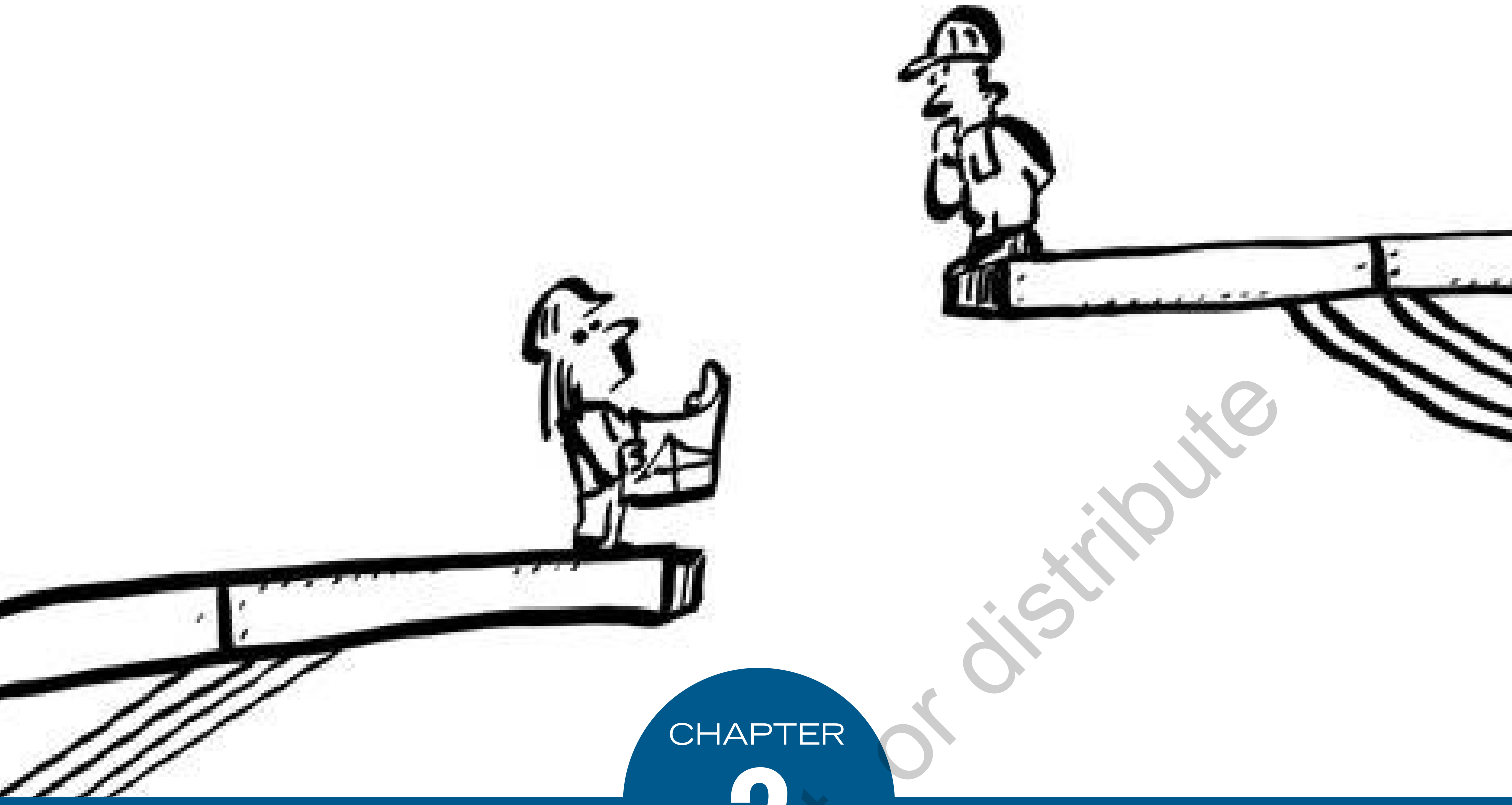


CHAPTER

2

Setting the Stage

Anupa Bir

Steven Sheingold

Learning Objectives

This chapter presents an overview of three evaluation typologies: formative, summative, and process evaluations. It introduces logic models as the intermediary between program goals and evaluation design, guiding the type and timing of data collection. Many challenges complicate researchers' ability to properly evaluate

