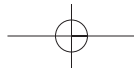


CHAPTER 1

KEY CONCEPTS AND ISSUES IN PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

Introduction	3
Integrating Program Evaluation and Performance Measurement	4
Connecting Evaluation and Performance Management	5
The Practice of Program Evaluation: The Art and Craft of Fitting Round Pegs Into Square Holes	8
A Typical Program Evaluation: Assessing the Neighbourhood Integrated Service Team Program	10
Implementation Concerns	11
The Evaluation	12
Connecting the NIST Evaluation to This Book	13



2 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

What Is a Program?	15
Key Concepts in Program Evaluation	15
Key Evaluation Questions	17
Formative and Summative Program Evaluations	21
<i>Ex Ante and Ex Post Program Evaluations</i>	22
Analyzing Cause and Effect Linkages in Program Evaluations	23
The Process of Conducting a Program Evaluation	24
General Steps in Conducting a Program Evaluation	25
The Evaluation Assessment	25
The Evaluation Study	32
Summary	34
Discussion Questions	35
References	36

INTRODUCTION ●

Program evaluation is a rich and varied combination of theory and practice. It is widely used in public, nonprofit, and private sector organizations to create information for planning, designing, implementing, and assessing the results of our efforts to address and solve problems using policies and programs. **Evaluation** can be viewed as a structured process that creates and synthesizes information intended to reduce the level of uncertainty for stakeholders about a given program or policy. It is intended to answer questions or test hypotheses, the results of which are then incorporated into the information bases used by those who have a stake in the program or policy.

This book will introduce a broad range of evaluation approaches and practices, reflecting the richness of the field. An important, but not exclusive theme of this textbook is evaluating the effectiveness of programs and policies, that is, constructing ways of providing defensible information to stakeholders as they assess whether and how a program accomplished its intended outcomes.

As you read this chapter, you will notice words and phrases in bold. These bolded terms are defined in a glossary at the end of the book. These terms are intended to be your reference guides as you learn or review the language of evaluation. Because this chapter is introductory, it is also appropriate to define a number of terms in the text that will help you get some sense of the “lay of the land” in the field of evaluation.

The richness of the evaluation field is reflected in the diversity of its methods. At one end of the spectrum, students of evaluation will encounter **randomized experiments** where people have been randomly assigned to a group that receives a program that is being evaluated, and others have been randomly assigned to a control group that does not get the program. Comparisons of the two groups are usually intended to estimate the incremental effects of programs. Although these are relatively rare in the practice of program evaluation and there is some controversy around making them the **benchmark** for sound evaluations, they are still often considered as exemplars of “good” evaluations (Scriven, in progress, unpublished).

More frequently, program evaluators do not have the resources, time, or control over program design or implementation situations to conduct experiments. In some cases, an experimental design may not even be the most appropriate for the evaluation at hand. A typical scenario is to be asked to evaluate a program that has already been implemented, with no real ways to create **control groups** and usually no baseline (preprogram) data to construct before-after comparisons. Often, measurement of program outcomes is challenging—there may be no data available and resources available to collect information are scarce.

4 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

Alternatively, data may exist (program records would be a typical situation) but closer scrutiny of these data indicate that they measure program characteristics that only partly overlap with the key questions that need to be addressed in the evaluation. Using these data can raise substantial questions about their **validity**.

Integrating Program Evaluation and Performance Measurement

Evaluation as a field has been transformed in the last 15 years by the broad-based movement in public and nonprofit organizations to construct and implement systems that measure program and organizational performance. Often, governments or boards of directors have embraced the idea that increased accountability is a good thing and have mandated performance measurement. Measuring performance is often accompanied by requirements to report performance results for programs.

Performance measurement is controversial among evaluation experts—some advocate that the profession embrace performance measurement (Bernstein, 1999) while others are skeptical (Perrin, 1998). A skeptic's view of the performance measurement enterprise might characterize performance measurement this way:

Performance measurement is not really a part of the evaluation field. It is a tool that managers (not evaluators) use. Unlike program evaluation, which can call upon a substantial methodological repertoire and requires the expertise of professional evaluators, performance measurement is straightforward: program **objectives** and corresponding outcomes are identified; measures are found to track outcomes, and data are gathered which permit managers to monitor program performance. Because managers are usually expected to play a key role in measuring and reporting performance, performance measurement is really just an aspect of organizational management.

This skeptic's view has been exaggerated to make the point that some evaluators would not see a place for performance measurement in a textbook on program evaluation. But this textbook shows how sound performance measurement, regardless of who does it, depends on an understanding of program evaluation principles and practices. Evaluators who are involved in developing and implementing performance measurement systems for

programs or organizations typically encounter similar problems to program evaluators. A scarcity of resources often means that key program outcomes that require specific data collection efforts are either not measured or are measured with data that may or may not be intended for that purpose. Questions of the validity of **performance measures** are important, as are the limitations to the uses of performance data.

