CHAPTER

1

What Is Program Evaluation and Why Is It Needed?

CHAPTER OUTLINE

Program evaluation is the systematic assessment of programs designed to improve social conditions and our individual and collective well-being. Programs are designed to address social problems, but most social problems resist efforts to remedy them. To answer key questions about the performance of such programs, evaluators apply social science research methods to provide answers to stakeholders. To be effective, a social program must correctly diagnose the problem it is intended to address, adopt a feasible design capable of ameliorating the problem, be well implemented in a manner consistent with the design, actually improve the outcomes for the population targeted by the program, and do so at an acceptable cost to society. Different domains of program evaluation address questions related to each of these aspects of social programs using concepts and methods appropriate to those questions.

This book is rooted in the tradition of scientific study of social problems—a tradition that has aspired to improve the quality of social conditions and our physical environment and enhance our individual and collective well-being through the systematic creation and application of knowledge. Although the terms *program evaluation* and *evaluation research* are relatively recent inventions, the activities we will consider under these rubrics are not. They can be traced to the very beginnings of modern science. Three centuries ago, as Cronbach and colleagues (1980) point out, Thomas Hobbes and his contemporaries tried to use numerical measures to assess social conditions and identify the causes of mortality, morbidity, and social disorganization. Since the latter part of the 20th century, the resistance of many social problems to efforts to bring about change for the better and developments in empirical social sciences have combined to make program evaluation an important and commonplace undertaking.

WHAT IS PROGRAM EVALUATION?

Our focus is on social programs, also referred to as social interventions, especially human service programs in such areas as health, education, employment, housing, community development, poverty, criminal justice, and international development. At various times, policymakers, funding organizations, planners, program managers, taxpayers, or program clientele need to distinguish worthwhile social programs from ineffective ones, or perhaps launch new programs or revise existing ones so that the programs may achieve better outcomes. Informing and guiding the relevant stakeholders in their deliberations and decisions about such matters is the work of program evaluation. (Note that throughout this book we use the terms *evaluation*, *program evaluation*, and *evaluation research* interchangeably.)

Although this text emphasizes evaluation of social programs, evaluation research is not restricted to that arena. The broad scope of program evaluation can be seen in the evaluations of the U.S. Government Accountability Office (GAO), which have covered the procurement and testing of military hardware, quality control for drinking water, the maintenance of major highways, the use of hormones to stimulate growth in cattle, and other organized activities far afield from human services. Indeed, the techniques described in this text are useful in virtually all spheres of activity in which issues are raised about the effectiveness of organized social action. For example, the mass communication and advertising industries use essentially the same approaches in developing media programs and marketing products. Political candidates develop their campaigns by evaluating the voter appeal of different strategies. Consumer products are tested for performance, durability, and safety. This list of examples could be extended indefinitely.

To illustrate the evaluation of social programs more concretely, we offer below a few examples of diverse programs with different aims that have been evaluated in various settings and social sectors.

- In 2010, malaria was responsible for 1 million deaths per year worldwide according to the World Health Organization, and in Kenya it was responsible for one quarter of all children's deaths. Bed nets treated with insecticide have been shown to be effective in reducing maternal anemia and infant mortality, but in Kenya fewer than 5% of children and 3% of pregnant women slept under them. In 16 Kenyan health clinics, pregnant women were randomly given an opportunity to obtain bed nets at no cost instead of the regular price. The acquisition and use of bed nets increased by 75%

when they were free compared with the regular cost of 75 cents. In part because of the availability and use of bed nets, deaths attributable to malaria have been reduced by 29% since 2010 (Cohen & Dupas, 2010).

- Since the initiation of federal requirements for monitoring students' proficiency in reading, mathematics, and science as well as graduation rates, the issue of chronically low performing schools has garnered much public attention. In Tennessee some of the lowest performing schools were taken into a special district controlled by the state. Others were placed in special "district-within-districts," known as iZones, and granted greater autonomy and additional resources. In the first 3 years of operation, an evaluation showed that student achievement increased in the iZone schools, but not in the schools taken over by the state, which were run primarily by charter school organizations (Zimmer, Henry, & Kho, 2017).
- Acceptance and commitment theory (ACT) is a treatment program for individuals who engage in aggressive behavior with their domestic partners. Delivered in a group format, ACT targets such problematic characteristics of abusive partners as low tolerance for emotional distress, low empathy for the abused partner, and limited ability to recognize emotional states. An evaluation of ACT compared outcomes for ACT participants with comparable participants in a general support-and-discussion group that met for the same length of time. Outcomes measured 6 months later showed that ACT participants reported less physical and psychological aggression than participants in the discussion group (Zarling, Lawrence, & Marchman, 2015).
- The threat of infectious disease is high in office settings where employees work in close proximity, with implications for absenteeism, productivity, and health care insurance claims. A large company in the American Midwest attempted to reduce these adverse effects by placing hand sanitizer wipes in each office and liquid hand sanitizer dispensers in high-traffic common areas. This intervention was implemented in two of the three office buildings on the company's campus, with the third and largest building held back for comparison purposes. They found that during the 1st year there were 24% fewer health care claims for preventable infectious diseases among the employees in the treated buildings than in the prior year, and no change for the employees in the untreated building. Those employees also had fewer absences from work, and an employee survey revealed increases in the perception of company concern for employee well-being (Arbogast et al., 2016).

These examples illustrate the diversity of social interventions that have been systematically evaluated and the globalization of evaluation research. However, all of them involve one particular evaluation activity: evaluating the effects of programs on relevant outcomes. As we will discuss later, evaluation may also focus on the need for a program; its design, operation, and service delivery; or its efficiency.

WHY IS PROGRAM EVALUATION NEEDED?

Most social programs are well intentioned and take what seem like quite reasonable approaches to improving the problematic situations they address. If that were sufficient to ensure their success, there would be little need for any systematic evaluation of their performance. Unfortunately, good intentions and intuitively plausible interventions do not necessarily lead to better outcomes. Indeed, they can sometimes backfire, with what seem to be promising programs having harmful effects that were not anticipated. For example, the popular Scared Straight program, which spawned a television series that lasted for nine seasons, involved taking juvenile delinquents to see prison conditions and interact with the adult inmates in order to deter crime. However, evaluations of the program found that it actually resulted in increased criminal activity among the participants (Petrosino, Turpin-Petrosino, Hollis-Peel, & Lavenberg, 2013). This example and countless others show that the problems social programs attack are rarely ones easily influenced by efforts to resolve them. They tend to be complex, dynamic, and rooted in entrenched behavior patterns and social conditions resistant to change.

Under these circumstances, there are many ways for intervention programs to come up short. They may be based on an action theory (more about this later) that is not well aligned with the nature or root causes of the problem, or one that assumes an unrealistic process for changing the conditions it addresses. Furthermore, any program with at least some potential to improve the pertinent outcomes must be well enough implemented to achieve that potential. A service that is not delivered or is poorly delivered relative to what is intended has little chance of accomplishing its goals. With an inherently effective intervention strategy that is adequately implemented and then actually has the intended beneficial effects, there can still be issues that keep the program from being a complete success. For example, the program may also have effects in addition to those intended that are not beneficial, that is, adverse side effects. And there is the issue of cost, whether to government and ultimately taxpayers or to private sponsors. A program may produce the intended benefits, but at such high cost that it is not viable or sustainable. Or there may be alternative program strategies that would be equally effective at lower cost.

In short, there are many ways for a program to fail to produce the intended benefits without unanticipated negative side effects, or to do so in a sustainable, cost-effective way. Good intentions and a plausible program concept are not sufficient. If they were, we could be confident that most social programs are effective at delivering the expected benefits without conducting any evaluation of their theories of action, quality of implementation, positive and adverse effects, or benefit-cost relationships. Unfortunately, that is not the world we live in. When programs are evaluated, it is all too common for the results to reveal that they are not effective in producing the intended outcomes. If those outcomes are worth achieving, it is especially important under these circumstances to identify successful programs. But it is equally important to identify the unsuccessful ones so that they may be improved or replaced by better programs. Assessing the effectiveness of social programs and identifying the factors that drive or undermine their effectiveness are the tasks of program evaluation.

Why Systematic Evaluation?

The subtitle of this evaluation text is “A Systematic Approach.” There are many approaches that might be taken to evaluate a social program. We could, for example, simply ask individuals familiar with the program if they think it is a good program. Or, we could rely on the opinions of experts who review a program and render judgment, rather the way sommeliers rate wine. Or, we could assess the status of the recipients on the outcomes the program addresses to see how well they are doing and somehow judge whether that is satisfactory. Although any of these approaches would be informative, none are what we mean by *systematic*. The next section of this chapter will discuss this in more detail, but for now we focus on the challenges any evaluation approach must deal with if it is to produce valid, objective answers to critical questions about the nature and effects of a program. It is those challenges that motivate a systematic approach to evaluation.

