CHAPTER ONE

PLANNING AND DESIGNING USEFUL EVALUATIONS

Kathryn E. Newcomer, Harry P. Hatry, Joseph S. Wholey

The demand for systematic data on the performance of public and nonprofit programs continues to rise across the world. The supply of such data rarely matches the level of demand of the requestors. Diversity in the types of providers of pertinent data also continues to rise.

Increasingly, elected officials, foundations and other nonprofit funders, oversight agencies, and citizens want to know what value is provided to the public by the programs they fund. Members of program staff want to know how their programs are performing so that they can improve them and learn from the information they gather. Increasingly, executives want to lead *learning organizations*, where staff systematically collect data, learn what works and does not work in their programs, and use this information to improve their organizational capacity and services provided. Leaders and managers also want to make evidence-based policy and management decisions, informed by data evaluating past program performance.

As we use the term in this handbook, a *program* is *a set of resources and activities directed toward one or more common goals, typically under the direction of a single manager or management team.* A program may consist of a limited set of activities in one agency or a complex set of activities implemented at many sites by two or more levels of government and by a set of public, nonprofit, and even private providers.

Program evaluation is *the application of systematic methods to address questions about program operations and results. It may include ongoing monitoring of a program as well as one-shot studies of program processes or program impact. The approaches used are based on social science research methodologies and professional standards.* The field of program evaluation provides processes and tools that agencies of all kinds can apply to obtain valid, reliable, and credible data to address a variety of questions about the performance of public and nonprofit programs.

Program evaluation is presented here as a valuable learning strategy for enhancing knowledge about the underlying logic of programs and the program activities under way as well as about the results of programs. We use the term *practical program evaluation* because most of the procedures presented here are intended for application at reasonable cost and without extensive involvement of outside experts. We believe that resource constraints should not rule out evaluation. Ingenuity and leveraging of expertise can and should be used to produce useful, but not overly expensive, evaluation information. Knowledge of how trade-offs in methodological choices affect what we learn is critical.

A major theme throughout this handbook is that evaluation, to be useful and worth its cost, should not only assess program implementation and results but also identify ways to improve the program evaluated. Although accountability continues to be an important goal of program evaluation, the major goal should be to improve program performance, thereby giving the public and funders better value for money. When program evaluation is used only for external accountability purposes and does not help managers learn and improve their programs, the results are often not worth the cost of the evaluation.

The objective of this handbook is to strengthen program managers' and staff members' abilities to meet the increasing demand for evaluation information, in particular information to improve the program evaluated. This introductory chapter identifies fundamental elements that evaluators and organizations sponsoring evaluations should consider before undertaking any evaluation work, including how to match the evaluation approach to information needs, identify key contextual elements shaping the conduct and use of evaluation, produce methodological rigor needed to support credible findings, and design responsive and useful evaluations. A glossary of some key evaluation terms is provided at the end of this chapter.

Matching the Evaluation Approach to Information Needs

Selecting among evaluation options is a challenge to program personnel and evaluators interested in allocating resources efficiently and effectively. The value of program evaluation endeavors will be enhanced when clients for the information know what they are looking for. Clients, program managers, and evaluators all face many choices.

Since the turn of the twenty-first century, the demand for evidence to inform policymaking both inside the United States and internationally has grown, as has the sophistication of the public dialogue about what qualifies as strong evidence. Relatedly, the program evaluation profession has grown in terms of both numbers and professional guidance. There are many influential organizations that provide useful standards for evaluation practice and identify competencies needed in the conduct of evaluation work. Three key sources of guidance that organizations and evaluators should consult before entering into evaluation work include:

- *Joint Committee on Standards for Educational Evaluation* (2010). This organization
  has provided four key watch words for evaluators for many years: *utility*, *feasibility*, *propriety*, and *accuracy* (see the committee's website, www.jcsee.org/program-evaluation-standards, for more information on the standards).
- *American Evaluation Association* (2004). The AEA's *Guiding Principles for Evaluators* is a detailed list of guidelines that has been vetted regularly by evaluators to ensure its usefulness (see www.eval.org/p/cm/ld/fid=51)
- *Essential Competencies for Program Evaluators Self-Assessment* at www.cehd.umn.edu/OLPD/MESI/resources/ECPESelfAssessmentInstrument709.pdf

Select Programs to Evaluate

Resources for evaluation and monitoring are typically constrained. Prioritization among evaluation approaches should therefore reflect the most urgent information needs of decision makers. There may be many demands for information on program performance. Not all of these can likely be met at reasonable cost. What criteria can guide choices?

Five basic questions should be asked when any program is being considered for evaluation or monitoring:

- Can the results of the evaluation influence decisions about the program?
- Can the evaluation be done in time to be useful?
- Is the program significant enough to merit evaluation?
- Is program performance viewed as problematic?
- Where is the program in its development?

One watchword of the evaluation profession has been *utilization-focused evaluation* (see Patton, 2008). An evaluation that is *utilization-focused* is designed to answer specific questions raised by those in charge of a program so that the information provided by these answers can affect decisions about the program's future. This test is the first criterion for an evaluation. Programs for which decisions must be made about continuation, modification, or termination are good candidates for evaluation, at least in terms of this first criterion. Programs for which there is considerable political support are less likely candidates under this criterion.

Timing is important in evaluation. If an evaluation cannot be completed in time to affect decisions to be made about the program (the second criterion), evaluation will not be useful. Some questions about a program may be unanswerable at the time needed because the data are not currently available and cannot be collected in time.

Significance can be defined in many ways. Programs that consume a large amount of resources or are perceived to be marginal in performance are likely candidates for evaluation using this third test, assuming that evaluation results can be useful and evaluation can be done in a reasonable amount of time.

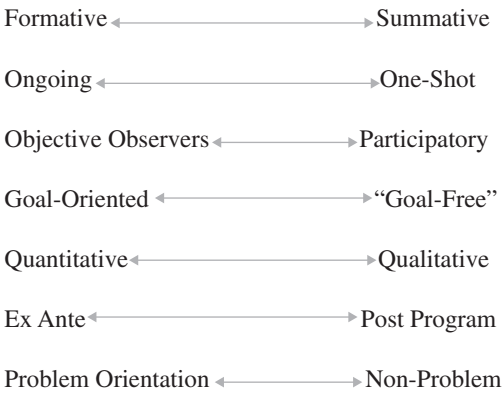
The fourth criterion, perceptions of problems by at least some program stakeholders, matters as well. When citizens or interest groups publicly make accusations about program performance or management, evaluation can play a pivotal role. Evaluation findings and performance data may be used to justify decisions to cut, maintain, or expand programs in order to respond to the complaints.

