education. This preference existed previously in medicine, where the term *RCT* (for *randomized clinical trial*) was popularized (users of term outside of medicine often refer to *control* or *controlled* rather than *clinical* trials). In other areas of practice, such as "development" evaluation (that is, the evaluation of international aid to developing countries), there have also been efforts to increase attention to RCTs. This includes the efforts of the International Initiative for Impact Evaluation or 3ie. 3ie, with support from multiple foundations and national development agencies, advocates for and funds impact evaluations and systematic reviews, with a preference for randomized experiments and strong quasi-experiments. More generally, evidence hierarchies are one of the ways that preferences for RCTs are expressed. An evidence hierarchy essentially rank orders methods from best to worst, and RCTs typically fall at or near the top.

There may currently be less heat than was initially the case around these and other instances of advocacy about RCTs and related methods. At the same time, there may not yet be the desired increase in light to illuminate the important task of choosing methods for evaluation and applied social research. By thinking about the various characteristics in Figure 13.1 and the different views held by advocates and opponents of a priority for RCT, we might find a way to more productive conversations.

Method Choices in Light of Relevance

The underlying differences that help motivate the two sides in the RCT debate can be clarified in terms of the factors listed in Figure 13.1. First, consider relevance. Advocates of more widespread use of RCTs presume that the findings from such studies are generally relevant for key decisions about programs and policies. Keep in mind that although RCTs can be expanded to address other questions, their basic contribution—if the experiment is successfully conducted—is an unbiased estimate of the average effect of the treatment (e.g., the mean difference in the outcome variable between the program group and the control/comparison group).

Advocates of RCTs hold the question of average program effects in high regard. They believe this question is commonly quite relevant in deliberations about programs and policies. Specifically, advocates of RCTs believe that a valuable form of use (if not *the most* valuable form) occurs if decision makers draw upon evidence of an intervention's average effects when they make choices about program or policy initiation, expansion, maintenance, or cessation. For example, Bickman and Reich (Chapter 5) make this point when talking about "the cost of making a wrong decision about causality." They state,

> We know that there are costs in making the wrong decision. To call a program effective when it is not means that valuable resources may be wasted and the search for other means to solve the problem will be hindered. Some programs may not only be ineffective but also harmful. In such

> cases, the costs of a wrong decision would be higher. On the other hand, falsely labeling a program as ineffective would mean that clients who would have benefited from the intervention would not have that benefit. (pp. 85–86)

Henry expands this position by explicitly locating it in the context of democratic decision making within representative democracies. The choice of better programs and policies, Henry argues, and especially the debunking of bad ideas are key needs within representative democracies. Henry states,

> Public policies and programs are intentional acts undertaken with the coercive force of government behind them. Authoritative bodies, such as legislatures and administrators, make public policy and program choices in most cases. The choices of these bodies are binding on the governed and therefore, to inform citizens' beliefs about public programs' positive, negative, or null results and potentially to influence voting decisions, democracies require that both the intended and unintended consequences of programs and policies be evaluated. (p. 70)

In short, Henry makes a case that findings from RCTs are highly relevant to major decisions in democracies.

At the same time, Henry does not see the RCT and the question it addresses as the only option for evaluation:

> It is possible for program administrators who also control evaluation funds to focus evaluations on questions of program coverage, studies of variations in implementation, single-case studies of a program or organization, descriptive studies of program participants, or subjective assessments of satisfaction. At a given time and in a specific situation, any of these could be the best choice for evaluation funding. (pp. 77–78)

Henry does, however, see a clear place for the RCT and its cousins: "But continually and indefinitely postponing addressing the public program effectiveness question cannot be in the interest of society."

In the earlier years of the debate about RCTs, it was easy to find statements from those opposed to the priority that were based (sometimes implicitly) on the claim that findings from RCTs usually were not relevant to the key questions that should be addressed. In the previous edition of this book, for example, Rallis (2009) expressed the view that evaluation should generally address questions other than the average effectiveness of programs or policies. In lauding evaluators who follow what she called *nonconsequentialist ethics*, Rallis stated that "[t]hese evaluators attend to the means and context more than to the outcome of a program. They ask, What does the experience mean to the individual?" (p. 218). Rallis further said, "The nonconsequentialist theories with which I am comfortable shape research ethics that turn away from experimental type of studies that seek to know outcomes and turn toward the qualitative approaches that inform process and meanings" (p. 218). For Rallis at that time, the findings from RCTs seemed generally less relevant than those from other methods. The same appeared to be true of Greene (2009):

> [Q]uestions about the causal effects of social interventions are characteristically those of policy and decision makers, while other stakeholders have other legitimate and important questions. . . . This

> privileging of the interests of the elite in evaluation and research is radically undemocratic. (p. 157)

Rallis's and Greene's positions on the relevance of findings from RCTs appear to be tempered in this edition relative to the previous one. At least, their positions as stated in this edition seem more open to the kind of evidence that comes from RCTs, under certain conditions. Rallis notes that the evaluations she does typically focus on program implementation, in which case findings from an RCT would not be relevant. However, she also states that if the question of interest is whether the program caused a change in outcomes, then a method such as an experiment would probably be chosen. Greene gives us a completely different chapter in this edition than in the previous one. In the current chapter, she describes practices that, in her view, increase credibility, which she illustrates in the context of a randomized experiment.

The two sides seem less polarized than in the past, at least with respect to the possible relevance of evidence from RCTs. Whether there is agreement about the conditions under which to use RCTs or other methods, however, is not clear. To the extent these disagreements continue, we may understand them better by thinking about them in the context of the criteria in Figure 13.1 and in a broader framework related to the use of evidence.