One such challenge is the relativity of program effects. With rare exceptions, some program participants will show improvement on the outcomes the program targets, such as less depression, higher academic achievement, obtaining employment, fewer arrests, and the like, depending on the focus of the program. But that does not necessarily mean these gains were caused by participation in the program. Improvement for at least some individuals is quite likely to have occurred anyway in the natural course of events even without the help of the program. Crediting the program with all the improvement participants make will generally overstate the program effects. Indeed, there may be circumstances in which participation in the program results in less gain than recipients would have made otherwise, such as in the Scared Straight example. Thus program effects must be assessed relative to the outcomes expected without program participation, and those are usually difficult to determine.

It follows that program effects are often hard to discern. Take the example of a smoking cessation program. If every participant is a 20-year smoker who has tried unsuccessfully multiple times to quit before joining such a program, and none of them ever smoke again afterward, it is not a great leap to interpret this as largely a program effect. It seems reasonably predictable that all of the participants would not have quit smoking in the absence of the program. But what if 60% start smoking again? Relapse rates are high for addictive behaviors, but could there be a program effect in that high rate? Maybe 70% would start smoking again without the program. Or maybe only 50%. Most program effects are not black or white, but in the gray area where the influence of the program is not obvious.

A direct approach to this ambiguity would be to ask the participants if the program helped them. They will almost certainly have opinions to offer, but they will not be reliable informants about program effects. Those who have done well will likely give exaggerated credit to the program, but it is as much a matter of speculation for them as it is for evaluators to rule out the possibility that they would have done as well without the program. The clearest indication of this inclination for participants to credit the program for their successes is the ready availability of testimonials for virtually every program. Even programs found to be ineffective in rigorous evaluations can generally find participants who did well and will attribute their success to the program. It is simply very difficult for people to accurately account retrospectively for the factors that actually caused their behavior to change.

Alternatively, we might ask the program providers about how effective the program is. The line staff who deliver the services and interact directly with recipients certainly seem to be in a position to provide a good assessment of how well the program is working. Here, however, we encounter the problem of **confirmation bias**: the tendency to see things in ways favoring preexisting beliefs. Consider the medical practitioners in bygone eras who were convinced by the evidence of their own eyes and the wisdom of their clinical judgment that treatments we now know to be harmful, such as bloodletting and mercury therapy, were actually effective. They did not intend to harm their patients, but they believed in those treatments and gave much greater weight in their assessment to patients who recovered than those who did not. Similarly, program providers generally believe the services they provide are beneficial, and confirmation bias nudges them to high awareness of evidence consistent with that belief and to discount contrary evidence.

The approaches to evaluating the performance of a program that may seem most natural and straightforward, therefore, cannot be counted on to provide a valid assessment. If program evaluation is to reach valid conclusions about program performance, systematic methods structured to avoid bias and misrepresentation as much as possible must be used.

SYSTEMATIC PROGRAM EVALUATION

We begin with the definition of program evaluation that guides the orientation of this text and then elaborate on each component of this definition to highlight the major themes we believe are integral to the practice of program evaluation.

Program evaluation is the application of social research methods to systematically investigate the effectiveness of social intervention programs in ways that are adapted to their political and organizational environments and are designed to inform social action to improve social conditions.

One of the pioneers of systematic program evaluation, who developed and refined many of the practices and methods used in the field today, was the first author of this text, Peter H. Rossi. Rossi, who passed away in 2006, was a leading sociologist who served on the faculty of Harvard, the University of Chicago, Johns Hopkins, and the University of Massachusetts–Amherst and conducted research on social problems and evaluated social programs. His vision for systematic program evaluation and some of his contributions to the field are noted in Exhibit 1-A.

Application of Social Research Methods

The concept of evaluation entails, on one hand, a description of the performance of the entity being evaluated and, on the other, some standards or criteria for judging that performance (see Exhibit 1-B). It follows that a central task of the program evaluator is to construct a valid description of program performance in a form that permits comparison with applicable criteria. Failing to describe program performance with a reasonable degree of validity may distort a program's accomplishments, deny it credit for its successes, or overlook shortcomings for which it should be accountable. Moreover, an acceptable description of program performance must be detailed and precise. An unduly vague or equivocal description will make it difficult to determine with confidence whether the performance actually meets the appropriate standard.

Social research methods and the accompanying standards of methodological quality have been developed and refined explicitly for the purpose of constructing sound factual descriptions of social phenomena. In particular, contemporary social science techniques of systematic observation, measurement, sampling, research design, and data analysis represent highly refined procedures for producing valid, reliable, and precise characterizations of social behavior. Social research methods thus provide an especially appropriate approach to the task of describing program performance in ways that will be as credible and defensible as possible.

Regardless of the type of social intervention under study, therefore, evaluators will typically use social research procedures for gathering, analyzing, and interpreting evidence about the performance of a program. This is not to say, however, that we believe that program evaluation must use some particular social research methods or combination of methods,

EXHIBIT 1-A

PETER H. ROSSI: AN EVALUATION CHAMPION AND LEGENDARY EVALUATOR

The major reason why public social programs fail is that effective programs are difficult to design. . . . The major sources of program design failures are: (a) incorrect understanding of the social problem being addressed, (b) interventions that are inappropriate, and (c) faulty implementation of the intervention.

. . . I believe that we can make the following generalization: **The findings of the majority of evaluations purporting to be impact assessments are not credible.**

They are not credible because they are built upon research designs that cannot be safely used for impact assessments. I believe that in most instances, the fatal design defects are not possible to remedy within the time and budget constraints faced by the evaluator.

Source: Rossi (2003).

One example of Peter Rossi's systematic approach to evaluation was his application of sampling theory and social science data collection methods to assess the needs of the homeless in Chicago. He became the first to obtain a credible estimate of the number of homeless individuals in the city, distinguishing residents of shelters and those living on the streets. For counts of shelter residents, his research team visited all the homeless shelters in Chicago for 2 weeks in the fall and 2 weeks in the winter. To collect additional data, he sampled shelters and residents within them for participation in a survey. For the homeless living on the streets, he sampled city blocks and then canvased the homeless individuals on each sampled block between 1 a.m. and 6 a.m. to reduce duplicate counts of shelter residents. The researchers were accompanied by out-of-uniform police officers for their safety, and respondents were paid for their participation in the study. Rossi's research revealed that the homeless population was much smaller than claimed by advocates for the homeless and that it had changed to include more women and minorities than in earlier homeless populations. He found that structural factors, such as the decline of jobs for low-skilled individuals, contributed to homelessness, but it was personal factors like alcoholism and physical health problems that separated the homeless from other extremely poor individuals. This is but one example of his influential contributions to evaluation, which also included evaluations of federal food programs, public welfare programs, and anticrime programs.

Source: Rossi (1990).

EXHIBIT 1-B

THE TWO ARMS OF EVALUATION

Evaluation is the process of determining the merit, worth, and value of things, and evaluations are the products of that process. . . . Evaluation is not the mere accumulation and summarizing of data that are clearly relevant for decision making, although there are still evaluation theorists who take that to be its definition. . . . In all contexts, gathering and analyzing the data that are needed for decision making—difficult though that often is—comprises only one of the two key components in evaluation; absent the other component, and absent a procedure for combining them, we simply lack anything that qualifies as an evaluation. *Consumer Reports* does not just test products and report the test scores; it (i) *rates or ranks* by (ii) *merit or cost-effectiveness*. To get to that kind of conclusion requires an input of something besides data, in the usual sense of that term. The second element is required to get to conclusions about merit or net benefits, and it consists of evaluative premises or standards. . . . A more straightforward approach is just to say that evaluation has two arms, only one of which is engaged in data-gathering. The other arm collects, clarifies, and verifies relevant values and standards.

Source: Scriven (1991, pp. 1, 4–5).

whether quantitative or qualitative, experimental or ethnographic, positivist or naturalist. Nor does this commitment to the methods of social science mean that we think current methods are beyond improvement. Evaluators must often innovate and improvise as they attempt to find ways to gather credible, compelling evidence about social programs. In fact, evaluators have made many novel contributions to methodological development in applied social research in their quest to improve the evidence they can provide about social programs and their effectiveness.

Nor does this view imply that methodological quality is necessarily the most important aspect of an evaluation or that only the highest technical standards, without compromise, are always appropriate. As Carol Weiss (1972) observed long ago, social programs are inherently inhospitable environments for research purposes. The people operating social programs tend to focus attention on providing the services they are expected to provide to the members of the target population specified to receive them. Gathering data is often viewed as a distraction from that central task. The circumstances surrounding specific programs and the issues the evaluator is called on to address frequently compel them to adapt textbook methodological standards, develop innovative methods, and make compromises that allow for the realities of program operations and the time and resources allocated for the evaluation. The challenges to the evaluator are to match the research procedures to the evaluation questions and circumstances as well as possible and, whatever procedures are used, to apply them at the highest standard possible to those questions and circumstances.

The Effectiveness of Social Programs

Social programs are generally undertaken to "do good," that is, to ameliorate social problems or improve social conditions. It follows that it is appropriate for the parties who invest in social programs to hold them accountable for their contribution to the social good. Correspondingly, any evaluation of such programs worthy of the name must evaluate—that is, judge—the quality of a program's performance as it relates to some aspect of its effectiveness in producing social benefits. More specifically, the evaluation of a program generally involves assessing one or more of five domains: (a) the need for the program, (b) its design and theory, (c) its implementation and service delivery, (d) its outcome and impact, and (e) its efficiency (more about these domains later in the chapter).