Placement of a program in its life cycle, the fifth criterion, makes a big difference in determining need for evaluation. New programs, and in particular pilot programs for which costs and benefits are unknown, are good candidates for evaluation.

Select the Type of Evaluation

Once a decision has been made to design an evaluation study or a monitoring system for a program, there are many choices to be made about the type of approach that will be most appropriate and useful. Figure 1.1 displays six important continua on which evaluation approaches differ.

FIGURE 1.1. SELECT AN EVALUATION APPROACH THAT IS APPROPRIATE GIVEN THE INTENDED USE.



Formative evaluation uses evaluation methods to improve the way a program is delivered. At the other end of this continuum is *summative* evaluation, which measures program outcomes and impacts during ongoing operations or after program completion. Most evaluation work will examine program implementation to some extent, if only to ensure that the assessment of outcomes or impacts can be logically linked to program activities. There are a variety of designs for formative evaluation, including *implementation evaluation*, *process studies*, and *evaluability assessment*, and they are covered later in this handbook. And there are a variety of specific designs intended to capture outcomes and impacts, and they are covered later in this text as well.

The timing of the evaluation can range across a continuum from a one-shot study of a specific aspect of implementation or one set of outcomes to an ongoing assessment system. The routine measurement of program inputs, outputs, or intermediate outcomes may be extremely useful for assessment of trends and should provide data that will be useful for more focused one-shot studies.

Traditional social science research methods have called for objective, neutral, and detached observers to measure the results of experiments and studies. However, as professional evaluation standards prescribe, program stakeholders should also be involved to ensure that the results of evaluation work of any kind will be used. The issue really is the level of participation of these stakeholders, who can include program staff, clients, beneficiaries, funders, and volunteers, to name a few. For example, various stakeholders could be consulted or given some degree of decision-making authority in evaluation design, data collection, interpretation of findings, and framing of recommendations.

Evaluators make judgments about the value, or worth, of programs (Scriven, 1980). When making determinations about the appropriateness, adequacy, quality, efficiency, or effectiveness of program operations and results, evaluators may rely on existing criteria provided in laws, regulations, mission statements, or grant applications. Goals may be clarified, and targets for performance may be given in such documentation. But in some cases evaluators are not given such criteria, and may have to seek guidance from stakeholders, professional standards, or other evaluation studies to help them make judgments. When there are no explicit expectations for program outcomes given, or unclear goals are espoused for a program (i.e., it appears to be "goal-free"), evaluators find themselves constructing the evaluation criteria. In any case, if the evaluators find unexpected outcomes (whether good or bad), these should be considered in the evaluation.

The terms *qualitative* and *quantitative* have a variety of connotations in the social sciences. For example, a qualitative research approach or mind-set means taking an inductive and open-ended approach in research and broadening questions as the research evolves. Qualitative data are typically words or visual images whereas quantitative data are typically numbers. The most common qualitative data collection methods are interviews (other than highly structured interviews), focus groups, and participant observation. Open-ended responses to survey questions can provide qualitative data as well. The most common sources of quantitative data are administrative records and structured surveys conducted via Internet and mail. Mixed-method approaches in evaluation are very common, and that means that both quantitative and qualitative data are used, and quantitative and qualitative data collection methods are used in combination (see Greene, 2007, for more on use of mixed methods). The extent to which an evaluation uses more quantitative or more qualitative methods and the relative reliance on quantitative or qualitative data should be driven by the questions the evaluation needs to answer and the audiences for the work.

And finally, the relative importance of the primary reason for the evaluation matters. That is, are assumptions that problems exist driving the demand for the application of evaluation methods the driver? When evaluators are asked to investigate problems, especially if they work for government bodies such as the U.S. Government Accountability Office, state audit agencies, or inspector general offices, the approaches and strategies they use for engaging stakeholders, and collecting data may be different from those used by evaluators in situations in which they are not perceived as collecting data due to preconceptions of fault.

Identify Contextual Elements That May Affect Evaluation Conduct and Use

The context for employing evaluation matters. The *context* includes both the broader environment surrounding evaluation and the immediate situation in which an evaluation study is planned. Since the beginning of the twenty-first century, daunting standards for evaluation of social programs have been espoused by proponents of *evidence-based* policy, management, and practice. Nonprofit organizations have promoted the use of evaluation to inform policy deliberations at all level of governments (For example, see Pew-MacArthur, 2014). The Cochrane and Campbell Collaborations and similar organizations have given guidance that randomized controlled trials (RCTs) are the "gold standard" for evaluation. Yet, ethical prohibitions, logistical impossibilities, and constrained resources frequently do not allow random assignment of subjects in evaluation of some social services, and some government programs with broad public mandates, such as environmental protection and national security. In such situations, less sophisticated approaches can provide useful estimates of program impact.

The key question facing evaluators is what type and how much evidence will be sufficient? Will the evidence be convincing to the intended audiences—be they nonprofit boards, legislators, or the public? The stakes have risen for what constitutes adequate evidence, and for many social service providers the term *evidence-based practice* is intimidating. There is not full agreement in virtually any field about when evidence is sufficient. And funders are likely to be aware of the rising standards for hard evidence and some may be unrealistic about what can be achieved by evaluators operating with finite resources.

It is usually difficult to establish causal links between program interventions and behavioral change. Numerous factors affect outcomes. Human as well as natural systems are complex and adaptive; they evolve in ways that evaluators may not be able to predict. Increasingly, attention has been drawn to using systems theory to inform evaluations of interventions designed to change behaviors in such complex systems.

Programs are typically located in multicultural environments. *Cultural competence* (also discussed as cultural humility) is a skill that has become more crucial for evaluators to develop than ever before. There are many important differences across program stakeholders, and expectation for evaluators to understand and address these differences in their work are high. Adequate knowledge of the social, religious, ethnic, and cultural norms and values of program stakeholders, especially beneficiaries who may present a large number of different backgrounds, presents another very important challenge to evaluators trying to understand the complex context in which a program

operates. Evaluators need to understand the human environment of programs so that data collection and interpretation are appropriate and realistic. Chapter Twelve describes culturally responsive evaluation and provides guidance on incorporating cultural competency into evaluation work.