(Continued)

(Continued)

program impacts, and among these challenges is the question of when to plan and implement the evaluation. The timing of evaluation planning and implementation can substantially influence the rigor, statistical validity, and credibility of the evaluation findings. This chapter presents the idea of a prospectively planned and integrated program evaluation, offering a series of related evaluation-planning tools to help neutralize these evaluation limitations.

Distinguishing the many types of evaluation discussed in the literature is important to set the stage for the chapters that follow. This book primarily targets what is often called impact evaluation, but this is just one among many types of program evaluation. In this chapter, we describe and distinguish between the different types. While these are typically thought of as alternative evaluation designs, they can also be envisioned as successive stages in a comprehensive evaluation process.

Typology for Program Evaluation[1,2]

Most program evaluations can be classified as one of three types: formative, process, or summative. Although the vast literature on evaluation typology encompasses frameworks different from the three-legged framework we describe here, each has essentially the same purpose: to help researchers determine the best approach for answering their programmatic and intervention-related questions. We therefore frame the three types of evaluations presented in relation to the central research questions each serves to answer.

1. **Formative Evaluation.** Generally, a formative evaluation gleans information that can contribute to further developing a program or intervention. It is most commonly used to answer questions that arise during program design and development, for example, "What components of the program should be included?"[3] Policymakers and program designers can find formative evaluation quite useful. Naturally this type of evaluation typically precedes implementation and as such, it has been termed *pretesting* or *developmental research*.[4] However, as discussed in later chapters, recent advances in both statistical methodologies and evaluative thinking have incorporated formative evaluation approaches into both rapid-cycle evaluation and theory-driven evaluation to aid in program improvement and in identifying what levers are central to change.

2. **Process Evaluation.** Process evaluation, as its name suggests, "investigates the procedures that were used to implement a policy program"[5] or an intervention. To answer the *what* and *how* questions that may follow a formative evaluation,

a process evaluation is a good approach. One of the foundational findings in evaluation was that services are rarely implemented as planned (or not implemented at all) and that the clients served are often not those for whom the services are intended. Process evaluation describes what is being implemented, to what extent it is being implemented, and who is actually receiving the intervention. As such, process evaluations also assess the fidelity with which a program is implemented.

3. **Summative Evaluation.** In contrast to formative evaluation, which precedes program implementation, summative evaluation answers questions about program outcomes and impacts. Research questions seeking to determine whether the program/intervention achieves the goal for which it was originally designed can be best answered by summative evaluation. This type of evaluation answers a fundamental research question: Is the program/intervention achieving the goals it is intended to achieve?[6] Even though summative evaluation requires data that are unavailable in full until program completions, the most complete and most efficient data collection requires even a summative evaluation to be planned prior to program implementation.

As in many areas related to evaluation and research, different individuals and groups can use different terminology to describe the same concept or method. For example, in this book, as in others, we use the term *impact evaluation* to describe evaluations that attempt to establish a causal link between a program and desired outcomes.[7] Such evaluations clearly fall under the general umbrella of summative evaluation. Often, these evaluations also separate outcomes and impacts, distinguished by timing and performance measures studied. Outcome evaluation measures program effects in the target population by assessing the progress in the outcomes or outcome objectives that the program is to achieve. It requires understanding of the kind of changes desired for participants, such as in learning, skills, behavior, and actions. Thus, evaluators must identify appropriate indicators that can measure these changes. Impacts are those effects consistent with the overall objectives of the intervention, such as measures of changes in health. For example, a program might train physicians in new ways to council patients on nutrition and weight loss. Outcome measures might be specific changes in physician practice, such as the number or content of counseling sessions, while impact might be measured by patient weight loss. As discussed in the chapters ahead, specifying the full range of performance measures for each stage in evaluation will be important. For the purposes of this book, we use the term *impact evaluation* in the general sense, that is, assessing all important program effects that might be associated with summative evaluation.