Adapting to the Political and Organizational Context

Program evaluation is not a cut-and-dried activity like putting up a prefabricated house or checking a student's paper with a computer program that detects plagiarism. Rather, evaluators must tailor the evaluation to the particular program and its circumstances. The specific form and scope of an evaluation depend primarily on its purposes and audience, the nature of the program being evaluated, and, not least, the political and organizational context within which the evaluation is conducted. Here we focus on the last of these factors, the context of the evaluation.

The evaluation plan is generally organized around questions posed about the program by the **evaluation sponsor**, who commissions the evaluation, and other pertinent **stakeholders**: individuals, groups, or organizations with a significant interest in how well a program is working. These questions may be stipulated in specific, fixed terms that allow little flexibility, as in a detailed contract for evaluation services. However, it is not unusual for the initial questions to be vague, overly general, or phrased in program jargon that must be translated for more general consumption. Occasionally, the evaluation questions put forward are essentially pro forma (e.g., is the program effective?) and have not emerged from careful reflection regarding the relevant issues. In such cases, the evaluator must probe thoroughly to determine what the questions mean to the evaluation sponsor and stakeholders.

Equally important are the reasons the questions are being asked, especially the uses that are intended for the answers. An evaluation must provide information that addresses issues that matter for the key stakeholders and communicate it in a form that is usable for their purposes. For example, an evaluation might be designed one way if it is to provide information about the quality of service as feedback to the program director, who will use the results to incrementally improve the program, and quite another way if it is to provide information to a program sponsor, who will use it to decide whether to renew the program's funding.

These assertions assume that an evaluation would not be undertaken unless there was an audience interested in receiving and at least potentially using the findings. Unfortunately, evaluations are sometimes commissioned with little intention of using the findings. For instance, an evaluation may be conducted solely because it is mandated by program funders and then used only to demonstrate compliance with that requirement. Responsible evaluators try to avoid being drawn into such situations of ritualistic evaluation. An early step in planning an evaluation, therefore, is an inquiry into the motivation of the evaluation sponsors, the intended purposes of the evaluation, and the uses to be made of the findings.

As a practical matter, an evaluation must also be tailored to the organizational makeup of the program. In designing an evaluation, the evaluator must take into account such organizational factors as the availability of administrative cooperation and support; the ways in which program files and data are kept and the access permitted to them; the character of the services provided; and the nature, frequency, duration, and location of the contact between the program and its clients. Once the evaluation is under way, modifications may be necessary in the types, quantity, or quality of the data collected as a result of unanticipated practical or political obstacles, changes in the operation of the program, or shifts in the interests of the stakeholders.

Influencing Social Action to Improve Social Conditions

We have emphasized that the role of evaluation is to provide answers to questions about a program that will be useful and will be used. This point is fundamental to evaluation: its purpose is to influence action. An evaluation, therefore, primarily addresses the audiences with the potential to make decisions and take action on the basis of the evaluation results. The evaluation findings may assist in making go/no-go decisions about specific program modifications or, perhaps, about initiation or continuation of entire programs. The evaluation may have direct effects on judgments of a program's value as part of an oversight process that holds the program accountable for results. Or it may have indirect effects in shaping the way program issues are framed and the nature of the debate about them.

Program evaluations may also have social action purposes beyond those of the particular programs being evaluated. What is learned from an evaluation of one program, say, a drug use prevention program at a particular high school, says something about the whole category of similar programs. Many of the parties involved with social interventions must make decisions and take action that relates to types of programs rather than individual programs. A congressional committee may debate the merits of privatizing public education, a state correctional department may consider instituting community-based substance abuse treatment programs, or a philanthropic foundation may deliberate about whether to provide contingent incentives to parents that encourage their children to remain in school. The body of evaluation findings for programs of each of these types is very pertinent to discussions and decisions at this broader level.

One important form of evaluation research is conducted on **demonstration programs**, which are social intervention projects designed and implemented explicitly to test the value of an innovative program concept. In such cases, the findings are significant because of what they reveal about the program concept and how promising it is for broader implementation. Another significant evaluation-related activity is the integration of the findings of multiple evaluations of a particular type of program into a synthesis that can inform policy making and program planning. Whether focused on an individual program or a collection of programs, the common denominator in all evaluation research is that it is intended to be both useful and used, either directly and immediately or as an incremental contribution to a cumulative body of practical knowledge.

THE CENTRAL ROLE OF EVALUATION QUESTIONS

One of the most challenging aspects of evaluation is that there is no one-size-fits-all approach. Every evaluation situation has a different and unique profile of characteristics. A good evaluation design is one that adapts the evaluator's repertoire of approaches, techniques, and concepts to the program circumstances in a way that yields credible and useful answers to the questions that motivate it. The nature of those evaluation questions and the way they are developed and formulated are not only the starting point for any program evaluation but the organizing themes around which the evaluation is structured. In this section we review some of the key features of evaluation questions and the factors that shape them.

The Purpose of the Evaluation

Evaluations are initiated for many reasons. They may be intended to help management improve a program; support advocacy by proponents or critics; gain knowledge about the program's effects; provide input to decisions about the program's funding, structure, or administration; or respond to political pressures. One of the first determinations the evaluator must make to identify the most relevant evaluation questions is the purpose of the evaluation. This is not always a simple matter. A statement of the purposes may accompany the request for an evaluation, but those announced purposes rarely tell the whole story and sometimes are only rhetorical. The evaluator often must dig deeper to determine who wants the evaluation, what they want, and why they want it. There is no cut-and-dried method for doing this, but it is usually best to approach the task the way a journalist would dig out a story. The evaluator can examine source documents, interview key informants with different vantage points, and uncover pertinent history and background. Generally, the purposes of the evaluation will relate mainly to program improvement, accountability, or knowledge generation, but sometimes quite different motivations are in play.

Program Improvement

An evaluation intended to furnish information for guiding program improvement is called a **formative evaluation** (Scriven, 1991) because its purpose is to help form or shape the program to perform better. The audiences for formative evaluations typically are program planners, administrators, oversight boards, or funders with an interest in optimizing the program's effectiveness. The information desired may relate to the need for the program, the program's design, its implementation, its impact, or its costs, but often tends to focus on program operations, service delivery, and take-up of services by the program's target population. The evaluator in this situation will usually work closely with program management and other stakeholders in designing, conducting, and reporting the evaluation. Evaluation for program improvement characteristically emphasizes findings that are timely, concrete, and immediately useful. Correspondingly, the communication between the evaluator and the respective audiences may occur regularly throughout the evaluation and can be relatively informal.

Accountability

The investment of social resources such as taxpayer dollars by human service programs is justified by the presumption that the programs will make beneficial contributions to society. Program managers are thus expected to use resources effectively and efficiently and actually produce the intended benefits. An evaluation conducted to determine whether these expectations are met is called a **summative evaluation** (Scriven, 1991) because its purpose is to render a summary judgment on the program's performance. The findings of summative evaluations are usually intended for decision makers with major roles in program oversight, for example, the funding agency, governing board, legislative committee, political decision makers, or organizational leaders. Such evaluations may influence significant decisions about the continuation of the program, allocation of resources, restructuring, or legislative action. For this reason, they require information that is sufficiently credible under scientific standards to provide a confident basis for action and to withstand criticism aimed at discrediting the results. The evaluator may be expected to function relatively independently in planning, conducting, and reporting the evaluation, with stakeholders providing input but not participating directly in decision making. In these situations, it may be important to avoid premature or careless conclusions, so communication of the evaluation findings may be relatively formal, rely chiefly on written reports, and occur primarily at the end of the evaluation.

Knowledge Generation

Some evaluations are undertaken to describe the nature and effects of an intervention as a contribution to knowledge. For instance, an academic researcher might initiate an evaluation to test whether a program designed on the basis of theory, say, a behavioral nudge to undertake a socially desirable behavior, is workable and effective. Similarly, a government agency or private foundation may mount and evaluate a demonstration program to investigate a new approach to a social problem, which, if successful, could then be implemented more widely. Because evaluations of this sort are intended to make contributions to the social science knowledge base or be a basis for significant program innovation, they are usually conducted using the most rigorous methods feasible. The audience for the findings will include the sponsors of the research as well as a broader audience of interested scholars and policymakers. In these situations, the findings of the evaluation are most likely to be disseminated through scholarly journals, research monographs, conference papers, and other professional outlets.

Hidden Agendas

Sometimes the true purpose of the evaluation, at least for those who initiate it, has little to do with actually obtaining information about the program's performance. Program administrators or boards may launch an evaluation because they believe it will be good for public relations and might impress funders or political decision makers. Occasionally, an evaluation is commissioned to provide a rationale for a decision that has already been made behind the scenes to terminate a program, fire an administrator, or the like. Or the evaluation may be commissioned as a delaying tactic to appease critics and defer difficult decisions.