Characteristics of the particular program to be evaluated can also affect the evaluation approach to be used. Evaluators may find themselves working with program staff who lack any experience with evaluation or, worse, have had bad experiences with evaluation or evaluators. Many organizations are simply not evaluation-friendly. A compliance culture has grown up in many quarters in which funders' requirements for data have risen, and so managers and administrators may feel that providing data to meet reporting demands is simply part of business as usual but has nothing to do with organizational learning to improve programs (for example, see Dahler-Larsen, 2012).

Finally, the operational issues facing evaluators vary across context. Challenging institutional processes may need to be navigated. Institutional review board processes and other clearances, such as the U.S. federal requirements for clearance of survey instruments when more than nine persons will be surveyed, take time and institutional knowledge. Site-specific obstacles to obtaining records and addressing confidentiality concerns can arise. Obtaining useful and sufficient data is not easy, yet it is necessary for producing quality evaluation work.

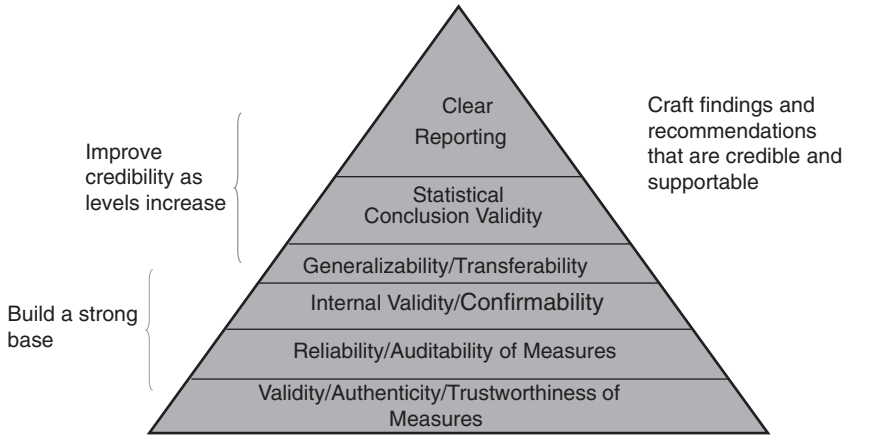
Produce the Methodological Rigor Needed to Support Credible Findings

The strength of findings, conclusions, and recommendations about program implementation and results depends on well-founded decisions regarding evaluation design and measurement. Figure 1.2 presents a graphical depiction of the way that credibility is supported by the methodological rigor ensured by wise decisions about measurement and design. This section focuses first on getting the most appropriate and reliable measures for a given evaluation and then on designing the evaluation to assess, to the extent possible, the extent to which the program being evaluated affected the measured outcomes.

Choose Appropriate Measures

Credible evaluation work requires clear, valid measures that are collected in a reliable, consistent fashion. Strong, well-founded measurement provides the foundation for methodological rigor in evaluation as well as in research and is the first requirement for useful evaluation findings. Evaluators must begin with credible measures and strong procedures in place to ensure that both quantitative and qualitative measurement is rigorous. The criteria used to assess

FIGURE 1.2. DESIGN EVALUATION STUDIES TO PROVIDE CREDIBLE FINDINGS: THE PYRAMID OF STRENGTH.



the rigor of quantitative and qualitative data collection, and inferences based on the two types of data, vary in terminology, but the fundamental similarities across the criteria are emphasized here.

The *validity or authenticity of measurement* is concerned with the accuracy of measurement, so that the measure accurately assesses what the evaluator intends to evaluate. Are the data collection procedures appropriate, and are they likely to provide reasonably accurate information? (See Part Two for discussions of various data collection procedures.) In practical evaluation endeavors, evaluators will likely use both quantitative and qualitative measures, and for both the relevance, legitimacy, and clarity of measures to program stakeholders and to citizens will matter. Often the items or concepts to measure will not be simple, nor will measurement processes be easy. Programs are composed of complex sets of activities to be measured. Outcomes to be measured may include both individual and group behaviors and may be viewed as falling on a short-term to long-term continuum, depending on their proximity to program implementation.

Measures may be *validated*, that is, tested for their accuracy, through several different processes. For example, experts may be asked to comment on the *face validity* of the measures. In evaluation work the term *experts* means the persons with the most pertinent knowledge about and experience with the behaviors to be measured. They may be case workers involved in service delivery, they may be principals and teachers, or they may be the program's customers, who

provide information on what is important to them. Box 1.1 provides tips for probing the validity and authenticity of measures.

Box 1.1. Questions to Ask When Choosing Measures

- Are the measures relevant to the activity, process, or behavior being assessed?
- Are the measures important to citizens and public officials?
- What measures have other experts and evaluators in the field used?
- What do program staff, customers, and other stakeholders believe is important to measure?
- Are newly constructed measures needed, and are they credible?
- Do the measures selected adequately represent the potential pool of similar measures used in other locations and jurisdictions?

Credibility can also be bolstered through testing the measures after data are collected. For example, evaluators can address the following questions with the data:

- Do the measures correlate to a specific agreed-upon standard or criterion measure that is credible in the field?
- Do the measures correlate with other measures in ways consistent with existing theory and knowledge?
- Do the measures predict subsequent behaviors in ways consistent with existing theory and knowledge?

Choose Reliable Ways to Obtain the Chosen Measures

The measures should be reliable. For quantitative data, r*eliability* refers to the extent to which a measure can be expected to produce similar results on repeated observations of the same condition or event. Having reliable measures means that operations consistently measure the same phenomena and consistently record data with the same decision criteria. For example, when questions are translated into multiple languages for respondents of different cultural backgrounds, evaluators should consider whether the questions will still elicit comparable responses from all. Data entry can also be a major source of error. Evaluators need to take steps to minimize the likelihood of errors in data entry.

For qualitative data, the relevant criterion is the *auditability* of measurement procedures. Auditability entails clearly documenting the procedures

used to collect and record qualitative data, such as documenting the circumstances in which data were obtained and the coding procedures employed. See Chapter Twenty-Two for more on coding qualitative data in a clear and credible manner.

In order to strengthen reliability or auditability of measures and measurement procedures, evaluators should adequately pretest data collection instruments and procedures and then plan for quality control procedures when in the field and when processing the information back home. (Also see Box 1.2.)