Generalizability as an Aspect of Relevance

Cartwright and Hardie (2012) emphasize that RCTs (and other forms of impact or outcome evaluation) answer a limited question: whether the program or policy "worked there." That is, in the ideal, the RCT gives an unbiased estimate of the effect of the specific way the program was implemented in the specific setting and time in which the RCT took place with the particular study participants. Cartwright and Hardie further emphasize that policy deliberations need to ask a different question—*Will it work here?*—that is, in the circumstances in which the policy or program may be implemented. Readers familiar with the earlier evaluation literature will recognize the *It worked there* versus *Will it work here?* formulation as an echo of Cronbach's (1982) critique of the limited reach of internal validity.

In their relatively detailed and thoughtful discussion of the limits of RCTs, Bickman and Reich point out that the "most often-cited criticism of this design is its reduced external validity." This refers to the possibility that you may not be able to generalize from the perhaps artificial circumstances that allow random assignment to other circumstances about which decisions are to be made. Bickman and Reich cite Berk (2005):

> It cannot be overemphasized that unless an experiment can be generalized at least a bit, time and resources have been wasted. One does not really care about the results of a study unless its conclusions can be used to guide future decisions. (p. 428)

Bickman and Reich applaud signs of possible increased attention to external validity in evaluation and applied social research. However, it is not yet clear how (and when) to give comfort to potential evidence users that the results of an RCT should indeed apply to future decisions.

Greene (Chapter 10) emphasizes the importance of relevance and suggests that relevance is required for

credibility:

> The evaluation is directed toward issues, concerns, and questions of *contextual meaning and relevance;* the evaluation is designed to generate information of import and consequence to stakeholders *in the contexts* being evaluated. Evidence that is relevant only for remote stakeholders does not have the potential for credibility. (p. 208)

In contrast, the model in Figure 13.1 allows for the possibility that evidence could be credible but not relevant. Regardless of whether relevance is viewed as a separate characteristic or as an aspect (or precursor) of credibility, Bickman and Reich, Greene, and others in this volume see relevance as an important concern.

In her chapter in the previous edition of this book, Greene pointed to the power of contexts, including the possibility that relevant processes may be different at one point in history than another. Citing Berliner (2002), she noted that "it is because of the need to understand the particularities of each local context that 'qualitative inquiry has become so important in educational research'" (p. 160). One response to this point would be to emphasize the importance of mixed methods, melding qualitative and quantitative to seek both unbiased estimates of treatment effects and sensitivity to local context. Absent that, advocates of RCTs would be well advised to engage in work to improve our understanding of the conditions under which generalizations from such studies are successful. And those who wish to debate the value of alternative methods would be well advised to try to assess the *relative* capacity of different methods to provide findings that support conclusions across different contexts, or put differently, different settings, persons, and times. For instance, does a case study with, say, eight sites provide more support than an RCT for generalizing to new locations in the future?

Henry (Chapter 4) reminds us that there often are limits on how microscopic this attention to generalizability—and thus to relevance—needs to be. He notes that governments commonly take actions that apply across a wide scope of persons, settings, and times. A math curriculum is selected for an entire school district, not for a given child in a specific classroom with a particular teacher. Classroom size laws may be implemented at the state level, again cutting across a wide range of contexts. The challenges of generalizability presumably are greater if choices among alternative programs are to be made at the individual level than at the state level. That is not to say the challenges are absent for, say, state-level choices but only that they are lesser in magnitude.

Method Choices in Light of Credibility

Preferences for methods are often framed in terms of credibility and its companion, inferential potency. For instance, Henry addresses credibility in the context of the view that a democracy needs to evaluate the consequences of the programs and policies that are binding for its citizens. Henry contends that to obtain findings about program consequences that are "as conclusive as possible," the "most conclusive and widely regarded as compelling means for producing findings that eliminate these alternative explanations are RCTs and FEs." In contrast, others, including Chelimsky, Rallis, Mathison, and Scriven, explicitly endorse the credibility of methods other than RCTs. These disparate views appear, at least in part, to be based on differing

views about relevance. Even though these two criteria are listed separately in Figure 13.1, they may be linked in people's judgments.

Credibility assessments can be influenced by numerous factors, as discussed especially by Miller (Chapter 3) and as addressed later in this chapter in the context of a broader framework that includes Figure 13.1. The advocacy of RCTs that surrounded the IES priority, the creation of 3ie, and related initiatives can be seen in this light. These activities could be viewed as part of a social movement intended in part to enhance the credibility of RCTs as seen by government officials, foundation staff, and others.

Method Choices in Light of Inferential Potency

Assume for the moment that an estimate of average program effects is relevant for an upcoming decision. Disagreements about the desirability of conducting an RCT can also rest on differing views about inferential potency. Note that, when considering a priority for one method over another, the real question is about *relative* inferential potency. Although not using such terminology, Henry contends that RCTs have more inferential potency than alternative methods. He bases this on the idea that random assignment experiments are superior in terms of eliminating alternative explanations.

Scriven (Chapter 6), in contrast, suggests that other methods are equally and sometimes better suited for estimating program effects. As a result, Scriven argues forcefully against giving a general priority to RCTs. Scriven's argument is considered in detail in Appendix I. In short, Scriven's position regarding RCTs is that "the usual claimed intrinsic advantage of the RCT design is only a theoretical advantage, and its actual achievements are frequently matched in practice by many other designs."