Virtually all evaluations involve some political maneuvering and public relations, but when these are the principal purposes, the prospective evaluator is presented with a difficult dilemma. The evaluation must either be guided by the political or public relations purposes, which will likely compromise its integrity, or focus on program performance issues that are of little real interest to those commissioning the evaluation and may even be threatening. In either case, the evaluator is well advised to try to avoid such situations.

The Evaluator-Stakeholder Relationship

Every program is necessarily a social structure in which various individuals and groups engage in the roles and activities that constitute the program. In addition, every program is a nexus in a set of political and social relationships among those with involvement or interest in the program, such as relevant decision makers, competing programs, and advocacy groups. The nature of the evaluator's relationship with these and other stakeholders who may participate in the evaluation or have an interest in it will shape the way the evaluation questions are framed. The primary stakeholders potentially influential in this process may include the following:

Decision makers: Persons responsible for deciding whether the program is to be initiated, continued, discontinued, expanded, modified, restructured, or curtailed.

Program sponsors: Individuals with positions of responsibility in public agencies or private organizations that initiate and fund the program; they may overlap with decision makers.

Evaluation sponsors: Individuals in public agencies or private organizations who initiate and fund the evaluation (the evaluation sponsors and program sponsors may be the same).

Target participants: Persons, households, or other units that are intended to receive the intervention or services being evaluated.

Program managers: Personnel responsible for overseeing and administering the intervention program.

Program staff: Personnel responsible for delivering the program services or functioning in supporting roles.

Program competitors: Organizations or groups that compete with the program. For instance, a private organization receiving public funds to operate charter schools will be in competition with public schools also supported by public funds.

Contextual stakeholders: Organizations, groups, and individuals in the environment of a program with interests in what the program is doing or what happens to it (e.g., other agencies or programs, journalists, public officials, advocacy organizations, citizens' groups in the jurisdiction in which the program operates).

Evaluation and research community: Evaluation professionals who read evaluations and review their technical quality and credibility along with researchers who work in areas related to that type of program.

The most influential stakeholder will typically be the evaluation sponsor, the agent that initiates the evaluation, usually provides the funding, and makes decisions about how and when it will be done and who will do it. Various relationships with the evaluation sponsor and other stakeholders are possible and will depend largely on the sponsor's preferences and whatever negotiation takes place with the evaluator. The evaluator's relationship to stakeholders is so influential for shaping the evaluation process that a special vocabulary has arisen to describe the major variants.

In an **independent evaluation**, the evaluator has the primary responsibility for developing the evaluation questions in collaboration with key stakeholders, conducting the evaluation, and disseminating the results. The evaluator may initiate and direct the evaluation quite autonomously, as when a social scientist undertakes an evaluation for purposes of knowledge generation with research funding that leaves the particulars to the researcher's discretion. More often, the independent evaluator is commissioned by a sponsoring agency that stipulates the purposes and nature of the evaluation but leaves it to the evaluator to do the detailed planning and conduct the evaluation. For instance, program funders often commission evaluations by publishing a request for proposals or applications, to which evaluators respond with statements of their capability, proposed design, budget, and time line, as requested. The evaluation sponsor then selects an evaluator from among those responding and establishes a contractual arrangement for the agreed-on work. In such cases, however, the evaluator nonetheless generally confers with a range of stakeholders to give them some influence in shaping the evaluation.

A **participatory or collaborative evaluation** is organized as a team project with the evaluator and representatives of one or more stakeholder groups jointly making decisions about the evaluation and how it is conducted. The participating stakeholders are directly involved in formulating the evaluation questions, and planning, conducting, and analyzing the data collected for the evaluation in collaboration with the evaluator. The evaluator's role might range from project leader or coordinator to that of resource person called on only as needed. Variations on this form of relationship are typical for internal evaluators who are part of the organization whose program is being evaluated. In such cases, the evaluator generally works closely with management in formulating the evaluation questions and planning and conducting the evaluation. One well-known form of participatory evaluation is Patton's (2008) utilization-focused evaluation. Patton's approach emphasizes close collaboration with the individuals who will use the evaluation findings to ensure that it is responsive to their needs and produces information they can and will actually use.

In an **empowerment evaluation**, the evaluator-stakeholder relationship is participatory and collaborative. In addition, however, the evaluator's role includes consultation and facilitation directed toward democratic participation and building the capacities of the participating stakeholders to conduct evaluations on their own, to use the results effectively for advocacy and change, and to take ownership of a program that affects their lives. For instance, some recipients of program services may be asked to take a primary role in planning, setting priorities, collecting information, and interpreting the results of the evaluation. The evaluation process in this arrangement, therefore, is directed not only at producing informative and useful findings but also at enhancing the development and political influence of the participants. As these themes imply, empowerment evaluation most appropriately includes stakeholders who otherwise have little power in the context of the program, usually the program recipients or intended beneficiaries. In their most recent contribution, three pioneers of empowerment evaluation document examples in contexts as diverse as a tobacco prevention program and an organizational transformation initiative that have used this approach (Fetterman, Kaftarian, & Wandersman, 2015).

Criteria for Program Performance

Beginning a study with a set of research questions is customary in the social sciences (often framed as hypotheses). What distinguishes evaluation questions is that they have to do with performance and are associated, at least implicitly, with some criteria by which that performance can be judged. When program managers or evaluation sponsors ask such things as "Are we targeting the right client population?" or "Do our services benefit the recipients?" they are not only asking for a description of the program's performance, they are also asking if that performance is good enough according to some standard or judgment.

One implication of this distinctive feature of evaluation is that good evaluation questions will, when possible, convey the applicable **performance criterion** or standard as well as the performance dimension that is at issue. Thus, evaluation questions may be much like this: "Does the program serve at least 75% of the individuals eligible to receive the services?" (by some explicit eligibility criteria) or "Do the majority of those who receive the employment services get jobs within 30 days of the conclusion of training that they keep at least 3 months?" To be meaningful, there should be some rationale for the standard that is related to the ability of the program to accomplish its overall goal of improving the target social conditions.

The applicable performance criteria may take different forms for various dimensions of program performance (Exhibit 1-C). In some instances, there are established professional standards that are applicable to program performance. This is particularly likely in medical and health programs, in which practice guidelines and managed care standards may be relevant. Perhaps the most common criteria are those based directly on program design, goals, and objectives. In this case, program officials and sponsors identify certain desirable accomplishments as the program aims. Often these statements are not very specific with regard to the nature or level of program performance they represent. One of the goals of a shelter for battered women, for instance, might be to "empower women to take control of their own lives." Although reflecting commendable values, this statement gives no indication of the tangible manifestations of such empowerment that would constitute attainment of this goal. Considerable discussion with stakeholders may be necessary to translate such statements into mutually acceptable terminology that describes the intended outcomes concretely, identifies the observable indicators of those outcomes, and specifies the level of accomplishment that would be considered a success in accomplishing the stated goal.

Some program objectives, on the other hand, may be very specific. These often come in the form of administrative objectives adopted as targets according to past experience, benchmarking against the experience of comparable programs, a judgment of what is reasonable and desirable, or maybe only an informed guess as to what is needed. Examples of administrative objectives may be to complete intake for 90% of the referrals within 30 days, to have 75% of the clients complete the full term of service, to have 85% "good" or "outstanding" ratings on a client satisfaction questionnaire, to provide at least three appropriate services to each person under case management, and the like. There is typically some arbitrariness in these criterion levels. But if they are administratively stipulated, can be established through stakeholder consensus, represent attainable targets for improvement over past practice, or can be supported by evidence of levels associated with positive outcomes, they may be quite serviceable in the formulation of evaluation questions and interpretation of the subsequent findings. However, it is not generally wise for the evaluator to press for specific statements of target performance levels if the program does not have them or cannot readily and confidently develop them.

Establishing a performance criterion can be particularly difficult when the performance dimension in an evaluation question involves outcome or impact issues. Program stakeholders and evaluators alike may have little idea about how much change on an outcome (e.g., frequency of alcohol or drug use) is large enough to have practical significance. In practice, the standard for performance is often set in relation to the outcome expected in the absence of the program and a related judgment about whether the program has improved on that at all. By default, these judgments are often made on the basis of statistical criteria, that is, whether the measured effects are statistically significant. This is a poor practice for reasons that will

EXHIBIT 1-C

MANY CRITERIA MAY BE RELEVANT TO PROGRAM PERFORMANCE

The standards by which program performance may be judged in an evaluation include the following:
The needs or wants of the target population
Stated program goals and objectives
Professional standards
Customary practice; norms for other programs
Legal requirements
Ethical or moral values; social justice, equity
Past performance; historical data
Targets set by program managers
Expert opinion
Preintervention baseline levels for the target population
Conditions expected in the absence of the program (counterfactual)
Cost or relative cost

be more fully examined in Chapter 9. Statistical criteria have no intrinsic relationship to the practical significance of a change on an important outcome and can be misleading. A juvenile delinquency program that is found to have the statistically significant effect of lowering subsequent reoffense rates by 2%, for example, may not make a large enough difference to be judged worthwhile relative to its costs.