Box 1.2. Tips on Enhancing Reliability

- Pretest data collection instruments with representative samples of intended respondents before going into the field.
- Implement adequate quality control procedures to identify inconsistencies in interpretation of words by respondents in surveys and interviews.
- When problems with the clarity of questions are uncovered, the questions should be revised, and evaluators should go back to resurvey or re-interview if the responses are vital.
- Adequately train observers and interviewers so that they consistently apply comparable criteria and enter data correctly.
- Implement adequate and frequent quality control procedures to identify obstacles to consistent measurement in the field.
- Test levels of consistency among coders by asking all of them to code the same sample of the materials.

There are statistical tests that can be used to test for intercoder and interobserver reliability of quantitative data, such as Cronbach's alpha. When statistical tests are desired, research texts or Web sites should be consulted (for example, see the Sage Research Methods website at http://srmo.sagepub.com/view/encyclopedia-of-survey-research-methods/n228.xml).

Supporting Causal Inferences

In order to test the effectiveness of programs, researchers must ensure their ability to make well-founded inferences about (1) relationships between a program and the observed effects (internal validity) and (2) generalizability or

transferability of the findings. With quantitative data this may include testing for the statistical conclusion validity of findings.

Internal Validity

Internal validity is concerned with the ability to determine whether a program or intervention has produced an outcome and to determine the magnitude of that effect. When considering the internal validity of an evaluation, the evaluator should assess whether a causal connection can be established between the program and an intended effect and what the extent is of this relationship. Internal validity is also an issue when identifying the unintended effects (good or bad) of the program. When employing case studies and other qualitative research approaches in an evaluation, the challenge is typically to identify and characterize causal mechanisms needed to produce desired outcomes, and the term *confirmability* is more often applied to this process.

When making causal inferences, evaluators must measure several elements:

- The timing of the outcomes, to ensure that observed outcomes occurred after the program was implemented;
- The extent to which the changes in outcomes occurred after the program was implemented; and
- The presence of confounding factors: that is, factors that could also have produced desired outcomes.

In addition, observed relationships should be in accordance with expectations from previous research or evaluation work. It can be very difficult to draw causal inferences. There are several challenges in capturing the *net impacts* of a program, because other events and processes are occurring that affect achievement of desired outcomes. The time needed for the intervention to change attitudes or behavior may be longer than the time given to measure outcomes. And there may be flaws in the program design or implementation that reduce the ability of the program to produce desired outcomes. For such reasons, it may be difficult to establish causation credibly. It may be desirable to use terms such as *plausible attribution* when drawing conclusions about the effects of programs on intended behaviors. Box 1.3 offers tips about strengthening causal inferences about program results.

Some evaluations may be intended to be relevant to and used by only the site where the evaluation was conducted. However, in other situations the evaluation is expected to be relevant to other sites as well. This situation is discussed in the next section, on generalizing findings.

Box 1.3. Tips on Strengthening Inferences About Program Effects

- Measure the extent to which the program was actually implemented as intended.
- Ask key stakeholders about other events or experiences they may have had that also affected decisions relevant to the program—before and during the evaluation time frame.
- Given existing knowledge about the likely time period needed to see effects, explore whether enough time has elapsed between implementation of the program and measurement of intended effects.
- Review previous evaluation findings for similar programs to identify external factors and unintended effects, and build in capacity to measure them.

Generalizability

Evaluation findings possess *generalizability* when they can be applied beyond the groups or context being studied. With quantitative data collection the ability to generalize findings from a statistical sample to a larger population (or other program sites or future clients) refers to statistical conclusion validity (discussed below). For qualitative data, the *transferability* of findings from one site to another (or the future) may present different, or additional, challenges. Concluding that findings from work involving qualitative data are fit to be transferred elsewhere likely require more extensive contextual understanding of both the evaluation setting and the intended site for replication (see Cartwright, 2013 and Patton, 2011, for guidance on replicating and scaling up interventions). All the conditions discussed previously for internal validity also need to be met for generalizing evaluation findings. In addition, it is desirable that the evaluation be conducted in multiple sites, but at the least, evaluators should select the site and individuals so they are representative of the populations to which the evaluators hope to generalize their results.

Special care should be taken when trying to generalize results to other sites in evaluations of programs that may have differential effects on particular subpopulations such as youths, rural groups, or racial or ethnic groups. In order to enhance generalizability, evaluators should make sampling choices to identify subpopulations of interest and should ensure that subsamples of the groups are large enough to analyze. However, evaluators should still examine each sample to ensure that it is truly representative of the larger population to which they hope to generalize on demographic variables of interest (for example, age or ethnic grouping). Box 1.4 offers tips about strengthening the generalizability of findings.

Statistical Conclusion Validity

Statistical generalizability requires testing the statistical significance of findings from probability samples, and is greatly dependent on the size of the samples used in an evaluation. Chapter Twenty-Three provides more background on the use of statistics in evaluation. But it bears noting that the criterion of statistical significance and the tests related to it have been borrowed from the physical sciences, where the concern is to have the highest levels of confidence possible. In program evaluation practice, where obstacles may exist to obtaining large samples, it is reasonable to consider confidence levels lower than the 95 or 99 percent often used in social science research. For instance, it may be reasonable to accept a 90 percent level of confidence. It is entirely appropriate to report deliberations on this issue, reasons why a certain level was chosen, and the exact level of significance the findings were able to obtain. This is more realistic and productive than assuming that evaluation results will not be discussed unless a, perhaps unrealistically, high level of confidence is reached.

Box 1.4. Questions to Ask to Strengthen the Generalizability of Findings

- To what groups or sites will generalization be desired?
- What are the key demographic (or other) groups to be represented in the sample?
- What sample size, with adequate sampling of important subgroups, is needed to make generalizations about the outcomes of the intervention?
- What aspects of the intervention and context in which it was implemented merit careful measurement to enable generalizability or transferability of findings?

In order to report properly on an evaluation, evaluators should report both on the statistical significance of the findings (or whether the sample size allows conclusions to be drawn about the evaluation's findings), and on the importance and relevance of the size of the measured effects. Because statistical significance is strongly affected by sheer sample size, other pertinent criteria should be identified to characterize the policy relevance of the measured effects.

Reporting

In the end, even careful planning and reasoned decision making about both measurement and design will not ensure that all evaluations will

produce perfectly credible results. There are a variety of pitfalls that frequently constrain evaluation findings, as described in Chapter Twenty-Six. Clarity in reporting findings and open discussion about methodological decisions and any obstacles encountered during data collection will bolster confidence in findings.