A modern, integrated, and comprehensive evaluation will draw on the lessons, practices, and principles of all three types of evaluation. For any given intervention, formative evaluation practices will inform program design; a process evaluation will assess if the design is implemented with fidelity and integrity; a summative evaluation will examine the program's success in achieving its goals. In rapid cycle evaluation, planned and unplanned variation can be documented by the process evaluation and combined with summative or impact evaluation findings to determine what components of the program or intervention are most significant in their contribution to particular outcomes. The quantitative and qualitative data collected in the observations required for process or diagnostic evaluation must be carefully selected, based on a theory of change that identifies the critical elements of a program and specifies their relation to proximal, intermediate, and long-term outcomes (as discussed later in this chapter). As programs are almost certainly adapted to local conditions when taken to scale, comprehensive integrated evaluations provide evidence on what are likely to be the core components of the intervention and which components are not associated with effectiveness—and can therefore be modified to better align, for example, with existing services.

After assessing the research questions, selecting the evaluation type will inform the steps that follow: timing of evaluation design, methods for actually conducting the evaluation (discussed in Part II), and measures involved in the ongoing monitoring required to conduct the evaluation (discussed in Part III), these measures being determinants of the data required to conduct the evaluation (as discussed in Chapter 3).

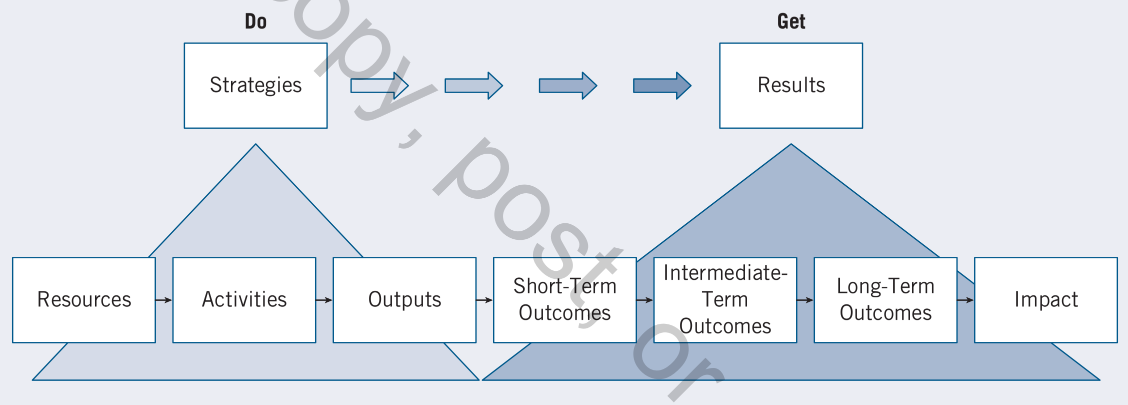
We emphasize again that although these are presented for simplicity as discrete forms of evaluation and may occur in isolation, an integrated comprehensive evaluation will draw on the lessons and practices of each, to create the evidence necessary for decision making that is maximally data informed. Integrating evaluation principles and practices into organizational practice will assist health care (and other service) providers to identify effective practices and the conditions necessary to support the implementation of those practices.

Planning an Evaluation: How Are the Changes Expected to Occur?

Programs and policies are implemented to achieve specific outcomes, generally by providing resources and incentives to change behaviors. Designing comprehensive evaluations therefore should begin by detailing the resources provided and exactly how the program is intended to change behaviors to achieve the desired outcomes. Such models of change are often called logic models or results chains.[8] They are typically a visual tool—usually represented as a flow diagram—intended to communicate the logic, or rationale, behind an effective program. These models detail a sequence of input, outputs, and activities that are expected to improve outcomes and final impacts. Essentially, the model should tell the story about what program resources are available to be used, how they are to be used by the program and participants, what short- or medium-term results are to be achieved, and the final outcomes.