THE FIVE DOMAINS OF EVALUATION QUESTIONS AND METHODS

A carefully developed set of **evaluation questions** gives structure to an evaluation, leads to appropriate and thoughtful planning, and serves as a basis for discussions about who is interested in the answers and how they are to be used. Although appropriate evaluation questions will be rather specific to the program to be evaluated, it is useful to recognize that they generally fall into categories according to the program issues they address. Five such domains of evaluation questions can be distinguished:

Need for the program: Questions about the social conditions a program is intended to ameliorate and the need for the program.

Program theory and design: Questions about program conceptualization and design.

Program process: Questions about program operations, implementation, service delivery, and the way recipients experience the program services.

Program impact: Questions about program change in the targeted outcomes and the program's impact on those changes.

Program efficiency: Questions about program cost and cost-effectiveness.

Evaluators have developed concepts and methods for addressing the kinds of questions in each of these categories, and those combinations of questions, concepts, and methods constitute the primary domains of evaluation practice. Below we provide an overview of each of those five domains.

Need for the Program: Needs Assessment

The primary rationale for a social program is to alleviate a social problem. The impetus for a new program to increase adult literacy, for example, is likely to be recognition that a significant proportion of persons in a given population are deficient in reading skills. Similarly, an ongoing program may be justified by the persistence of a social problem: Driver education in high schools receives public support because of the continuing high rates of automobile accidents among adolescent drivers.

One important form of evaluation, therefore, assesses the nature, magnitude, and distribution of a social problem; the extent to which there is a need for intervention; and the implications of these circumstances for the design of the intervention. These diagnostic activities are referred to as **needs assessment** in the evaluation field (Altschuld & Kumar, 2010; Watkins, Meiers, & Visser, 2012) but overlap with what is called social epidemiology and social indicators research in other fields. Critical to the process of conducting a needs assessment is determination of the gap between the current social condition and the condition judged to be acceptable to society or a particular community.

Examples of the kinds of questions addressed by needs assessment, stated in summary form, are as follows:

- What are the nature and magnitude of the problem to be addressed?
- What are the characteristics of the population in need?
- What are the needs of the population? What has created that need?
- What kinds of assistance might address those needs? What outcomes would be desirable?
- What characteristics of the population in need would influence the ability to provide assistance or the way in which it should be provided?

Needs assessment to provide information about the nature of the social condition at issue and the implications for the ways in which it might be effectively addressed is often a first step in planning a new program. Needs assessment may also be appropriate to examine whether an established program is responsive to the current needs of its target population and provide guidance for improvement. Exhibit 1-D provides an example of one of the several approaches that can be taken. Chapter 2 discusses the various aspects of needs assessment in detail.

EXHIBIT 1-D

ASSESSING THE NEEDS OF OLDER CAREGIVERS FOR YOUNG PERSONS INFECTED OR AFFECTED BY HIV OR AIDS

In South Africa, many aspects of the reduction of the incidence of HIV infection and AIDS and management of care for HIV-infected individuals and those with AIDS have been the focus of government interventions. However, the needs of older persons who are the primary caregivers for children or grandchildren affected by HIV or AIDS had not been previously assessed. In one arm of a mixed-methods study, evaluators selected and surveyed individuals, 50 years of age or older who were giving care to younger persons who received HIV- or AIDS-related services from one of seven randomly selected nongovernmental organizations (NGOs) in three of South Africa's nine provinces. In addition to the survey data, the evaluators selected 10 survey respondents for in-depth interviews and 9 key informants who managed government HIV/AIDS interventions or NGO programs.

Quantitative data were collected to assess the extent of the problem of caregiving by older persons, and qualitative data were collected to understand the burden of caregiving on the caregivers and to identify areas of need for formal support. A semi-structured survey instrument was tested, refined, piloted, and then used to assess demographic and household data, health status, knowledge and awareness of HIV and AIDS, caregiving to persons living with the disease, caregiving to children and orphaned grandchildren, and support received from the government and other community institutions. Interview schedules were used to interview a purposive sample of caregivers, government officials, and managers of NGOs.

The evaluators collected data on the challenges and support needs of older caregivers and the gaps in public policy responses to the burden of care on those caregivers. The 305 respondents were 91% older women with a mean age of 66 years. Results highlighted that caregiving was largely femininized, and a majority of the caregivers (59%) relied on informal support from NGOs and family members. Lack of formal support was identified across all three provinces. The study was used to formulate a policy framework to inform the design and implementation of policy and programmatic responses aimed at supporting the caregivers.

Source: Adapted from Petros (2011).

Assessment of Program Theory and Design

Given a recognized problem and need for intervention, another domain for evaluation involves questions about the design of the program or intervention that is expected to address that need. The conceptualization and operational plan of a program must reflect valid assumptions about the nature of the problem and represent a feasible approach to reducing the gap between current and acceptable levels of the problematic condition. This program plan may not be written out in detail, but exists nonetheless as a shared conceptualization among the principal stakeholders. The critical part of program design consists of assumptions and expectations about how the program should operate in order to have the intended effects and is referred to as the program theory or theory of action. If this theory is faulty, the intervention will fail no matter how elegantly it is conceived or how well it is implemented.

Examples of questions that may guide an **assessment of program theory and design** in summary form are the following:

- What outcomes does the program intend to affect, and how do they relate to the nature of the problem or conditions the program aims to change?
- What is the theory of action that supports the expectation that the program can have the intended effects on the targeted outcomes?
- Is the program directed to an appropriate population, and does it incorporate procedures capable of recruiting and sustaining their participation in the program?
- What services does the program intend to provide, and is there a plausible rationale for the expectation that they will be effective?
- What delivery systems for the services are to be used, and are they aligned with the nature and circumstances of the target population?
- How will the program be resourced, organized, and staffed, and does that scheme provide an adequate platform for recruiting and serving the target population?

This type of assessment involves, first, describing the program theory in explicit and detailed form, often in the form of a logic model or a theory of behavioral or social change rooted in social science. Logic models are generally organized around the inputs required for a program, the actions or activities to be undertaken, the outputs from those activities, and the immediate, intermediate, and ultimate outcomes the program aims to influence (Knowlton & Phillips, 2013). Programs designed around social science concepts are often drawn from theories of behavioral change, such as outsider theory that begins with dissatisfaction with one's current state and continues through anticipation of the benefits of changing behavior to the adoption of new behavior (Pawson, 2013). Once the program theory is formulated, various approaches are used to examine how reasonable, feasible, ethical, and otherwise appropriate it is. The sponsors of this form of evaluation are generally funding agencies or other decision makers attempting to launch a new program. Exhibit 1-E provides an example and Chapter 3 offers further discussion of program theory and design as well as the ways in which it can be evaluated.

Assessment of Program Process

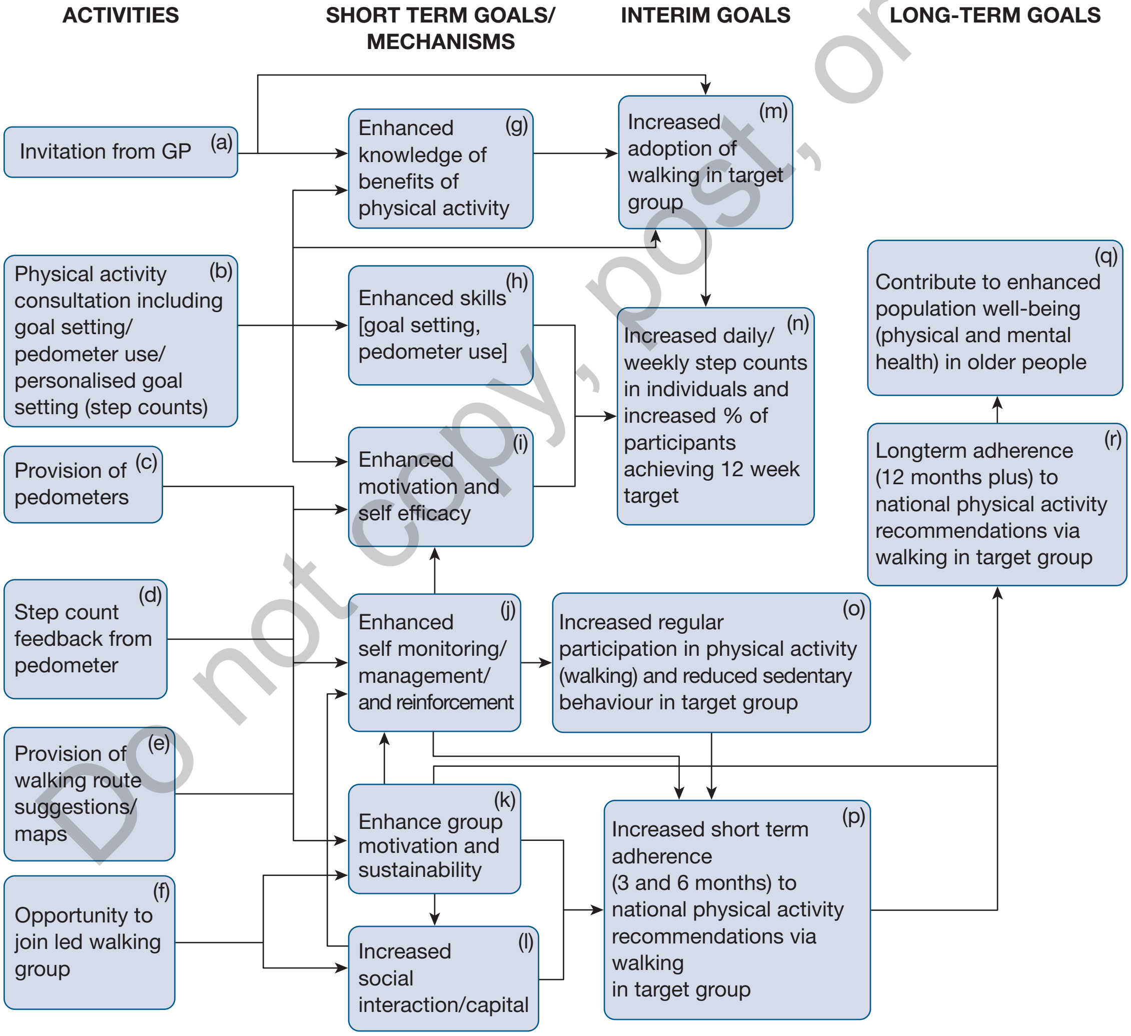
Given a plausible theory about how to intervene to ameliorate an accurately diagnosed social problem, a program must still be implemented well to have a reasonable chance of actually improving the situation. It is not unusual to find that programs are not implemented and executed according to their intended designs. A program may be poorly managed, compromised by political interference, or designed in ways that are impossible to carry out. Sometimes appropriate personnel are not available, facilities or resources are inadequate, or program staff lack motivation, expertise, or training. Possibly the intended program participants do not exist in the numbers required, cannot be identified precisely, or are difficult to engage.