Planning a Responsive and Useful Evaluation

Even with the explosion of quantitative and qualitative evaluation methodologies since the 1970s, designing evaluation work requires both social science knowledge and skills and cultivated professional judgment. The planning of each evaluation effort requires difficult trade-off decisions as the evaluator attempts to balance the feasibility and cost of alternative evaluation designs against the likely benefits of the resulting evaluation work. Methodological rigor must be balanced with resources, and the evaluator's professional judgment will arbitrate the trade-offs.

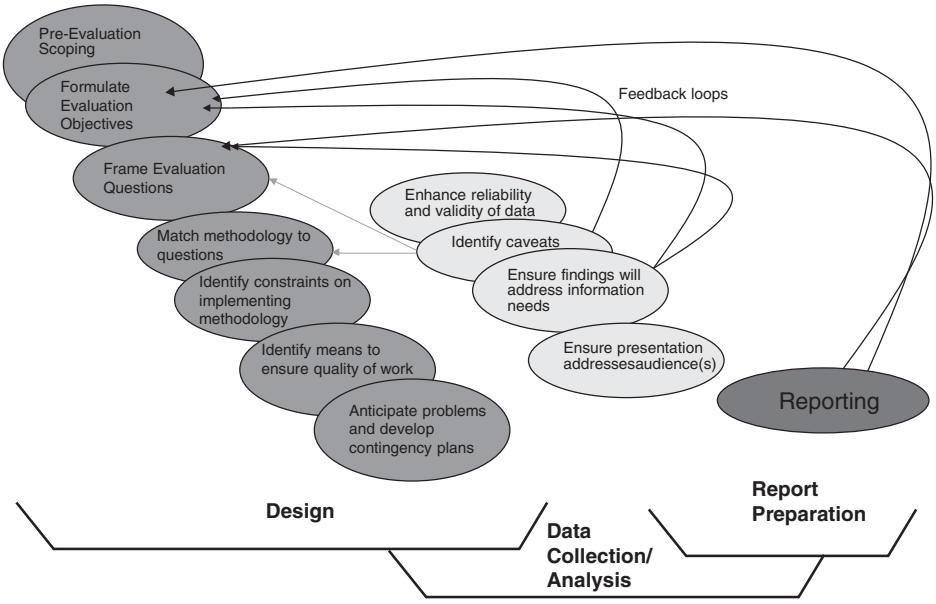
Wherever possible, evaluation planning should begin before the program does. The most desirable window of opportunity for evaluation planning opens when new programs are being designed. Desired data can be more readily obtained if provision is made for data collection from the start of the program, particularly for such information as clients' pre-program attitudes and experiences. These sorts of data might be very difficult, if not impossible, to obtain later.

Planning an evaluation project requires selecting the measures that should be used, an evaluation design, and the methods of data collection and data analysis that will best meet information needs. To best inform choices, evaluators learn how the evaluation results might be used and how decision making might be shaped by the availability of the performance data collected. However, it is important to recognize that evaluation plans are organic and likely to evolve. Figure 1.3 displays the key steps in planning and conducting an evaluation. It highlights many feedback loops in order to stress how important it is for evaluators to be responsive to changes in context, data availability, and their own evolving understanding of context.

Planning Evaluation Processes

Identification of the key evaluation questions is the first, and frequently quite challenging, task faced during the design phase. Anticipating what clients need

FIGURE 1.3. REVISE QUESTIONS AND APPROACHES AS YOU LEARN MORE DURING THE EVALUATION PROCESS.



to know is essential to effective evaluation planning. For example, the U.S. Government Accountability Office (GAO) conducts many program evaluations in response to legislative requests. These requests, however, are frequently fairly broad in their identification of the issues to be addressed. The first task of GAO evaluators is to more specifically identify what the committees or members of Congress want to know, and then to explore what questions should be asked to acquire this information. (See Box 1.5 for more information on the GAO's evaluation design process.)

Box 1.5. GAO's Evaluation Design Process

Stephanie Shipman

U.S. Government Accountability Office

Each year, GAO receives hundreds of requests to conduct a wide variety of studies, from brief descriptions of program activities to in-depth evaluative assessments of program or policy effectiveness. Over time, GAO has drawn lessons from its experience to develop a systematic, risk-based process for selecting the most appropriate

approach for each study. Policies and procedures have been created to ensure that GAO provides timely, quality information to meet congressional needs at reasonable cost; they are summarized in the following four steps: (1) clarify the study objectives; (2) obtain background information on the issue and design options; (3) develop and test the proposed approach; and (4) reach agreement on the proposed approach.

Clarify the Study Objectives

The evaluator's first step is to meet with the congressional requester's staff to gain a better understanding of the requester's need for information and the nature of the research questions and to discuss GAO's ability to respond within the desired time frame. Discussions clarify whether the questions are primarily descriptive—such as how often something occurs—or evaluative—involving assessment against a criterion. It is important to learn how the information is intended to be used and when that information will be needed. Is it expected to inform a particular decision or simply to explore whether a topic warrants a more comprehensive examination? Once the project team has a clearer understanding of the requester's needs, the team can begin to assess whether additional information will be needed to formulate the study approach or whether the team has enough information to commit to an evaluation plan and schedule.

In a limited number of cases, GAO initiates work on its own to address significant emerging issues or issues of broad interest to the Congress. In these studies, GAO addresses the same considerations in internal deliberations and informs majority and minority staff of the relevant congressional committees of the planned approach.

Obtain Background Information

GAO staff review the literature and other work to understand the nature and background of the program or agency under review. The project team will consult prior GAO and inspector general work to identify previous approaches and recommendations, agency contacts, and legislative histories for areas in which GAO has done recent work. The team reviews the literature and consults with external experts and program stakeholders to gather information about the program and related issues, approaches used in prior studies, and existing data sources. Evaluators discuss the request with agency officials to explore their perspectives on these issues.

GAO evaluators explore the relevance of existing data sources to the research questions and learn how data are obtained or developed in order to assess their completeness and reliability. Evaluators search for potential evaluative criteria in legislation, program design materials, agency performance plans, professional standards, and elsewhere, and assess their appropriateness to the research

(*Continued*)

question, objectivity, suitability for measurement, and credibility to key program stakeholders.

Develop and Test the Proposed Approach

The strengths and limitations of potential data sources and design approaches are considered in terms of which ones will best answer the research questions within available resource and time constraints. Existing data sources are tested to assess their reliability and validity. Proposed data collection approaches are designed, reviewed, and pretested for feasibility given conditions in the field. Evaluators outline work schedules and staff assignments in project plans to assess what resources will be required to meet the desired reporting timelines. Alternative options are compared to identify the trade-offs involved in feasibility, data validity, and the completeness of the answer likely to be obtained.