Bickman and Reich add another voice to this discussion. They consider a list of potential problems that can undermine the inferential potency of RCTs. To have a fair discussion about relative inferential potency, one would need to do that same kind of analysis for whatever method alternatives are under consideration. As Scriven's analysis reminds us, an analysis of the *general* inferential potency of any method in the abstract may not apply to a *specific* study. As an example, several of the problems that Bickman and Reich describe for RCTs may occur in one study, while another RCT may be free from these problems. Unfortunately, evaluators/researchers and funders often need to try to assess, *in advance,* whether a study with a particular design will be subject to flaws. To the extent this can be surmised in advance, review panels could make defensible judgments about the relative value of alternative research designs. However, if there is little ability to foresee the presence or potency of validity threats, then one might prefer to bet on the *general* inferential potency of a design.

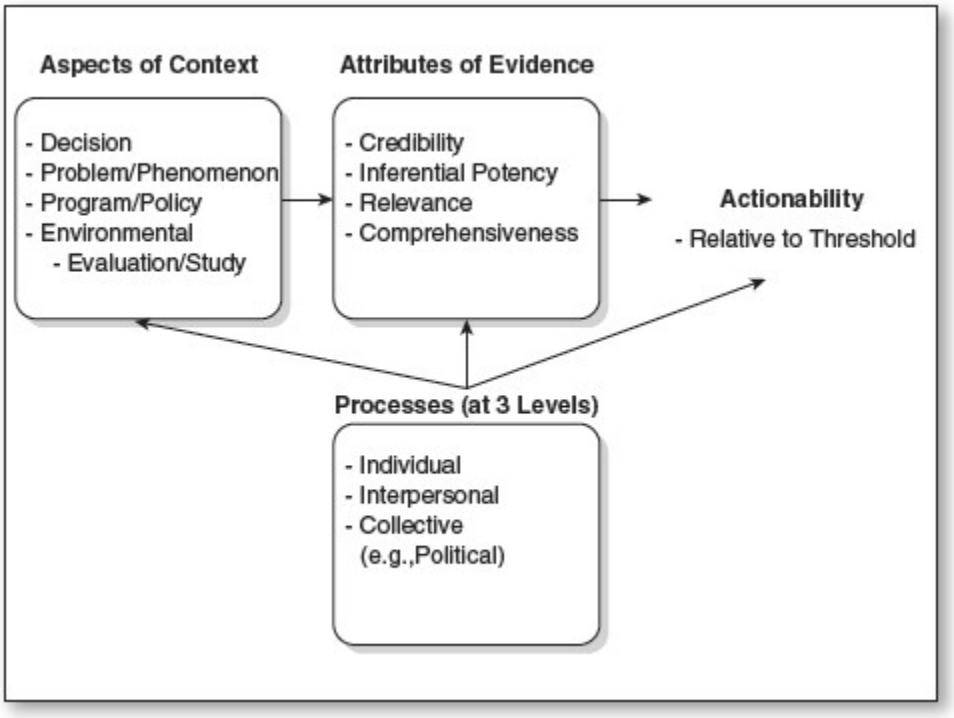
Credibility and Actionability in a General Framework of Evidence Use, Part 2

Figure 13.1 provides a simplified model in which four characteristics of evidence—credibility, inferential potency, relevance, and comprehensiveness (both specific and general)—combine to result in judgments about actionability. As noted previously, one way that Figure 13.1 is simplified is that it ignores the possible interplay between the four characteristics on the left. For instance, perceived relevance may affect judgments of credibility, as suggested by Greene's quote about the two.

A second way in which Figure 13.1 is simplified is that it represents a limited slice of a larger, prior sequence. In particular, judgments about credibility, inferential potency, relevance, and comprehensiveness can be affected by a variety of factors. Accordingly, Figure 13.2 places the evidence attributes of Figure 13.1 within a broader framework with two additional sets of factors. On the left of Figure 13.2 are four of the aspects of context from Julnes and Rog. The fifth, *evaluation* or *study context*, sits slightly to the right of the other aspects of context, because the other four generally are present prior to the study. For example, information needs and characteristics of the problem exist before the evaluation begins in most cases. Indeed, one hopes that thoughtful analysis of the other contextual areas took place as part of the planning of the evaluation.

At the bottom of Figure 13.2, in bold, are three levels of processes that may influence judgments of credibility and the other attributes of evidence. These are individual, interpersonal, and collective levels. Along with the areas of context, the three levels of process are discussed in the current section.* The arrows in Figure 13.2 from the processes are intended to show that processes at the three levels can affect, for example, the relationships between context and the perceived attributes of evidence. Figure 13.2 would be technically more accurate if there were arrows emanating from each of the areas of context and from each of the levels of processes. It would, however, also be less legible.

Figure 13.2 A Broader Framework With Five Aspects of Context, Three Levels of Process, Four Attributes of Evidence, and Judged Actionability



Five Aspects of Context

As noted during the discussion of Figure 13.1, judgments about credibility, relevance, and the other characteristics of evidence generally do not take place in a vacuum. Rather, they take place in the context of, for example, a specific evaluation conducted of a particular program, which was implemented on a pilot basis to address a given social problem, in the face of the information needs of certain stakeholders, with all of this taking place within a broader social context.