EXHIBIT 1-E

ASSESSING THE PROGRAM THEORY FOR A PHYSICAL ACTIVITY INTERVENTION

Research indicates that physical activity can improve mental well-being, help with weight maintenance, and reduce the risk for chronic diseases such as diabetes. Despite such evidence, it was reported in 2011 that 67% of women and 55% of men in Scotland did not reach the minimum level of activity needed to attain such health benefits. As a result, an intervention known as West End Walkers 65+ (WEW65+) was developed in Scotland to increase walking and reduce sedentary behavior in adults older than 65 years. The design of the intervention relied heavily on empirically supported theories underlying behavioral change and prior activity interventions that had demonstrated effectiveness. Before implementation, the intervention design and underlying theory, depicted below, was assessed as part of a pilot and feasibility assessment of the program.

Theory for WEW65+ intervention


ACTIVITIES
SHORT TERM GOALS/ MECHANISMS
INTERIM GOALS
LONG-TERM GOALS
Invitation from GP (a)
Physical activity consultation including goal setting/ pedometer use/ personalised goal setting (step counts) (b)
Provision of pedometers (c)
Step count feedback from pedometer (d)
Provision of walking route suggestions/ maps (e)
Opportunity to join led walking group (f)
Enhanced knowledge of benefits of physical activity (g)
Enhanced skills [goal setting, pedometer use] (h)
Enhanced motivation and self efficacy (i)
Enhanced self monitoring/ management/ and reinforcement (j)
Enhance group motivation and sustainability (k)
Increased social interaction/capital (l)
Increased adoption of walking in target group (m)
Increased daily/ weekly step counts in individuals and increased % of participants achieving 12 week target (n)
Increased regular participation in physical activity (walking) and reduced sedentary behaviour in target group (o)
Increased short term adherence (3 and 6 months) to national physical activity recommendations via walking in target group (p)
Contribute to enhanced population well-being (physical and mental health) in older people (q)
Longterm adherence (12 months plus) to national physical activity recommendations via walking in target group (r)

While assessing the program theory, the evaluators examined the underlying assumptions and the triggers for the psychological mechanisms expected lead to achieving the outcomes goals set for the intervention. They confirmed the reasonableness of assumptions such as the focus on an older population of adults, the appropriateness of walking as a sufficient physical activity to enhance health outcomes and reduce sedentariness, and the likelihood that information provided in a clinical setting to influence attitudes and behaviors. They also noted the addition of a program activity based on previously tested behavioral theory—a physical activity consultation to enhance the participants knowledge of the benefits of walking and enhance their motivation and self-efficacy—to the intervention design.

Source: Adapted from Blamey, Macmillan, Fitzsimons, Shaw, and Mutrie (2013).

A basic and widely used form of evaluation, **assessment of program process**, evaluates the fidelity and quality of a program's implementation. Such process assessments may be done as a freestanding evaluation of the activities and operations of the program, commonly referred to as a process evaluation or an implementation assessment. When the process evaluation is an ongoing function that occurs regularly, it will usually be referred to as **program monitoring**. A program monitoring function may also include information about the status of program participants on targeted outcomes after they have completed the program and thus also include **outcome monitoring**. Process evaluation investigates how well the program is operating. It might examine how consistent the services actually delivered are with the design for the program, whether services are delivered to appropriate recipients, how well service delivery is organized, the effectiveness of program management, the use of program resources, the well-being of participants after receipt of program services, and other such matters (Exhibit 1-F provides an example). Examples of the kind of evaluation questions that guide process evaluations are:

- Are the intended services being delivered to the intended persons?
- Are administrative and service objectives being met?
- Are there eligible but unserved persons the program is not reaching?
- Once beginning service, do sufficient numbers of participants complete service?
- Are the participants satisfied with the services?
- Are the participants doing well in the ways intended after receipt of the program services?
- Are administrative, organizational, and personnel functions managed well?

Process evaluation is the most common form of program evaluation. It is used both as a stand-alone evaluation and in conjunction with impact assessment as part of a more comprehensive evaluation. As a stand-alone evaluation, it yields quality assurance information, assessing the extent to which a program is implemented as intended and operating according to the standards established for it. When the program model used is one of established

EXHIBIT 1-F

ASSESSING THE IMPLEMENTATION FIDELITY AND PROCESS QUALITY OF A YOUTH VIOLENCE PREVENTION PROGRAM

After a pilot study proved successful, a community-level violence prevention and positive youth development program, Youth Empowerment Solutions (YES), was rolled out, and a process evaluation was conducted to measure implementation fidelity and quality of delivery. The process evaluation was conducted in 12 middle and elementary schools in Flint, Michigan, and surrounding Genesee County. Data were collected from 25 YES groups from 12 schools over 4 years. Four groups were eliminated from the analysis because of incomplete data. Data collection covered the measurement of implementation fidelity, the dose delivered to participants, the dose received from participants, and program quality. The evaluators summarized multiple methods adopted to measure each component in the table below.

Source	Participant–Teacher Interaction: Observation	Core Content Components: Teacher Self-Report	Sessions Offered: School Records and Teacher Self-Report	Attendance: School Records and Teacher Self-Report	Participant Engagement: Participant Self-Report	Participant Satisfaction: Participant Self-Report	Teacher Training: Study Records	Quality Summary: Score Calculated
Fidelity	X	X						
Dose delivered			X	X				
Dose received					X	X		
Program quality							X	X

Results measuring implementation fidelity found that although teachers scored well on their adherence to program protocol, there was large variation in the proportion of curriculum core content components covered by each group, ranging from 8% to 86%. Additionally, dose delivered also varied widely, with the number of sessions offered ranging from 7 to 46. Finally, despite high participant satisfaction, with 84% of students stating that they would recommend the program to others, there were large variations in the quality summary scores of program delivery. Overall, the evaluation findings reinforced the program, including enhancements to the curriculum, teacher training, and technical assistance. The evaluators noted the limitations of collecting self-reported data, but they also acknowledged the value of collecting data from multiple sources, allowing the triangulation of findings.

Source: Adapted from Morrel-Samuels et al. (2017).

effectiveness, establishing that the program is well implemented can be presumptive evidence that the expected outcomes are produced as well. When the program is new, a process evaluation provides valuable feedback to administrators and other stakeholders about progress implementing the program design. From a management perspective, process evaluation provides the feedback that allows a program to be managed for high performance, and the associated data collection and reporting of key indicators may be institutionalized in the form of a data dashboard to provide routine, ongoing feedback on key performance indicators.

In its other common application, process evaluation is an indispensable adjunct to impact assessment. The information about a program's effects on its target outcomes that evaluations of impact provide is incomplete and ambiguous without knowledge of the program activities and services that produced those outcomes. When no impact is found, process evaluation has significant diagnostic value, indicating whether this was because of implementation failure, that is, the intended services were not provided hence the expected benefits could not have occurred, or theory failure, that is, the program was implemented as intended but failed to produce the expected effects. Process evaluation and program monitoring are described in more detail in Chapter 4, and outcome monitoring is described in Chapter 5.