Evaluation plans are outlined in a design matrix to articulate the proposed approach in table format for discussion with senior management (see Figure 1.4 later in this chapter). The project team outlines, for each research question, the information desired, data sources, how the data will be collected and analyzed, the data's limitations, and what this information will and will not allow the evaluators to say. Discussions of alternative design options focus on the implications that any limitations identified will have on the analysis and the evaluator's ability to answer the research questions. What steps might be taken to address (reduce or counterbalance) such limitations? For example, if the primary data source relies on subjective self-reports, can the findings be verified through more objective and reliable documentary evidence?

Discussion of "what the analysis will allow GAO to say" concerns not what the likely answer will be but what sort of conclusion one can draw with confidence. How complete or definitive will the answer be to the research question? Alternatively, one might characterize the types of statements one will not be able to make: for example, statements that generalize the findings from observed cases to the larger population or to time periods preceding or following the period examined.

Reach Agreement on the Proposed Approach

Finally, the proposed approach is discussed both with GAO senior management in terms of the conclusiveness of the answers provided for the resources expended and with the congressional requester's staff in terms of whether the proposed information and timelines will meet the requester's needs. GAO managers review the design matrix and accompanying materials to determine whether the proposed approach adequately addresses the requester's objectives, the study's risks have been adequately identified and addressed, and the proposed resources are appropriate given the importance of the issues involved and other work requests. The GAO team then meets with the requester's staff to discuss the engagement

methodology and approach, including details on the scope of work to be performed and the product delivery date. The agreed-upon terms of work are then formalized in a commitment letter.

Matching evaluation questions to a client's information needs can be a tricky task. When there is more than one client, as is frequently the case, there may be multiple information needs, and one evaluation may not be able to answer all the questions raised. This is frequently a problem for nonprofit service providers, who may need to address multiple evaluation questions for multiple funders.

Setting goals for information gathering can be like aiming at a moving target, for information needs change as programs and environmental conditions change. Negotiating evaluable questions with clients can be fraught with difficulties for evaluators as well as for managers who may be affected by the findings.

The selection of questions should drive decisions on appropriate data collection and analysis. As seen in Figure 1.4, the GAO employs a design tool it calls the *design matrix* that arrays the decisions on data collection and analysis by each question. This brief, typically one-page blueprint for the evaluation is used to secure agreement from various stakeholders within the GAO, such as technical experts and substantive experts, and to ensure that answers to the questions will address the information needs of the client, in this case the congressional requestor. Although there is no one ideal format for a design matrix, or evaluation blueprint, the use of some sort of design tool to facilitate communication about evaluation design among stakeholders is very desirable. An abbreviated design matrix can be used to clarify how evaluation questions will be addressed through surveying (this is illustrated in Chapter Fourteen).

A great deal of evaluation work performed for public and nonprofit programs is contracted out, and given current pressures toward outsourcing along with internal evaluation resource constraints, this trend is likely to continue. Contracting out evaluation places even more importance on identifying sufficiently targeted evaluation questions. Statements of work are typically prepared by internal program staff working with contract professionals, and these documents may set in stone the questions the contractors will address, along with data collection and analysis specifications. Unfortunately, the contract process may not leave evaluators (or program staff) much leeway in reframing the questions in order to make desired adjustments when the project gets under way and confronts new issues or when political priorities shift. Efforts should be made to allow the contractual process to permit

FIGURE 1.4. SAMPLE DESIGN MATRIX.

Issue problem statement:
Guidance:

1. Put the issue into context.
2. Identify the potential users.

Researchable Question(s)	Criteria and Information Required and Source(s)	Scope and Methodology, Including Data Reliability	Limitations	What This Analysis Will Likely Allow GAO to Say
What question(s) is the team trying to answer?	What information does the team need to address the question? Where will they get it?	How will the team answer each question?	What are the engagement's design limitations and how will they affect the product?	What are the expected results of the work?
Question 1				
Question 2				
Question 3				
Question 4				

Source: U.S. Government Accountability Office.

contextually-driven revisions. See Chapter Twenty-Nine for more guidance on effectively contracting out evaluation work.

Balancing clients' information needs with resources affects selection of an evaluation design as well as specific strategies for data collection and analysis. Selecting a design requires the evaluator to anticipate the amount of rigor that will be required to produce convincing answers to the client's questions. Evaluators must specify the comparisons that will be needed to demonstrate whether a program has had the intended effects and the additional comparisons needed to clarify differential effects on different groups.

The actual nature of an evaluation design should reflect the objectives and the specific questions to be addressed. This text offers guidance on the wide variety of evaluation designs that are appropriate given certain objectives and questions to address. Table 1.1 arrays evaluation objectives with designs and also identifies the chapters in this text to consult for guidance on design. The wide range of questions that be framed about programs is matched by the variety of approaches and designs that are employed by professional evaluators.

Resource issues will almost always constrain design choices; staff costs, travel costs, data collection burdens on program staff, and political and bureaucratic costs may limit design options. Evaluation design decisions, in turn, affect where and how data will be collected. To help evaluators and program

TABLE 1.1. MATCHING DESIGNS AND DATA COLLECTION TO THE EVALUATION QUESTIONS.

Evaluation Objective	Illustrative Questions	Possible Design	Corresponding Handbook Chapter(s)
1. Describe program activities	Who does the program affect–both targeted organizations and affected populations? What activities are needed to implement the program (or policy)? By whom? How extensive and costly are the program components? How do implementation efforts vary across delivery sites, subgroups of beneficiaries, and/or across geographical regions? Has the program (policy) been implemented sufficiently to be evaluated?	Performance Measurement Exploratory Evaluations Evaluability Assessments Multiple Case Studies	Chapter 4 Chapter 5 Chapter 8 Chapter 11 Chapter 12
2. Probe implementation and targeting	To what extent has the program been implemented? When evidence-based interventions are implemented, how closely are the protocols implemented with fidelity to the original design? What key contextual factors tare likely to affect the ability of the program implementers to have the intended outcomes? What feasibility or management challenges hinder successful implementation of the program?	Multiple Case Studies Implementation or Process evaluations Performance Audits Compliance Audits	Chapter 4 Chapter 8 Chapter 10 Chapter 11 Chapter 12

(Continued)