Julnes and Rog (Chapter 11) employ a framework that specifies five areas of context: (a) decision context, which involves stakeholders' information needs and the alternative evaluation questions and tasks that might be addressed, such as formative versus summative evaluation purposes; (b) problem/phenomenon context, which involves key attributes of the phenomenon being investigated, such as whether change is likely to be easy and quick or difficult and slow; (c) program/policy intervention context, which includes characteristics of the program or policy, such as the level of program maturity; (d) evaluation context, which concerns specific circumstances of the evaluation or study, including funding and evaluation capacity; and (e) environmental context, that is, more general characteristics of the context in which the program or policy operates, which includes political constraints and ethical considerations. Four of these are represented on the left side of Figure 13.2, with the evaluation or study context placed slightly to the right of the others (which typically precede the study).

The Julnes and Rog model of alternative areas of context can be used to organize many points raised throughout this book and elsewhere. As one example, drawing on her experience leading a major unit of what is now called the Government Accountability Office (GAO), Chelimsky (Chapter 9) identifies different kinds of evaluation questions. According to Chelimsky, the questions that come to evaluators fall into four categories: management questions, knowledge questions, accountability questions, and development questions. This distinction could be incorporated within Julnes and Rog's decision context.

More important than its ability to incorporate others' conceptualization, Julnes and Rog's model of context can be useful in thinking about method choices. Julnes and Rog sensibly caution against the idea that any framework can effectively generate method choices automatically. But they also suggest that attention to relevant contextual factors can be useful when a preferred method is being selected. For example, in their Table 11.2 and thereabout, Julnes and Rog discuss a set of contextual factors that they see as making RCTs particularly valuable. These include information needs that center on the merit and worth of a program (such that an estimate of average program effects would be informative); the need for evidence that provides a high level of confidence about the program effects; a phenomenon that does not allow a confident estimation of effects with simpler methods; a program that is sufficiently mature; and adequate resources for an RCT, along with the absence of major ethical concerns.

It would be possible to engage in a similar exercise to try to identify the contextual factors that would make other methods particularly valuable. This includes methods that were discussed in this volume and ones that were not. Some readers might wish to take this on as an exercise (this might be a good exercise for graduate students, for example).

Returning to Figure 13.2, the specifics of context are likely to affect perceptions of credibility and the other attributes of evidence. For instance, the decision context should strongly affect judgments of relevance. Suppose the decision context is such that information is needed about a program's overall merit and worth. RCTs and related methods would then be more likely to be seen as relevant, while a study that provides descriptive information about implementation would be likely be viewed as less relevant.

Contextual factors could also affect the relative importance (or weighting) of the various criteria as precursors of actionability. Imagine, for example, a decision environment in which a high level of confidence is needed in the estimate of program effects. Inferential potency should be relatively more important in this context as a determinant of judged actionability. (For the sake of simplicity, Figure 13.2 does not have arrows showing all the possible effects of context on each of the perceived criteria of evidence.)

Three Levels of Processes

Figure 13.2 expands on Figure 13.1 in another way—by including three levels of processes that may be involved in generating judgments about credibility and the other attributes from Figure 13.1. Mark and Henry (2004) used the same levels in discussing processes that lead to the use or influence of evaluation findings. In short, these three are the individual, interpersonal, and collective levels (the collective level could involve

political process or other organizational processes, depending on the circumstances). Different distinctions could be made, either coarser or finer, but the three categories in Figure 13.2 appear to be useful without being too cumbersome. Contributors to this volume did not explicitly highlight the existence of these three levels. However, several authors focused on different levels, as noted in this section.

Individual level processes are, for the most part, psychological in nature. Miller's (Chapter 3) primary focus is on the individual psychological processes that can affect judgments of credibility. Although her emphasis is on credibility, the processes Miller reviews have implications for the other characteristics of evidence and potentially for the perception of the Julnes and Rog's aspects of context.

As Miller explains, the extent to which a potential user relies on the actual content of an evaluation depends on the user's ability and motivation to do so. Otherwise, various peripheral cues or heuristics may be employed. For example, the format and appearance of a report (or briefing) may matter. The prestige of the evaluator's affiliation, consistency of findings with deeply held beliefs, and various other factors (many reviewed by Miller) can affect perceived credibility and other perceived attributes of evidence at the individual level. Chelimsky makes a related point, drawing on her experience rather than the psychological literature on credibility. She shares the lesson that evaluators should attend to "the evidence of credibility" as well as the credibility of evidence. In this way, Chelimsky offers practice guidance that is conceptually related to Miller's research-based review. Chelimsky also calls for clear communication based on the actual strength of the evidence. Without careful attention to this, peripheral cues could lead to weak evidence being viewed as strong or vice versa. At the individual level, personally held values can also affect perceptions, for example, by making findings seem more credible if they agree with the person's values.

The *interpersonal level* involves processes by which one or more individuals influence another. For example, legislators' views of the credibility of evidence are often based on input from their staff or others. More generally, people with expertise and interest in a program or policy area often are among the first to become familiar with relevant evaluations or other studies; they may then engage in efforts to persuade others. Such intermediaries may undertake explicit efforts to shape others' views of the credibility and relevance of the evidence. Or their efforts may be more subtle.

Greene (Chapter 10) emphasizes the importance of interpersonal processes as determinants of credibility. Mathison makes a similar point, but in less detail. Like Mathison, Greene does not focus on congressional staff, experts, or other intermediaries. Instead, she focuses on the interactions between the evaluator and local stakeholders. Greene contends that, in addition to strong and appropriate methods, credibility depends on the evaluation questions (a) having "*contextual meaning and relevance,*" that is, being of interest "to stakeholders *in the contexts* being evaluated" and, further, (b) being of interest "to a *diversity of stakeholders* in the contexts being evaluated, with special inclusion of the interests of stakeholders who are least well served in those contexts." To achieve this, Greene focuses on the cultivation of respectful relationships with local stakeholders and quality communication. She also emphasizes the inclusion of local stakeholders and their values in generating judgments of a program's value. For Greene, these aspects of interpersonal relations are key to achieving credibility.