Effectiveness of the Program: Impact Evaluation

The effectiveness of a social program is gauged by the change it produces in outcomes that represent the intended improvements in the social conditions it addresses. The ability of a program to have that impact will depend in large part on whether it adequately operationalizes and implements an effective theory of action grounded in an understanding of the social conditions in which it intervenes. **Impact evaluation** asks whether the desired outcomes were actually affected and whether the changes included unintended side effects. Examples of evaluation questions that might be addressed by impact evaluation include:

- Are the outcome goals and objectives of the program being achieved?
- Are the trends in outcomes moving in the desired direction?
- Does the program have beneficial effects on the recipients and what are those effects?
- Are there any adverse effects on the recipients, and what are they?
- Are some recipients affected for better or worse than others, and who are they?
- Is the problem or situation the program addresses made better? How much better?

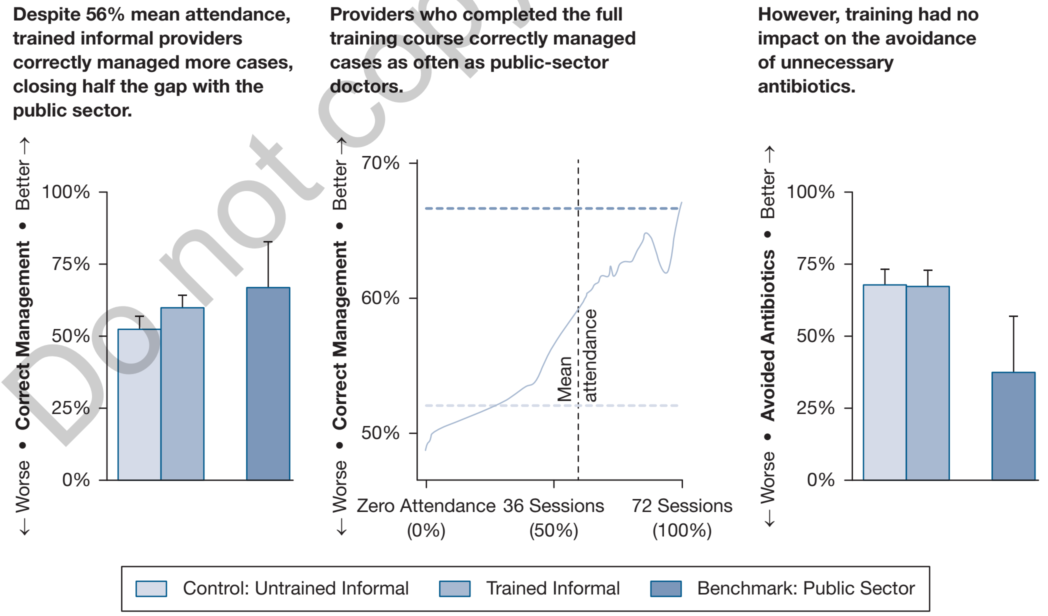
The major difficulty in assessing the impact of a program is that the desired outcomes can usually also be influenced by factors unrelated to the program. Accordingly, impact assessment involves producing an estimate of the *net effects* of a program—the changes brought about by the intervention above and beyond those resulting from other processes and events affecting the targeted social conditions. To conduct an impact assessment, the evaluator must thus design a study capable of establishing the status of program recipients on relevant outcome measures and also estimate what their status would have been had they not received the intervention. Much of the complexity of impact assessment is associated with obtaining a valid estimate of the latter, known as the *counterfactual* because it describes a condition contrary to what actually happened to program recipients (Exhibit 1-G presents an example of an impact evaluation).

EXHIBIT 1-G

EVALUATING THE EFFECTS OF TRAINING INFORMAL HEALTH CARE PROVIDERS IN INDIA

In many countries in the developing world, health care providers without formal medical training account for a large proportion of primary health care visits. Despite legal prohibitions in rural India, informal providers, who are estimated to exceed the number of trained physicians, provide up to three fourths of primary health care visits. Medical associations in India have taken the position that training informal providers may legitimize illegal practices and worsen public health outcomes, but there is little credible evidence on the benefits or adverse side effects of training informal providers. Because of the severe shortage of trained health care providers, an intervention to train informal health care providers was designed as stopgap measure to improve health care while reform of health care regulations and the public health care system was undertaken. The intervention took place in the Indian state of West Bengal and trained informal health care providers in 72 sessions over a period of 9 months on multiple topics, including basic medical conditions, triage, and the avoidance of harmful practices.

A randomized design was used to evaluate the impact of the training program. A sample of 304 providers who volunteered for the training was randomly split into treatment and control groups, the latter of which was offered the training program after the evaluation was complete. Daylong clinical observations that assessed the clinical practices of the providers and their treatment of unannounced standardized patients who were trained to present specific health conditions to the health care providers, were employed to test each participant on his or her delivery of treatment and utilization of skills taught in the training. The researchers withheld information about which group, treatment or control, the health care providers were in from the test patients. The researchers found that the training increased rates of correct case management by 14%, but the training had no effect on the use of unnecessary medicines and antibiotics. Overall, the results suggested that the intervention could offer an effective short-term strategy to improve health care provision. The graphic below provides a summary of the research results:


Despite 56% mean attendance, trained informal providers correctly managed more cases, closing half the gap with the public sector.

Providers who completed the full training course correctly managed cases as often as public-sector doctors.

However, training had no impact on the avoidance of unnecessary antibiotics.

The evaluators raised concerns about the failure of the training to reduce prescriptions of unnecessary medications, even though it had been explicitly included in the training. They noted that many of the informal providers made a profit on the sale of prescriptions and stated, "We believe these null results are directly tied to the revenue model of informal providers."

Source: Adapted from Das, Chowdhury, Hussam, and Banerjee (2016).

Determining when an impact assessment is appropriate and what evaluation design to use presents considerable challenges to the evaluator. Evaluation sponsors often believe they need an impact evaluation, and indeed, it is the only way to determine if the program is bringing about the intended changes. However, an impact assessment can be demanding of expertise, time, and resources and may be difficult to set up properly within the constraints of routine program operation. If the need for information about effects on outcomes is sufficient to justify an impact assessment, there is still a question of whether the program circumstances are suitable for conducting such an evaluation. For instance, it makes little sense to establish the impact of a program that is not well structured or cannot be adequately described. Impact assessment, therefore, is most appropriate for mature, stable programs with well-defined program models and a clear intention to use the results to justify the effort required. Impact assessment is also often appropriate for demonstration projects or pilots of programs that are under consideration for widespread adoption. Chapters 6 to 8 discuss impact assessment and the various ways in which it can be designed and conducted.

Cost Analysis and Efficiency Assessment

Finding that a program has positive effects on the intended outcomes is often insufficient for assessing its social value. Resources for social programs are limited, so their accomplishments must also be judged against their costs. The first requirement for evaluations assessing costs is to describe the specific costs incurred in operating a program. Although many programs have expenditure records, the actual costs of operating a program may include donated items, volunteer time, and opportunity costs (costs associated with spending time on the program rather than other uses of the time by leaders, staff, and participants). A careful description of the full costs of a program is referred to as a **cost analysis**. Beyond describing the costs needed to operate a program, an **efficiency assessment** takes account of the relationship between a program's costs and its effectiveness. Efficiency assessments may take the form of **cost-benefit analysis** or **cost-effectiveness analysis**, asking, respectively, whether a program produces sufficient benefits in relation to its costs and whether other interventions or delivery systems can produce the benefits at a lower cost. Examples of evaluation questions that might guide an efficiency assessment are as follows:

- What are the actual total costs of operating the program, and who pays those costs?
- Are resources used efficiently without waste or excess?
- Is the cost reasonable in relation to the magnitude or monetary value of the benefits?
- Would alternative approaches yield equivalent benefits at less cost?

EXHIBIT 1-H

ASSESSING THE COST-EFFECTIVENESS OF SUPPORTED EMPLOYMENT FOR INDIVIDUALS WITH AUTISM IN ENGLAND

In England, autism spectrum conditions affect approximately 1.1% of the population, and the costs of supporting adults with autism spectrum conditions is estimated to be £25 billion. Given that adults with autism experience difficulties in finding and retaining employment, and the employment rate for adults with autism is estimated to be 15%, the evaluators set out to estimate the cost-effectiveness of supported employment in comparison with standard care or day services.

The authors drew the data on program effectiveness from a prior evaluation, which found that a supported employment program specifically for individuals with autism in the United Kingdom increased employment and job retention in a follow-up study 7 to 8 years after the program was initiated. The program assessed the clients, supported them in obtaining jobs, supported them in coping with the requirements for maintaining employment, educated employers, and advised coworkers and supervisors on how to avoid or handle any problems. For the main analysis, the evaluators used cost data from a study of the unit costs for supported employment services and day services for adults with mental health problems.

TABLE 1 Results of main analysis: mean costs, number of weeks in employment and QALYs of supported employment and standard care per adult with autism seeking employment, over the time horizon of the analysis (17 months of intervention + 8 years follow-up)

	Mean total intervention cost over 17 months	Mean total day service cost over 8-year follow-up (incurred by unemployed only)	Mean total cost	Mean number of weeks in employment	Mean number of QALYs
Supported employment	£5044	£4193	£9237	136	5.42
Standard care (day services)	£2742	£5893	£8635	102	5.31
Difference	£2302	–£1700	£602	34	0.11
Cost-effectiveness	ICER of supported employment versus standard care: £I8/extra week in employment; £5600/QALY				

QALY: quality-adjusted life year; ICER: incremental cost-effectiveness ratio.