TABLE 1.1. MATCHING DESIGNS AND DATA COLLECTION TO THE EVALUATION QUESTIONS. ***(Continued)***

Evaluation Objective	Illustrative Questions	Possible Design	Corresponding Handbook Chapter(s)
	To what extent have activities undertaken affected the populations or organizations targeted by the regulation? To what extent are implementation efforts in compliance with the law and other pertinent regulations? To what extent does current program (or policy) targeting leave significant needs (problems) not addressed?		
3. Measure program impact	Has implementation of the program produced results consistent with its design (espoused purpose)? How have measured effects varied across implementation approaches, organizations, and/or jurisdictions? For which targeted populations has the program (or policy) consistently failed to show intended impact? Is the implementation strategy more (or less) effective in relation to its costs? Is the implementation strategy more cost effective than other implementation strategies also addressing the same problem?	Experimental Designs, that is Random Control Trials (RCTs) Difference-in-Difference Designs Propensity Score Matching (PSM) Statistical Adjustments with Regression Estimates of Effects Multiple Time Series Designs Regression Discontinuity Designs Cost-Effectiveness Studies Benefit-Cost Analysis Systematic Reviews Meta-Analyses	Chapter 6 Chapter 7 Chapter 25

(Continued)

TABLE 1.1. MATCHING DESIGNS AND DATA COLLECTION TO THE EVALUATION QUESTIONS. ***(Continued)***

Evaluation Objective	Illustrative Questions	Possible Design	Corresponding Handbook Chapter(s)
	What are the average effects across different implementations of the program (or policy)?		
4. Explain how and why programs produce intended and unintended effects	How and why did the program have the intended effects? Under what circumstances did the program produce the desired effects? To what extent have program activities had important unanticipated negative spillover effects? What are unanticipated positive effects of the program that emerge over time, given the complex web of interactions between the program and other programs, and who benefits? For whom (which targeted organizations and/or populations) is the program more likely to produce the desired effects? What is the likely impact trajectory of the program (over time)? How likely is it that the program will have similar effects in other contexts (beyond the context studied)? How likely is it that the program will have similar effects in the future?	Multiple Case Studies Meta-Analyses Impact Pathways and Process Tracing Contribution Analysis Non-Linear Modeling, System Dynamics Configurational Analysis, e.g., Qualitative Case Analysis (QCA) Realist-Based Synthesis	Chapter 8 Chapter 25

personnel make the best design decisions, a pilot test of proposed data collection procedures should be considered. Pilot tests may be valuable in refining evaluation designs; they can clarify the feasibility and costs of data collection as well as the likely utility of different data analysis strategies.

Data Collection

Data collection choices may be politically as well as bureaucratically tricky. Exploring the use of existing data involves identifying potential political barriers as well as more mundane constraints, such as incompatibility of computer systems. Planning for data collection in the field should be extensive in order to help evaluators obtain the most relevant data in the most efficient manner. Chapters Thirteen through Twenty-One present much detail on both selecting and implementing a variety of data collection strategies.

Data Analysis

Deciding how the data will be analyzed affects data collection, for it forces evaluators to clarify how each data element will be used. Collecting too much data is an error that evaluators frequently commit. Developing a detailed *data analysis plan* as part of the evaluation design can help evaluators decide which data elements are necessary and sufficient, thus avoiding the expense of gathering unneeded information.

An analysis plan helps evaluators structure the layout of a report, for it identifies the graphs and tables through which the findings will be presented. Anticipating how the findings might be used forces evaluators to think carefully about presentations that will address the original evaluation questions in a clear and logical manner.

Identifying relevant questions and answering them with data that have been analyzed and presented in a user-oriented format should help to ensure that evaluation results will be used. However, communicating evaluation results entails more than simply drafting attractive reports. If the findings are indeed to be used to improve program performance, as well as respond to funders' requests, the evaluators must understand the bureaucratic and political contexts of the program and craft their findings and recommendations in such a way as to facilitate their use in these contexts.

Using Evaluation Information

The goal of conducting any evaluation work is certainly to make positive change. When one undertakes any evaluation work, understanding from the

outset how the work may contribute to achieving important policy and program goals is important. Program improvement is the ultimate goal for most evaluators. Consequently, they should use their skills to produce useful, convincing evidence to support their recommendations for program and policy change.

Box 1.6. Anticipate These Challenges to the Use of Evaluation and Performance Data

1. Lack of visible appreciation and support for evaluation among leaders
2. Unrealistically high expectations of what can be measured and "proven"
3. A compliance mentality among staff regarding collection and reporting of program data and a corresponding disinterest in data use
4. Resistance to adding the burden of data collection to staff workloads
5. Lack of positive incentives for learning about and using evaluation and data
6. Lack of compelling examples of how evaluation findings or data have been used to make significant improvements in programs
7. Poor presentation of evaluation findings

Understanding how program managers and other stakeholders view evaluation is also important for evaluators who want to produce useful information. Box 1.6 lists some fairly typical reactions to evaluation in public and nonprofit organizations that may make it difficult for evaluators to develop their approaches and to promote the use of findings (for example, see Hatry, 2006; Mayne, 2010; Newcomer, 2008; Pawson, 2013; and Preskill and Torres, 1999). Clear and visible commitment by leadership is always critical, as are incentives within the organization that reward use. The anticipation that evaluation will place more burdens on program staff and clients is a perception that evaluators need to confront in any context.

The most effective evaluators are those who plan, design, and implement evaluations that are sufficiently relevant, responsive, and credible to stimulate program or policy improvement. Evaluation effectiveness may be enhanced by efficiency and the use of practical, low-cost evaluation approaches that encourage the evaluation clients (the management and staff of the program) to accept the findings and use them to improve their services.

Efforts to enhance the likelihood that evaluation results will be used should start during the planning and design phase. From the beginning, evaluators must focus on mediating obstacles and creating opportunities to promote use. Box 1.7 provides tips for increasing the likelihood that the findings will

be used. Six of these tips refer to actions that need to be taken during evaluation design. Evaluators must understand and typically shape their audiences' expectations, and then work consistently to ensure that the expectations are met. Producing methodologically sound findings and explaining why they are sound both matter.