The *collective level* involves processes that take place in identifiable and relatively stable social aggregates, such as private organizations and governments. I focus here on the political processes that operate in democratic governments. To be clear, the focus here is not on collective processes in the *use* of findings, such as the passage of a law that expands a pilot program to a national one. Rather, the focus here is on how collective (in this case, political) processes can affect relevant aspects of context and therefore influence the perceived characteristics of evidence, such as credibility. In particular, political deliberations frequently set the decision context, which in turn affects the judged relevance of different kinds of information. Political considerations also may affect the height of the hurdle required for a decision to act. For instance, if a program is well entrenched (think Social Security during much of its life), the threshold required to make a change is higher. Political processes do not fully determine judged relevance or other perceived attributes of evidence, however. Consider as an example the IES priority, which was established by appointed officials using standard procedures in response to legislation. The priority may have affected some people's view of what constituted relevant and credible evidence. But clearly, not all observers agreed.

Several contributors to this volume refer to political processes and their role in the evaluation of evidence. For example, Henry (Chapter 4) discusses political processes in a general sense. Chelimsky (Chapter 9) draws on her impressive experiences to describe the interactions between the GAO and Congress in several evaluation efforts involving synthesis of the available evidence.

The contrast between these two evaluators' typical work settings may point to an aspect of context that is not adequately appreciated relative to its importance. Sometimes evaluators work in circumstances that allow them to interact directly with key potential users of their work. Chelimsky at the GAO could engage in negotiations with members of congress and their staff about the scope of a synthesis or other evaluation. Greene presumes that the evaluator can interact with local stakeholders. In contrast, evaluators sometimes are several steps removed from the ultimate key potential users of their work. For instance, a national evaluation that could inform legislative action might be mandated by Congress but overseen by an agency in the executive branch. The number of steps between evaluation and evaluation users may alter the effectiveness of the use of various practices that an evaluator might undertake to try to increase credibility and actionability.

Selected Additional Issues Related to the Framework

The framework presented in Figure 13.2 can serve as background to thinking about several other issues raised, explicitly or implicitly, across the preceding chapters.

Divergent Views about Democracy and Evaluation

Greene (Chapter 10, 2009) and Henry (Chapter 4) both appeal to democratic values, but in support of rather different views of evaluation. Henry refers to contemporary theory about representative democracy. More specifically, Henry highlights the contribution that can occur when convincing information about program

effects is available to democratically elected and appointed officials (and to voters who may be disappointed in their representatives' choices).

Greene (2009), in contrast, emphasized that stakeholders other than policy makers "have other legitimate and important questions" and that the "privileging of the interests of the elite in evaluation and research is radically undemocratic." In Chapter 10, Greene states, "Evidence that is relevant only for remote stakeholders does not have the potential for credibility." Thus Greene emphasizes direct, participatory democracy. Related to this, she highlights interpersonal processes by which the evaluator includes, communicates with, and respects local stakeholders, especially the least well-off. Because Henry emphasizes representative democracy, he instead seems willing to address the information needs of elite and remote stakeholders such as legislators, who may be interested in assessing the effects of a program they have funded.

Perhaps the difference between Henry and Greene is smaller than it appears. That is, the questions that interest legislators may be of interest not *only* to these remote decision makers but also to local stakeholders. Even so, Henry and Greene would seem to diverge at least in terms of assumptions and preferred procedures. Henry appears to assume, in general, the legitimacy of addressing the question about average program effects when policy makers request such information. In contrast, Greene seems not to assume the legitimacy of this question, unless respectful interactions with local stakeholders reveal their interest in it.

Perhaps an alternative approach is to recognize that in the U.S. and most modern democracies, a mixed model of democracy operates. The reality is a mix of representative democracy, direct participatory democracy, and probably other abstract models. From this vantage, evaluators can serve democracy by addressing either the information needs of elected and appointed officials or the information needs of local stakeholders. Indeed, it can be argued that both democracy and the field of evaluation benefit from having evaluators who emphasize and serve different audiences. Further, for both approaches, a persuasive argument can be made for the evaluator to try to ensure that multiple value positions are considered and the needs of the least well-off are represented. But, from this perspective, there is not a single pathway by which evaluation can serve democratic ends.

Whose Views?

The issue raised by Henry and Greene of representative versus participatory democracy represents a more general concern: Whose views about credibility, actionability, and the like matter? The pragmatist is likely to say that the views that matter most are those of whoever is supposed to be influenced by the evaluation or applied social research and perhaps whoever else will influence them. So, for example, if the potential use of evaluation findings is by legislators voting on reauthorization of welfare, then for the pragmatist, the perceptions that matter most are those of the legislators as well as others who may affect their judgment, such as key legislative staff. Greene's (Chapter 10) comments indicate that not all researchers and evaluators would agree with this idea, however. Greene appears to give priority to the perspectives of various local

stakeholders directly involved with the program, especially the disadvantaged.

The question of whose views matter reminds us of an additional consideration that overlays Figure 13.2: Whose judgments are we considering when we assess the decision context and other contextual considerations? Whose perceptions of credibility, relevance, and so on matter? Left unconstrained, these views may well vary, as shown for example by the contrast of Henry's and Bickman and Reich's views with those of either Scriven or Rallis and Mathison.

With What Supports?