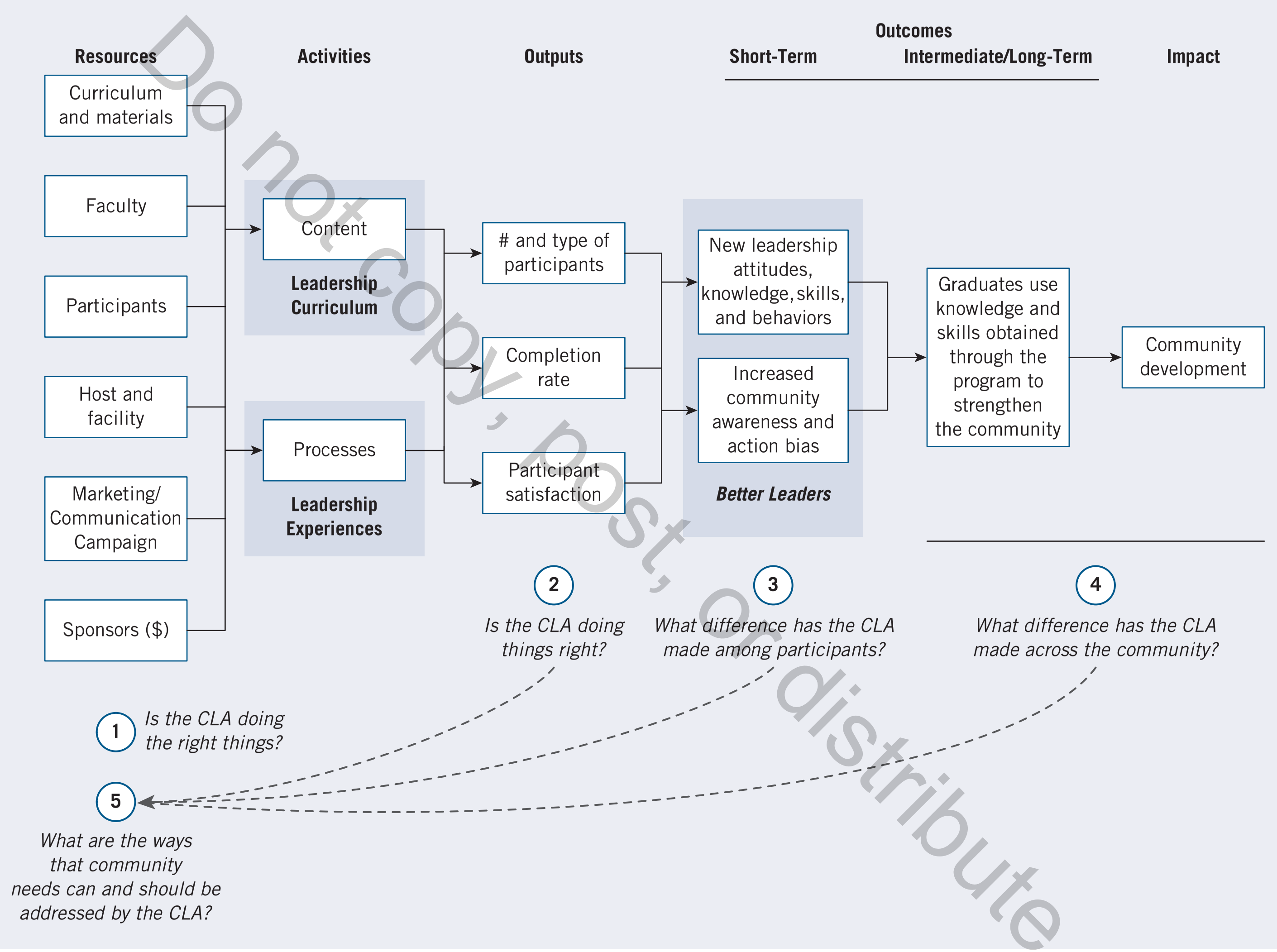
A good logic model or results chain can serve several important purposes for developing and conducting the evaluation. First, it makes developing an evaluation plan much easier by making explicit your expected outcomes throughout the

Figure 2.1 Generic Logic Model



Source: The Logic Model Guidebook, Lisa Wyatt Knowlton and Cynthia C. Phillips, SAGE Publishing © 2013.