Box 1.7. Tips on Using Evaluation Findings and Data

1. Understand and appreciate the relevant perspectives and preferences of the audience (or audiences!) to shape communication of evaluation findings and performance data.
2. Address the questions most relevant to the information needs of the audience.
3. Early in the design phase, envision what the final evaluation products should contain.
4. Design sampling procedures carefully to ensure that the findings can be generalized to whomever or wherever the key stakeholders wish.
5. Work to ensure the validity and authenticity of measures, and report on the efforts to do so.
6. Address plausible alternative explanations for the measured program outcomes.
7. Clearly communicate the competence of the evaluators and the methodology employed to enhance the credibility of findings.
8. When quantitative analytical techniques are used, clarify why these techniques were appropriate and that adequate sample sizes were used.
9. In recommendations, to the extent politically feasible, state who should take what actions, where, and when.
10. Tailor reporting vehicles to address the communication preferences of different target audiences.
11. Provide an executive summary and a report written clearly and without jargon.
12. Work consistently from the beginning to develop strong working relationships with program staff and other pertinent stakeholders so that they will be willing to implement recommendations.

Clear presentation of both findings and feasible recommendations is also necessary, and these skills are discussed in depth in Chapters Twenty-Seven and Twenty-Eight.

Credibility of evaluation work in the eyes of the audiences, especially those people who need to implement recommended changes, is the goal for all evaluators. In the end, production of credible performance data and evaluation study findings that are communicated to funders and the broader public can contribute to the public good through informing policy and program management decisions.

Glossary

Case study. A rich description and analysis of a program in its context, typically using multiple modes of qualitative data collection.

Comparison group design. An assessment design that compares outcomes for program participants with outcomes for people in a comparison group.

Cost-benefit study. An analysis that compares the dollar value of program costs with the dollar value of program impacts.

Evaluation design. A plan for conducting an evaluation that specifies (1) a set of evaluation questions, (2) the targeted groups from whom data will be collected, and the timing of collection, (3) the data that will be collected, (4) the analyses that will be undertaken to answer the evaluation questions, (5) the estimated costs and time schedule for the evaluation work, and (6) how the evaluation information may be used.

Evaluation stakeholders. The individuals, groups, or organizations that can affect or are affected by an evaluation process or its findings, or both.

Experimental design. An assessment design that tests the existence of causal relationships by comparing outcomes for those randomly assigned to program services with outcomes for those randomly assigned to alternative services or no services. Also called a randomized experiment or random control trial (RCT).

Implementation evaluation. An assessment that describes actual program activities, typically to find out what actually happened or is happening in the program.

Interrupted time-series design. An assessment design that tests the existence of causal relationships by comparing trends in outcomes before and after the program.

Logic model (or program logic model). A flowchart that summarizes key elements of a program: resources and other inputs, activities, outputs (products and services delivered), and intermediate outcomes and end outcomes (short-term and longer-term results) that the program hopes to achieve. Logic models should also identify key factors that are outside the control of program staff but are likely to affect the achievement of desired outcomes. A logic model shows assumed cause-and-effect linkages among model elements, showing which activities are expected to lead to which outcomes, and it may also show assumed cause-and-effect linkages between external factors and program outcomes.

Outcomes. Changes in clients or communities associated with program activities and outputs.

Outputs. Products and services delivered to a program's clients.

Pre-post design. An assessment design that compares outcomes before and after the program.

Process evaluation. An assessment that compares actual with intended inputs, activities, and outputs.

Program. A set of resources and activities directed toward one or more common goals, typically under the direction of a single manager or management team.

Program logic model. See *logic model*.

Quasi-experimental design. An assessment design that tests the existence of a causal relationship where random assignment is not possible. Typical quasi-experimental designs include pre-post designs, comparison group designs, and interrupted time-series designs.

Randomized experiment or Random Control Trial (RCT). See *Experimental design*.

Regression discontinuity design. An experiment that assigns units to a condition on the basis of a score cutoff on a particular variable.

Stakeholder. See *evaluation stakeholders*.

Theory-based evaluation (TBE). A family of approaches that seek to explicate and test policy-makers', managers', and other stakeholders' assumptions (or 'theories') about how a program intends to bring about a desired change. Core elements of these theories are mechanisms (the 'nuts and bolts' of an intervention) and how they relate to context and outcomes.

References

American Evaluation Association. "The AEA's Guiding Principles for Evaluators." www.eval.org/p/cm/ld/fid=51. 2004.

Cartwright, Nancy. "Knowing What We Are Talking About: Why Evidence Doesn't Always Travel." *Evidence & Policy*, 2013, *9*(1), 97–112.

Dahler-Larsen, Peter. *The Evaluation Society*. Stanford, CA: Stanford University Press, 2012.

Greene, Jennifer. *Mixed Methods in Social Inquiry*. San Francisco, CA: Jossey-Bass, 2007.

Hatry, Harry. *Performance Measurement: Getting Results* (2nd ed.). Washington, DC: Urban Institute Press, 2006.

Joint Committee on Standards for Educational Evaluation. "Program Evaluation Standards." www.jcsee.org/program-evaluation-standards. 2010.

Mayne, J. "Building an Evaluative Culture: The Key to Effective Evaluation and Results Management." *Canadian Journal of Program Evaluation*, 2009, 24, 1–30.

Newcomer, Kathryn. "Assessing Program Performance in Nonprofit Agencies." In Patria de Lancer Julnes, Frances Stokes Berry, Maria P. Aristigueta, and Kaifeng Yang (Eds.), *International Handbook of Practice-Based Performance and Management Review*. Thousand Oaks, CA: Sage, 2008.

Patton, Michael Quinn. *Utilization-Focused Evaluation* (4th ed.). Thousand Oaks, CA: Sage, 2008.

Patton, Michael Quinn. *Developmental Evaluation: Applying Complexity Concepts to Enhance Innovation and Use*. New York: The Guilford Press, 2011.

Pawson, Ray. *The Science of Evaluation: A Realist Manifesto*. Thousand Oaks, CA: Sage, 2013.

The Pew Charitable Trusts. "Evidence-Based Policymaking: A Guide for Effective Government Pew Charitable Trusts." www.pewtrusts.org/en/research-and-analysis/reports/2014/11/evidence-based-policymaking-a-guide-for-effective-government. November 13, 2014.

Preskill, Hallie S., and Torres, Rosalie. *Evaluative Inquiry for Learning in Organizations*. Thousand Oaks, CA: Sage, 1999.

Sage Research Methods. http://srmo.sagepub.com/view/encyclopedia-of-survey-research-methods/n228.xml.

Scriven, Michael. *The Logic of Evaluation*. Inverness, CA.: Edgepress, 1980.

Shadish, William., Cook, Thomas D., and Campbell, Donald Thomas. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin, 2002.

Williams, Bob, and Hummelbrunner, Richard. *Systems Concepts in Action: A Practitioner's Toolkit*. Stanford, CA: Stanford University Press, 2011.