Drawing on the psychological literature on credibility, Miller (Chapter 3) reminds us that people often rely on peripheral cues, such as a researcher affiliation or an evaluation report's appearance. For people instead to do a thoughtful assessment of the quality of a report and its findings, they need to have both the ability and the motivation to do so. Many potential users of applied research and evaluation do not have the training required to pass technical judgment of the quality of the work. Many also have too much else to do in their jobs and lives. This shortage of ability and/or motivation is among the reasons why legislators often rely on staff for guidance as to how heavily to weigh evidence.

A variety of professional arrangements can be seen as offering support for the sensemaking efforts of decision makers and other stakeholders who may have to decide whether and how to weight evaluative or other evidence in their decisions. These include but are not limited to peer review panels that select studies for funding, expert advisory panels that help guide the project, meta-evaluators who offer judgments on the quality of the work, and the comments from scholarly and professional experts who praise or criticize the study.

In addition, efforts are sometimes undertaken to increase capacity to make sense of evidence. Professional associations may hold "boot camps" for reporters and other intermediaries. Professional evaluators have held briefings about using evidence for staffers and government officials. In Chapter 9, Chelimsky's advice about "evidence of credibility" offers a concise roadmap for evaluators to create a guide to evidence quality within an evaluation briefing or report.

The creation and dissemination of *evidence hierarchies*, which rate methods in terms of quality, can be seen as another attempt at capacity building. That is, in a sense, an evidence hierarchy is an effort to assist stakeholders who lack training in research methods but who need to decide whether one or more studies provide actionable guidance. The problem is that in practice, evidence hierarchies tend to be too broad and too general. Too often, they seem to suggest that a preferred method, such as the RCT, is *generally* preferable; it would be better to suggest that the method tends to be valuable for certain purposes in specific circumstances. Thus evidence hierarchies tend not to give adequate attention to the kind of analysis of a specific context that Julnes and Rog call for. Many evidence hierarchies also tend to give short shrift to the strengths and weaknesses of a study *as conducted.* Thus they tend not to give adequate attention to the kind of detailed inventory of validity threats and related weaknesses presented by Bickman and Reich. An

important ongoing project for the future involves the construction of better forms of support for those who should consider research evidence in their decisions but do not have advanced training in the methods used.

Given What Assumptions?

As noted previously, people with different method predilections probably start with different default assumptions. They may, for example, assume that different decision contexts generally exist. Consider the earlier discussion of Henry and Rallis, who make different default assumptions about relevance. Or people may have opposing assumptions about the extent to which the problem and program context enable causal inference with methods other than RCTs; this would create differing views about credibility and inferential potency. A case can be made that these different default assumptions often arise, at least in part, from differences in training and in adherence to the kind of traditions and paradigms that Christie and Fleischer (Chapter 2) describe.

A case can also be made that if one could create greater agreement about the nature of the context in which evidence is to be generated, then greater agreement would probably result about method choice, credibility, and so on. This notion was a major thrust of the Julnes and Rog chapter. In the absence of detailed information about context, different evaluators may have quite varied views about what kind of evidence is typically relevant. On the other hand, if a particular situation is detailed, with the various aspects of Julnes and Rog's five aspects of context spelled out, evaluators may agree more. (Put differently, we can ask evaluators to make *conditional judgments*, assessing the credibility, actionability, etc. of evidence conditional on certain detailed contexts.) If this is true, future discussion could focus more on the factors that underlie opinions about credible and actionable evidence rather than on the more superficial matter of method choice *per se*.

Future Research

The framework in Figure 13.2 suggests a number of research questions that could be fruitfully addressed. Among the possible questions are the following: (1) One could investigate the interrelations among (and indeed, the construct validity of) credibility, inferential potency, relevance, and comprehensiveness. (2) The extent to which these four attributes affect judgments of actionability could also be investigated. (3) Another group of research questions involves the effects of the various aspects of context on judgments about credibility and the other judged attributes of evidence. (4) The effect of various individual, interpersonal, and collective processes on the other factors in the framework is another arena for research. A range of research methods could be employed, from reflective case studies of past evaluations to surveys measuring evaluators' and stakeholders' perceptions to simulations with random assignment to levels of a factor of interest.

Choices for IndividualStudies and for Portfolios of Studies

Scriven and Henry, in different ways, raise a very important consideration for a more general theory of

credibility and value of evaluation and applied research. Scriven thoughtfully and explicitly raises the question of funding priorities not for individual studies but rather for broader portfolios of studies. He suggests that even if RCTs are preferable in general, one would likely be better off not putting all the investment eggs in the basket of RCTs: “the best single investment—better than every alternative—is not the best bet for the whole portfolio budget.” This is a relatively compelling point, especially in light of the kind of external validity and other validity limits noted by Bickman and Reich. For example, RCTs might give the most internally valid, unbiased estimates of the effects of a program; nevertheless, because of the limited circumstances in which random assignment is feasible, RCTs might result in serious reservations about external validity. If true, funding a second RCT with the same external validity limits might purchase less of an increment in confidence relative to funding a strong quasi-experiment. Although the latter would allow less study-specific confidence about the program’s effects, it could better facilitate generalization to policy-relevant settings. Recognition that the considerations that guide method choices for one study might be quite different than the considerations that guide method choices for a second, third, or subsequent study is quite consistent with the literature on critical multiplism (e.g., Cook, 1985), but this point has not been adequately recognized in contemporary debates about method preferences and credibility, such as the debates about RCTs in education, development evaluation, and elsewhere.