Figure 2.2 Program-Specific Model: Community Leadership Academy Program


Resources
Activities
Outputs
Outcomes
Short-Term
Intermediate/Long-Term
Impact
Curriculum and materials
Faculty
Participants
Host and facility
Marketing/ Communication Campaign
Sponsors ($)
Content
Leadership Curriculum
Processes
Leadership Experiences
# and type of participants
Completion rate
Participant satisfaction
New leadership attitudes, knowledge, skills, and behaviors
Increased community awareness and action bias
Better Leaders
Graduates use knowledge and skills obtained through the program to strengthen the community
Community development
1
Is the CLA doing the right things?
2
Is the CLA doing things right?
3
What difference has the CLA made among participants?
4
What difference has the CLA made across the community?
5
What are the ways that community needs can and should be addressed by the CLA?

Source: The Logic Model Guidebook, Lisa Wyatt Knowlton and Cynthia C. Phillips, SAGE Publishing © 2013.

course of the program, as well as the program elements that will lead to these outcomes. Thus, the model is critical in providing guidance for choosing performance indicators to be used for the various types of evaluation, from formative to impact. In addition, the model should set out reasonable expectations for the timeframes expected for successful results to be achieved. These purposes will become apparent for the planning steps detailed in the text that follows.

Figures 2.1 and 2.2 are examples of a generic logic model and then one developed for a specific program.[9] The generic model outlines the basic components of the model from the resources put into the program through the various levels of outcomes. The detailed model, based on the Community Leadership Academy (CLA), demonstrates the details that can be added from each component of the model.[10] In the following chapter, we describe using a logic model to organize evaluation measures.

Developing Evaluations: Some Preliminary Methodological Thoughts

Whether one of the evaluation types described at the beginning of this chapter is desired for any particular program or policy, or whether a comprehensive evaluation strategy incorporating all of them is desired, resolving basic design and data issues is essential. Each evaluation type requires measures of progress, success, or effectiveness to assess. These in turn require that appropriate data are available at critical time points for the program.

This is especially true for impact evaluation. As will be discussed in the later chapters, the research questions posed by impact evaluation require the establishment of a counterfactual, that is, what outcomes the intervention group would have experienced in the absence of the intervention. Establishing the counterfactual generally means a valid comparison group must be available. As will become apparent as the book proceeds, developing rigorous impact evaluations depends on four interrelated factors: how the program itself is designed and implemented; the research design for the evaluation; the availability of a comparison group; and the availability of necessary data. These factors are fully interdependent. For example, program design and data availability often determine whether a valid comparison group is available. The data and comparison group in turn determine what research design and statistical methods can be used to infer the attribution of the outcomes to the intervention.

As discussed in the following section, when evaluations are planned as part of program design and implementation, there are many choices over these factors. In other words, with this type of planning, program administrators and evaluators enhance their opportunities to ensure valid results. When evaluations are implemented retrospectively and separately from program design, these choices become much more limited.

Prospectively Planned and Integrated Program Evaluation

Program implementation and administration can be combined with evaluation in three basic ways. The first is to take a purely retrospective approach—the evaluation is planned and undertaken after the program has been implemented. This approach by definition limits the data available for analysis to secondary sources (e.g., administrative data, clinical records). The second is to plan the evaluation prospectively but still fail to integrate its design and data requirements into program implementation plans. This approach may likewise limit the evaluation to data available from secondary sources, which may not adequately address all the program-specific concerns administrators and other evaluation stakeholders may have. In both cases, an impact evaluation can face obstacles to construction of appropriate comparison groups and use of best available impact estimation methods, as well as potentially suffering from the absence of key data elements and/or less than optimal observation points. As a result, these evaluations may not allow us to answer whether the program achieved its goals and whether the program is likely responsible for the changes observed.