Rather than seeing performance measurement as a quasi-independent enterprise, we *integrate* performance measurement into evaluation by grounding it in the same tools and methods that are essential to assess program processes and effectiveness. Thus, program **logic models** (Chapter 2), **research design** (Chapter 3), and **measurement** (Chapter 4) are important for both **program evaluation** and performance measurement. After laying the foundations for program evaluation, we turn to performance measurement as an outgrowth of our understanding of program evaluation (Chapters 8, 9, and 10).

We see performance measurement approaches as complementary to program evaluation, and not a replacement for evaluations. Analysts in the evaluation field (Newcomer, 1997; Mayne 2001) have generally recognized this complementarity, but in some jurisdictions, efforts to embrace performance measurement have eclipsed program evaluation (McDavid, 2001). We see an important need to balance these two approaches, and our approach in this textbook is to show how they can be combined.

Connecting Evaluation and Performance Management

Both program evaluation and performance measurement are increasingly seen as ways of contributing information that informs **performance management** decisions. Performance management, which is sometimes called **results-based management**, has emerged as an organizational management approach that depends on performance measurement.

Increasingly, there is an expectation that managers will be able to participate in evaluating their own programs, and will also be involved in developing, implementing, and reporting the results of performance measurement. Information from program evaluations and performance measurement systems is expected to play an important role in the way managers manage their programs. Changes to improve program operations and effectiveness are expected to be driven by evidence of how well programs are doing in relation to stated objectives.

Canadian and American governments at the federal, provincial (or state), and local levels are increasingly emphasizing the importance of accountability

6 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

for **program outcomes**. Central agencies including the U.S. Federal Office of Management and Budget [OMB] and the General Accounting Office [GAO] and the Canadian Federal Treasury Board of Canada) as well as state and provincial finance departments and auditors are developing policies and articulating expectations that shape the ways program managers are expected to be able to inform their administrative superiors and other stakeholders outside the organization about what they are doing and how well they are doing it.

In the United States, the OMB publishes reports that serve as policies/guidelines for evaluations across federal departments and agencies. An example is the “Program Evaluation” report that outlines criteria for assessing program effectiveness (Office of Management and Budget, 2004). It is worth noting that the OMB’s view is that randomized control treatment designs are the best for assessing whether and how programs achieved their intended outcomes. In Canada, publications such as “Evaluation Policy” (Treasury Board of Canada Secretariat, 2001) emphasize the importance of creating, implementing, and using an integrated approach for program evaluation and performance measurement and reporting.

The **performance management cycle** includes an iterative planning–implementation–evaluation program adjustments process in which program evaluation and performance measurement play important roles as ways of providing information to decision makers who are engaged in managing organizations to achieve results.

In this book, we will use the performance management cycle as a framework within which evaluation activities can be situated for managers in public sector and nonprofit organizations. Figure 1.1 shows how organizations integrate strategic planning, program and policy design, implementation and evaluation into a cycle. Although this example is taken from a Canadian jurisdiction (Auditor General of British Columbia and Deputy Ministers’ Council, 1996), the terminology and the look of the framework is similar to others that have been adopted by many North American, European, and Australasian jurisdictions.

The five stages in the performance management cycle begin and end with formulating clear objectives for organizations and, hence, programs and policies. “Effective strategies” includes program design, and “aligned management systems” incorporates implementation of programs.

“Performance measurement and reporting” includes both program evaluation and performance measurement, and is expected to contribute to “real consequences” for programs. Among these consequences include a range of possibilities from program adjustments to elections. All can be thought of as parts of the accountability phase of the performance management cycle.