Henry reminds us that any single funding stream is part of a broader portfolio. He acknowledges the appropriateness of “program administrators who also control evaluation funds to focus evaluations on questions of program coverage, studies of variations in implementation, single-case studies of a program or organization, descriptive studies of program participants, or subjective assessments of satisfaction.” But he also supports priorities at select agencies:

> This situation is likely to be best addressed by establishing and funding alternative institutions to provide priorities for evaluations of program effectiveness. It appears that independent or quasi-independent federal organizations such as IES, the National Institute of Justice, and the National Institutes of Health have rebalanced institutional priorities by providing resources for independent evaluations. While evaluations sponsored and conducted by the agencies administering the programs are likely to be biased toward questions that do not include an assessment of the program’s consequences, these quasi-independent institutes can help to ensure that the resources needed to implement designs at the top of the hierarchy to evaluate the consequences of public policies and programs are available. (p. 78)

Expanding on this point, we can raise a question about what the conceptual boundary is for the portfolio of studies. Is it, for example, the particular funding program that uses the IES priority for random assignment? Or is it the broader array of Department of Education funding programs, some of which rarely, if ever, fund RCTs? Or is it the still broader set of educational research and evaluation with numerous funding sources? Advocates of the priority for RCTs may, like Henry, take the broader view and see the priority as a rebalancing. Critics of the priority may take the narrower view and see the priority as an unjustified unbalancing.

Selected Suggestions for Practice

This chapter provides a framework combining (a) Julnes and Rog's five aspects of context with (b) four attributes of evidence (credibility, inferential potency, relevance, and comprehensiveness) believed to be precursors of judged actionability and (c) three levels of processes that can affect the judgments about credibility and the other factors in the model. Much of the chapter has focused on how the elements of this framework can help clarify why different views exist about method preferences. My hope is that if future debate can be focused on the underlying reasons, the resulting discussion may be more fruitful. Ideally, future discussion could lead to more contingent and consensual guidance statements, indicating which methods are preferable under which circumstances. The framework may also be useful in organizing and supporting suggestions for the planning of applied research and evaluation. The latter possibility is the focus of the current section.

When planning an upcoming evaluation or other applied study, the investigator should consider the areas of context described by Julnes and Rog, probably starting with the decision context. For example, are the issues of interest ones about the overall merit and worth of an intervention, and would these be addressed well by estimating the effects of an identifiable treatment? An example would be "Stakeholders want to know whether the new pre-algebra math curriculum actually leads to improved student performance and interest in math." *If* the upcoming research should estimate the effect of a specific intervention (e.g., the new math curriculum) on already identified outcomes (e.g., math performance and interest), then experiments, quasi-experiments, and other related cause-probing techniques would provide relevant evidence. In contrast, if the key question instead is "What variants in program implementation exist?" then experiments and their approximations would not be informative. As these examples suggest, the decision context should inform study design and should also inform subsequent judgments of the relevance of the resulting evidence.

Second, consider the other aspects of context. Assume for the moment that the decision context is such that the question of program effects is of interest. As Julnes and Rog discuss, the relative value of a method should be made in light of relevant aspects of context. Is the nature of the phenomenon and program such that effects should be large and quick and alternative explanations of change unlikely? If so, simpler methods, such as a simple pretest–posttest design should suffice. In contrast, if program effects are likely to be modest in size relative to the changes that are plausible from other factors (or validity threats), then an RCT has greater add-on value. Aspects of context will influence perceived credibility and the other attributes of evidence.

Third, be sure to think about comprehensiveness. If the question of program effects is of interest, for example, which outcomes need to be measured in order to meet stakeholder information needs (specific comprehensiveness)? For instance, can action be taken if relevant stakeholders learn of an effect of a job training program on employment? Or do they need to know about a broader range of outcomes, such as effects on income, family composition, health, and children's well-being? Also, which (if any) other questions, such as ease of quality implementation across sites, need to be addressed (general comprehensiveness)?

Even if the study design and budget have been set by others, the evaluator may be able to increase comprehensiveness, especially specific comprehensiveness, by adding other measures.

Fourth, think about generalizability (or external validity) as an aspect of relevance. The decisions that an evaluation are meant to inform may involve places and persons different than the ones examined in the study. At the least, the decisions will be about a future time. Thus practitioners should think about how to increase the presumed application to the settings that would be affected by the decision. For instance, if an RCT is planned, theory-based tests of moderation and mediation may be valuable, as would the sampling of different contexts and subgroups.

A fifth point of action—or an alternative starting point—involves thinking about whose views matter most in this case. For example, whose questions are given priority when assessing the decision context? Of course, the researcher sometimes comes to the scene after the funding agency or some other body has defined the scope of the study, effectively taking this kind of planning exercise out of the hands of the researcher. In other cases, the researcher or evaluator has considerable latitude to set, or at least influence, the key research questions and methods. Even when aspects of the study design are defined before the researcher is on the scene, there may be opportunities to modify or at least expand on the original plan. In those cases in which the researcher has more leeway, the pragmatic approach is to consider most strongly the views of the expected users of an evaluation, whether those are legislators, a local service delivery organization or school, a foundation, or others. As we have seen, some applied researchers and evaluators emphasize representative democracy (and, typically, providing information for elected and appointed officials), while others prefer to give voice to the least powerful. The evaluator or applied researcher should at least consider these alternatives. Depending on the answer, this issue may have important implications, including whether to engage deeply in the kind of stakeholder interactions described by Greene.