The third approach can be called prospectively planned and integrated evaluation (PPIPE), an example of which is presented in Chapter 9 and discussed further in Chapters 10–13. The PPIPE approach, which this book advocates, is a common-sense approach based on the editors' (Bir and Sheingold) many years of experience with research, evaluation, and policy issues. PPIPE integrates evaluation activities fully with program design and implementation from the very beginning. The PPIPE approach can be consistent with continuous monitoring of a program, feedback mechanisms, and methodologically credible rapid-cycle and final evaluations. It can be based on learning system approaches and continuous quality improvement. Indeed, the Institute of Medicine's report, "Rewarding Provider Performance,"[11] recommended that HHSs pay for performance programs be implemented within an active learning system that allowed monitoring, evaluation, and research:

> Monitoring, evaluation and research functions should not be divorced from program design and implementation or merely appended to pay for performance programs. Rather, their success depends on having a strong learning system that is intrinsic to the design and activities of the program. . . . Conversely, the absence of a scientifically valid, comprehensive, integrated, flexible system—one that facilitated learning from experience—would likely contribute to the failure of a pay for performance program.

The descriptors of most public programs could easily replace the words "pay for performance."

Taking a PPIPE approach can avoid some or all of the common evaluation obstacles. If implementing programs with truly experimental designs is thwarted for any of the many reasons already described, proper planning can enable program implementation in a way that allows for credible quasi-experimental designs. The PPIPE approach requires planning for comparison groups and statistical methods that are compatible, as well as ensuring all data flows are consistent with the chosen methods. The rest of this chapter provides planning tools for the PPIPE approach. Subsequent chapters address the methodologies that can be applied.

Evaluation Framework Elements: Suggested Planning Tools for the Prospectively Planned and Integrated Evaluation Approach

The logic model described above provides a theory of change for a program or policy from the program inputs, to changes in behaviors and finally desired impacts. A framework for establishing a PPIPE evaluation must include the following elements that should follow naturally from a good logic model or results chain:

1. **Clear Description of Program Objectives.** Framing the evaluation design requires a clear statement of program/intervention objectives. All the evaluation components described below will be based on these objectives. In particular, knowing what to measure, and therefore the choice of performance indicators, critically depends on program objectives. Choice of evaluation methods and comparison groups for impact evaluation should also depend on those objectives.

2. **Description of Target Population.** A clear description of the target population is needed to guide the choice of performance indicators and data sources for measurement and assessment at all stages of the evaluation process.

3. **Description of Measurable Performance.** One of the most important evaluation steps is to prospectively establish performance indicators that can demonstrate the impact of the program relative to the established objectives. These indicators, which are established in accordance with the stated objectives and target populations, are structured for use at different stages of the evaluation process. The three types of performance indicators are potentially useful to measure the short, intermediate, and long-term outcomes described in the logic model:

    a. *Measures for Process Evaluation*. These measures can be structured for the formative stages of an evaluation, as well as rapid and continuous monitoring of program progress. Rather than focusing on final or intermediate outcomes, process measures focus on the processes and infrastructure consistent with successful achievement of program objectives. Process measures may be related to recruitment, enrollment, changes in technology, and/or changes in organizational structure. Thus, whereas the outcome indicators discussed below are measured for the target population, process indicators are measured for those delivering the services or otherwise interacting with the target population.

b. *Measures for Outcomes Evaluation.* These measures would reflect changes in the target population that would be expected as a result of exposure to the intervention and that would be strongly related to changes in the final outcome measures. These measures may reflect changes in behaviors or in knowledge and abilities of the target population.

c. *Measures for Impact Evaluation.* The key measures that the program is ultimately intended to affect for the target population are the final outcome measures. They can include clinical, health, economic, or other performance outcomes that are most consistent with the program objectives.