Note that numbers have been rounded to the nearest £ (costs), to the nearest integer (weeks in employment) and to the nearest second decimal digit (QALYs).

The incremental cost-effectiveness analysis, or the cost of an extra week of employment, was £18, which led the authors to determine that supported employment programs for adults with autism were cost effective. The authors concluded, "Although the initial costs of such schemes are higher that standard care, these reduce over time, and ultimately supported employment results not only in individual gains in social integration and well-being but also in reductions of the economic burden to health and social services, the Exchequer and wider society."

Source: Adapted from Mavranezouli et al. (2014).

Efficiency assessment can be tricky and arguable because it requires making assumptions about the dollar value of program-related activities and, sometimes, imputing monetary value to program outcomes, both beneficial and adverse, that are difficult to represent with a dollar value. Nevertheless, such estimates are often germane for informing decisions about allocation of resources and identification of the program models that produce the strongest results with a given amount of funding. In certain cases, a descriptive cost analysis by itself may provide salient information to guide decisions about program adoption or consideration that involve fewer assumptions than efficiency assessments.

Like impact assessment, efficiency assessment is most appropriate for mature, stable programs with well-structured program models. This form of evaluation builds on process and impact assessment. A program must be well implemented and produce the desired effects before questions of how efficiently it accomplishes that become especially relevant. Given the specialized expertise required to conduct efficiency assessments, it is also apparent that it should be undertaken only when there is a clear need and identified use for the information. With the high level of concern about program costs in many contexts, however, this may not be an unusual circumstance. Chapter 10 discusses cost and efficiency assessment methods in more detail.

The Interplay Among the Evaluation Domains

As is apparent in the descriptions above of the issues that motivate the different domains of evaluation questions, they reflect a general logic about what constitutes an effective program. That logic says that a program must correctly diagnose and understand the problem or conditions it aims to improve, be designed around a feasible plan for addressing the problem that is based on a valid theory about how the intended changes can be brought about, and operationalize that design in the way it is implemented and sustained. Those qualities should position the program to be effective, that is, to have a beneficial impact on the respective outcomes for the population targeted by the program. Being effective, however, does not necessarily mean being efficient. To be efficient, the program must achieve its effects at an acceptable cost to its sponsors and funders, and at a cost that compares favorably with other means of attaining the same effects.

There is a parallel logic for evaluators attempting to assess these various aspects of a program. Each family of questions draws on or makes assumptions about the answers to the prior questions. A program's theory and design, for instance, cannot be adequately assessed without some knowledge of the nature of the need the program is intended to address. If a program addresses lack of economic resources, the appropriate program concepts and the evaluation questions will be different than if the program addresses drunken driving. Moreover, the most appropriate criteria for judging program design and theory is how responsive it is to the nature of the need and the circumstances of those in need. When an evaluation of a program's theory and design are undertaken in the absence of a prior needs assessment, the evaluator must make assumptions about the extent to which the program design reflects the actual needs and circumstances of the target population to be served. There may be good reason to have confidence in those assumptions, but that will not always be the case.

Similarly, the central questions about program process are about whether the program operations and service delivery are consistent with the program theory and design; that is, whether the program *as intended* has actually been implemented. This means that the criteria for assessing the quality of the implementation are based, at least in part, on how the program is intended to function as specified by its basic conceptualization and design. The evaluator assessing program process must therefore be aware of the nature of the intended implementation, perhaps from a prior assessment of the program theory and design, but more often by reviewing program documents, talking with key stakeholders, and the like. The quality of implementation for a program to feed the homeless through a soup kitchen cannot be assessed without knowing the aims of the program with regard to the population of homeless individuals targeted, the manner in which they are to be reached, the nature of the nutritional support to be provided, the number of individuals to be served, and other such specifics about the expectations and plans for the program.

Questions about program impact, in turn, are most meaningful and interpretable if the program is well implemented. Program services that are not actually delivered, are not fully or adequately delivered, or are not the intended services cannot generally be expected to produce the desired effects on the conditions the program is expected to impact. Evaluators call it **implementation failure** when the effects are null or weak because the program activities assumed necessary to bring about the desired improvements did not actually occur as intended. But a program may be well implemented and yet fail to achieve the desired impact because the program design and theory embodied in the corresponding program activities are faulty. When the program conceptualization and design are not capable of generating the desired outcomes no matter how well implemented, evaluators interpret the lack of impact as **theory failure**.

The results of an impact evaluation that does not find meaningful effects on the intended outcomes, therefore, are difficult to interpret when the program is not well implemented. The poor implementation may well explain the limited impact, and attaining and sustaining adequate implementation is a challenge for many programs. But it does not follow that better implementation would produce better outcomes; implementation failure and theory failure cannot be distinguished in that situation. Strong implementation, in contrast, allows the evaluator to draw inferences about the validity of the program theory, or lack thereof, according to whether the expected impacts occur. It is advisable, therefore, for the impact evaluator to obtain good information about program implementation along with the impact data.

Evaluation questions relating to program cost and efficiency also draw much of their significance from answers to prior evaluation questions. In particular, a program must have at least minimal impact on its intended outcomes before questions about the efficiency of attaining that impact become relevant to decisions about the program. If there are no program effects, there is little for an efficiency evaluation to say except that any cost is too much.

Needs assessments, assessments of program theory and design, assessments of program process, impact evaluations, and cost analysis and efficiency assessments can all be conducted as stand-alone evaluation studies, and the questions addressed in each case will be meaningful in many program contexts. As we have shown, however, there is an interplay among these evaluation domains such that information about the issues addressed in each have implications for the questions, answers, and interpretations in other domains. Some of this can be thought of in relation to the life cycle of a program, with assessments of need, program theory, and program process ideally feeding successively into the planning and initial implementation of a new program. When full implementation is attained, impact evaluation can then test the expectation that this sequence has resulted in a program that has beneficial effects for its target population. If so, an efficiency assessment can guide consideration of whether the cost of achieving those benefits is acceptable. In the rough-and-tumble world of social programs, however, the need for actionable information from an evaluation will not always hew to this logic, and evaluations centered on any of the domains may be appropriate at different stages in the life cycle of a program.

Most of the remainder of this text is devoted to further describing the nature of the issues and methods associated with each of the five evaluation domains and their interrelationships.

SUMMARY

- Program evaluation focuses on social programs, especially human service programs, but the concepts and methods are broadly applicable to any organized social action.
- Most social programs are well intended and take reasonable approaches to improving the social conditions they address, but that is not sufficient to ensure they are effective; systematic evaluation is needed to objectively assess their performance.
- Program evaluation involves the application of social research methods to systematically investigate the performance of social intervention programs and inform social action.
- Evaluation has two distinct but closely related components, a description of performance and standards or criteria for judging that performance.
- Most evaluations are undertaken for one of three reasons: program improvement, accountability, or knowledge generation.
- The evaluation of a program involves answering questions about the program that generally fall into one or more of five domains: (a) the need for the program, (b) its theory and design, (c) its implementation and service delivery, (d) its outcome and impact, and (e) its costs. Each domain is characterized by distinctive questions along with concepts and methods appropriate for addressing those questions.
- Although program evaluations fall into one of these five domains, any particular evaluation involves working with key stakeholders to adapt the evaluation to its political and organizational context.
- Ultimately, evaluation is undertaken to support decision making and influence action, usually for the specific program that is being evaluated, but evaluations may also inform broader understanding and policy for a type of program.

KEY CONCEPTS

Assessment of program process	21
Assessment of program theory and design	19
Confirmation bias	5
Cost analysis	25
Cost-benefit analysis	25
Cost-effectiveness analysis	25
Demonstration program	10
Efficiency assessment	25
Empowerment evaluation	14
Evaluation questions	16
Evaluation sponsor	9
Formative evaluation	11
Impact evaluation	23
Implementation failure	28
Independent evaluation	13
Needs assessment	17
Outcome monitoring	21
Participatory or collaborative evaluation	14
Performance criterion	15
Program evaluation	6
Program monitoring	21
Social research methods	6
Stakeholders	9
Summative evaluation	11
Theory failure	28

CRITICAL THINKING/DISCUSSION QUESTIONS

1. Explain the four different reasons evaluations are conducted. How does the reason an evaluation is undertaken change how the evaluation is conducted?
2. Explain what is meant by systematic evaluation and discuss what is necessary to conduct an evaluation in a systematic way.
3. There are five domains of evaluation questions. Describe each of the five domains and discuss the purpose of each. Provide examples of questions from each of the five domains.

APPLICATION EXERCISES

1. At the beginning of the chapter the authors provide a few examples of social interventions that have been evaluated. Locate a report of an evaluation of a social intervention and prepare a brief (3- to 5-minute) summary of the social intervention that was evaluated and the evaluation that was conducted.
2. This chapter discusses the role of stakeholders, which are individuals, groups, or organizations with a significant interest in how well a program is working. Think of a social program you are familiar with. Make a list of all of the possible stakeholders for that program. How could their interest in the program be the same? How could they differ? Which stakeholders do you believe are most important to engage in the evaluation process and why?