Sixth, consider the potential supports that will be provided to assist in relevant parties' assessments of credibility and inferential potency. If stakeholders lack the expertise or motivation to make technical judgments of credibility and inferential potency, they may not have strong initial preferences about design. But the practitioner should consider what judgments the stakeholders are likely to have after the researcher has undertaken an educative function with the stakeholders or what judgments would be made by intermediaries such as legislative staff. Chelimsky's points about the evidence of credibility apply.

Seventh, keep in mind that the goal is to facilitate the appropriate use of evidence, not to get all evidence used. Evaluators should not try to sell weak evidence as credible and actionable. Chelimsky reminds us of this in her discussion of telling policy makers what is *not* known. Related to this, the professional standards to which most researchers adhere commonly include statements about technical quality or accuracy—bad research is not justified, even if the relevant stakeholders would find it credible.

At the same time, the height of the threshold for a judgment of action should also be considered. Are high levels of credibility, relevance, inferential potency, and comprehensiveness needed to justify action? Or could action reasonably be guided by the evidence, even if credibility or inferential potency is somewhat lower?

Even if the threshold for action is lower, the researcher ought to obtain the strongest evidence possible in light of the available resources. And if credibility, inferential potency, relevance, or comprehensiveness are low, this should be honestly communicated.

Thoughtfully consider trade-offs. As noted previously, trade-offs may exist among credibility, relevance, inferential potency, and comprehensiveness as well as between these and resource constraints. The thoughtful researcher keeps in mind the costs of various trade-offs and attempts to make the choices that will best support judgments of actionability in light of the specific constraints faced.

Think about the next study in context of a portfolio of studies, if more than one study has been or will be done. Much of the past debate about RCTs implies that one size fits all and, therefore, that the fifth study should aspire to the same gold standard as the first four. This is undesirable. For example, the fifth highly credible study about a program's average treatment effect may provide less value relative to a first study that examines another important question. As an another example, a fifth study that is less credible in terms of internal validity could add great value if it extends external validity by examining clients and settings not included in the previous four studies.

Take appropriate advantage especially of the individual and interpersonal processes that can affect judgments of credibility and other attributes of evidence. Examples at the individual level again include Chelimsky's recommendations for providing evidence of credibility. At the interpersonal level, Greene's recommendations about respectful communication and involvement seem valuable, regardless of one's views of democracy.

Conclusion

Frameworks such as those summarized in Figure 13.2 can seem like an academic exercise, of value only to ivory-tower types who care about nothing but abstractions such as *theory*. In the ideal, however, such frameworks will have value both for scholarship and for *practice*. I hope that readers will be able to find both kinds of values in the framework offered here. I would hope that the framework might contribute to more fruitful debate and discussion about method choices, including highlighting common ground where such can be found; guide future research to increase understanding; and support thoughtful practice in generating credible and actionable evidence in service of evidence-informed decision making that may lead to better outcomes for the intended beneficiaries of the programs, policies, and practices that we study.

References

Berk, R. A. (2005). Randomized experiments as the bronze standard. *Journal of Experimental Criminology*, 1, 417–433.

Berliner, D. C. (2002). Educational research: The hardest science of them all. *Educational Researcher*, 31,

18–18.

Cartwright, N., & **Hardie, J.** (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford, England: Oxford University Press.

Cook, T. D. (1985). *Post-positivist critical multiplism*. In **R. L. Shotland** & **M. M. Mark** (Eds.), *Social science and social policy* (pp. 21–62). Beverly Hills, CA: SAGE.

Cook, T. D. (2003). Why have educational evaluators chosen not to do randomized experiments? *Annals of American Academy of Political and Social Science*, 589, 114–114.

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

Donaldson, S., **Christie, T. C.**, & **Mark, M. M.** (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: SAGE.

Gersten, R., & **Hitchcock, J.** (2009). *What is credible evidence in education? The role of the What Works Clearinghouse in informing the process*. In **S. Donaldson**, **T. C. Christie**, & **M. M. Mark** (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 78–95). Thousand Oaks, CA: SAGE.

Greene, J. C. (2009). *Evidence as "proof" and evidence as "inkling."* In **S. Donaldson**, **T. C. Christie**, & **M. M. Mark** (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 153–167). Thousand Oaks, CA: SAGE.

Henry, G. T., & **Mark, M. M.** (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation*, 24, 293–293.

Mark, M. M., & **Henry, G. T.** (2004). The mechanisms and outcomes of evaluation influence. *Evaluation*, 10, 35–35.

Rallis, S. F. (2009). *Reasoning with rigor and probity: Ethical premises for credible evidence*. In **S. Donaldson**, **T. C. Christie**, & **M. M. Mark** (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (pp. 168–180). Thousand Oaks, CA: SAGE.

Reichardt, C. S., & **Mark, M. M.** (1998). *Quasi-experimentation*. In **L. Bickman** & **D. Rog** (Eds.), *Handbook of applied social research* (pp.193–228). Newbury Park, CA: SAGE.

* A third way in which Figure 13.1 is simplified has not been addressed in Figure 13.2. Presumably, actionability judgments are not important in and of themselves. Instead, they—like credibility and other such judgments—are primarily important because they affect the likelihood of use. They presumably help determine whether evidence contributes to subsequent changes in action and understandings. Ultimately, the primary concern is even further downstream: Does the use of evidence help set in play a chain of events that

results in improved outcomes for the intended program and policy beneficiaries? In terms of Figure 13.2, this means there are important potential additions that could be made on the right-hand side. However, given the focus of this book, the already ambitious Figure 13.2 will not be further elaborated here. (For one approach to the kind of additions that might be made, see Henry & Mark, 2003; Mark & Henry, 2004.)

http://dx.doi.org/10.4135/9781483385839.n17