Changes in many final outcome measures may occur slowly, and in some cases either very late in, or after, the program's period of performance. Thus, periodic monitoring of these measures will not necessarily provide useful information regarding the program progress or effectiveness. Where this is the case, intermediate outcome measures should be identified that allow demonstration of progress toward the outcomes desired. These intermediate measures chosen should be linkable to the final outcomes, preferably by a strong evidence base. For example, changes in the number of strokes that result from a blood pressure monitoring program would likely occur well in the future, but reductions in blood pressure readings that can be linked to strokes can be monitored on a timelier basis.

4. **Expectations for Trajectory of Changes in Performance Measures.** These should be based on the logic model and be clear and prospectively established to the extent possible. They should include expected timeframes for improvement in all performance indicators, expected rate of improvement, and mileposts for success or corrective action. While establishing the trajectory of expected improvement in outcomes is necessarily imprecise, it serves a useful purpose in guiding expectations and choosing intervals for data reporting, monitoring, and analysis. Of equal importance, these expectations guide the optimal timing for the program and the evaluation phases. For example, if it is a reasonable assumption that the program would affect the final outcome measures in 3 years, then the impact evaluation cannot be completed for some time beyond that period. If it is necessary for decisions to be made before that time, they would need to be based on the process or outcomes evaluation phases.

5. **Timeframes and Intervals for Data Reporting.** These should be established for the various performance indicators (see Item 3 above) to be consistent with monitoring and evaluation needs throughout the life of the program.

6. **Data Infrastructure and Flows.** This critical aspect of prospectively planned evaluation is to specify in advance and ensure that all data will be available consistent with the indicators specified in Item 3 and the timeframes established in Item 5. Most importantly, the planning must ensure appropriate data are available for implementing the research designs chosen (see Item 7), including data on both program participants and established comparison groups. Data planning should also establish the infrastructure needs to ensure efficient and timely data processing.

7. **Appropriate Research Designs.** Perhaps the most critical step is to determine and specify the research design to be used for the impact evaluation. As described above, decisions on the research design, data availability, and an appropriate comparison group are interdependent. This is why we advocate building evaluation into program design and implementation. In that way, more options available for such decisions improve the evaluation's ability to rigorously demonstrate the program's impact on specified outcomes. In other words, the design should answer the following questions: Did the outcomes improve? For whom? And can some or all improvement be attributed to the program operations themselves?

   The research design description should include an analysis and justification for the statistical methods to be used, data needs, and how comparison groups will be chosen. Any subgroup analyses that are anticipated should be specified in advance, with the rationale for their inclusion. It is important (as described in Chapter 7) to assess potentially differential outcomes across subgroups. For this analysis, an *ex ante* conceptual framework is important to provide a defense against concerns regarding *ex post* "fishing" for results.

8. **Monitoring and Feedback Mechanisms.** These mechanisms need to evaluate data on an ongoing basis and reflect policy changes as needed. Good program management will require systems in place, and staff responsibilities assigned, to enable the information flowing from the steps above to be readily processed and displayed in reports accessible to specific audiences. The ability to produce "dashboards" and other reports that can monitor what is working well and what is not is critical to support decision making about potential program changes. While early monitoring often focuses on process, as described below, adding early outcome results where feasible can be most valuable.

9. **Rapid-Cycle Evaluation.** The special responsibilities of CMMI, as noted, have given rise to a new approach known as rapid-cycle evaluation (RCE). RCE describes methods that provide interim information on a program's progress in both process and outcome measures. *Rapid* denotes frequent assessment of program model effectiveness. *Cycle* denotes real-time data monitoring and mixed methods approaches, to provide regular feedback to participating providers about their performance to support continuous quality improvement.

It is important to be clear that RCE and final impact evaluation should not be considered as alternatives, but rather as different phases within a comprehensive evaluation process—ideally using the same methods and measures in each phase. In this way, key outcome measures can also be assessed at regular intervals to provide interim results—for example, quarterly (rapid cycle) as well as at the end of a program period (outcome and/or impact).

10. **Stopping Rules and Evidence Thresholds.** Occasionally a PPIPE evaluation incorporating robust RCE methods yields strong enough evaluation findings (positive or negative) to argue for terminating the evaluation before its scheduled end date. There is a strong caveat to such a step, however. Critics are likely to assume, in the case of positive preliminary findings, that the program sponsors may be terminating the evaluation to prevent the possibility of negative findings later on. To guard against such criticism, it is critical that the rules for possible early termination—including how the strength of evidence will be determined and statistical rules applied to any such decision—are specified and justified when the original design decisions are made.

Following the general steps above will greatly increase the chances of obtaining robust results within a timeframe useful for decision makers. But no one should count on good and timely results by themselves to ensure that evaluation results will be enough to persuade program administrators, policymakers, and other decision makers to use them. Finding ways to present evaluation results and match them to meet the unique needs of the specific decision-making audiences is also essential. Later chapters address alternative ways of analyzing and presenting results to decision makers, as well as the types of decision-making frameworks real-world program stakeholders use.

SUMMARY

As presented in this chapter, evaluations can be formative, summative, or process evaluations, or any combination of the three. While many challenges may complicate researchers' ability to properly evaluate program impacts, many can be avoided through early planning and implementation of the evaluation. Incorporating evaluation planning and integrating evaluation activities fully with program design and implementation can determine the rigor, statistical validity, and credibility of the evaluation's findings and ensure that all elements necessary for the timely, accurate, and valid monitoring and assessment of program implementation and performance are achieved. A prospectively planned and integrated program evaluation (PPIPE) ensures the evaluation can provide an evidence- and experience-based learning system that supports the design, activities, and goals of the program.

The PPIPE framework provides a context for the later chapters, which contain detailed information and analyses related to its key components. When used in conjunction with the information this primer describes in later chapters, the PPIPE framework provides a useful planning tool/checklist for program managers and their evaluation partners to prospectively build effective evaluation into their program structures. In addition, the PPIPE framework can provide a benchmark for programs already under way that have yet to plan evaluations or that have planned them separately from the program implementation. Comparing what they have available to them in the way of evaluation design and data sources to the 10 points in the PPIPE framework can enable program managers to assess the opportunities they may still have available to improve the rigor and value of their evaluation.

DISCUSSION QUESTIONS

1. What steps need to be taken to ensure that an evaluation is properly prospectively planned and integrated?
2. Apply the PPIPE framework to the expansion of Medicaid coverage to low-income men.
3. Should program planning take into account evaluation design? Why, or why not? If so, how?

NOTES

1. R. Bruce Hutton, and Dennis L. McNeil, "The Impact of Program Evaluation Needs on Research Methodology," *Advances in Consumer Research* [serial online] 8, no. 1 (January 1981):547–52, http://www.acrwebsite.org/volumes/5856/volumes/v08/NA-08.
2. Huey-tsyh Chen, "A Comprehensive Typology for Program Evaluation," *American Journal of Evaluation*, 17, no. 2 (1996): 121–30.
3. Hutton and McNeill, 1981.
4. Ibid.
5. Ibid.
6. Ibid.
7. Paul J. Gertler, Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch, *Impact Evaluation in Practice* (Washington, DC: World Bank, 2011).
8. Ibid.; Centers for Disease Control, Division for Heart Disease and Stroke Prevention, *Evaluation Guide: Developing and Using Logic Models*, https://www.cdc.gov/eval/tools/logic_models/index.html.
9. Both figures are from Lisa Wyatt Knowlton and Cynthia C. Phillips, *The Logic Model Guidebook: Better Strategies for Great Results* (Thousand Oaks, CA: Sage Publications, 2013).
10. Ibid.
11. Institute of Medicine, *Rewarding Provider Performance: Aligning Incentives in Medicare* (Washington, DC: National Academies Press, 2007).