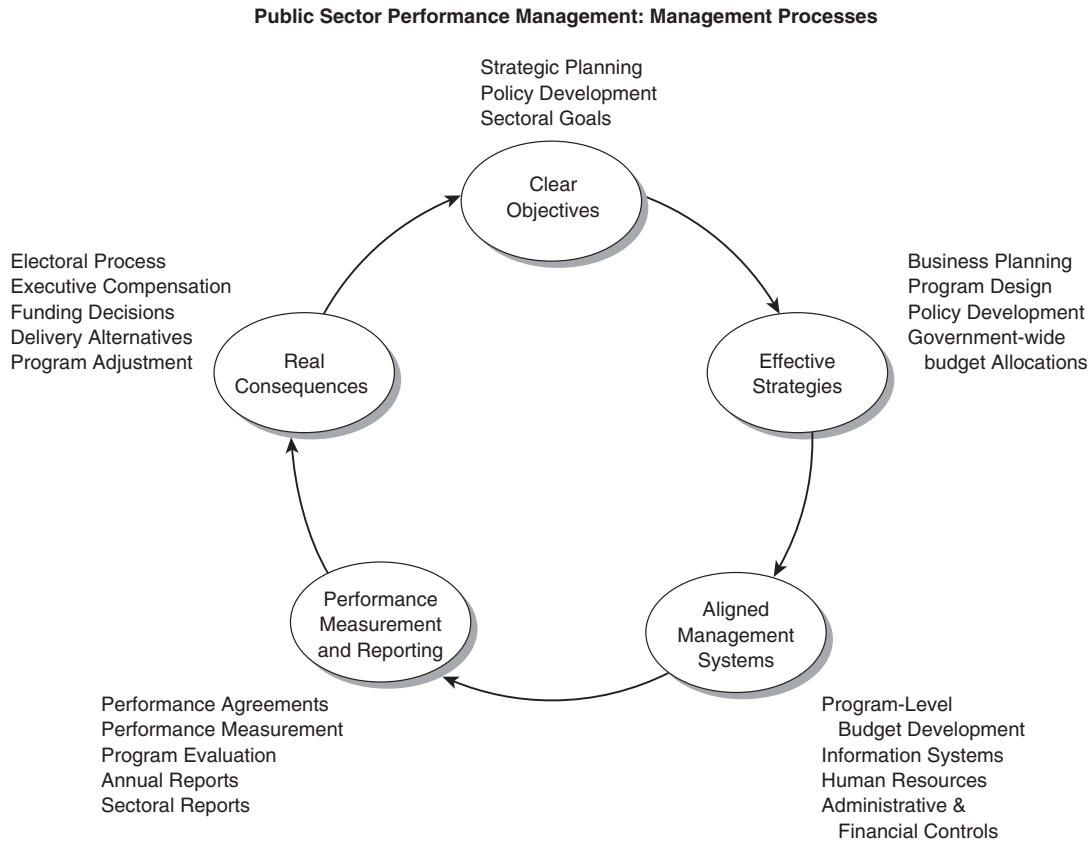


Figure 1.1 Performance Management Cycle

SOURCE: Adapted from Auditor General of British Columbia and Deputy Ministers' Council. April, 1996. *Enhancing Accountability for performance: A framework and an implementation plan*.

Finally, objectives are revisited, and the evidence from earlier phases in the cycle is among the inputs that result in “new” or, at least, revised objectives—usually through a strategic planning process.

In this book the performance management cycle illustrated in Figure 11 is used as a framework for organizing different evaluation topics and showing how the analytical approaches covered in key chapters map onto the performance management cycle. Performance measurement and reporting are the main focus of this textbook and include both program evaluation and performance measurement. Chapters 1, 2 (**logic modelling**), 3 (**research designs**), 4 (**measurement**), and 5 (**qualitative methods**) serve as foundations for both program evaluation and performance measurement.

8 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

Chapters 8 (introduction to performance measurement), 9 (designing and implementing performance measurement systems), and 10 (using performance measures) elaborate performance measurement.

Needs assessments (Chapter 6) build on topics covered in Chapter 4 (measurement) and can occur in several of the phases in the cycle: setting clear objectives, designing effective strategies, and measuring and reporting performance. As well, **cost-benefit analysis** and **cost-effectiveness analysis** (Chapter 7) build on topics in Chapter 3 (research designs) and can be conducted as we design programs (the effective strategies phase) or as we evaluate their outcomes (the performance measurement and reporting phase).

Finally, the relationships between organizational management and evaluation activities (Chapter 11) are key to understanding how performance management and evaluation are linked. Chapter 12 (the nature and practice of **professional judgment**) emphasizes that the roles of managers and evaluators depend on developing and exercising sound professional judgment.

The Practice of Program Evaluation: The Art and Craft of Fitting Round Pegs Into Square Holes

One of the principles underlying this book, and is referred to repeatedly, is the importance of exercising professional judgment as program evaluations are designed, executed, and acted on. The methodological tools we learn, and the pluses and minuses of applying each in our practice, are often intended for applications that are less constrained in time, money, and other resources than are typical of evaluations. One way to look at the fit between the methods we learn and the situations in which they are applied is to think of trying to fit round pegs into square holes. Even if our pegs fit, they often do not fully meet the criteria specified for their application. As evaluators, we need to learn to adapt the tools we know, to the uniqueness of each evaluation setting. In some situations, we find that no approach we know fits the circumstances, so we must improvise.

Our tools are indispensable—they help us to construct useful and defensible evaluations, but like craftspersons or artisans, we ultimately create a structure that combines what our tools can shape with what our own experience, beliefs, values, and expectations furnish and display.

The mix of technique and professional judgment will vary with each evaluation. In some, where causality is a key issue and we have the resources and the control needed to construct an experimental or perhaps **quasi-experimental** research design, we will be able to rely on well-understood methods, which the field of program evaluation shares with social science

disciplines. Even here, evaluators exercise professional judgment. There are no program evaluations that can be done without the evaluator's own experiences, values, beliefs, and expectations playing an important role.

In many situations, program evaluators are expected to "make do." We might be asked to conduct an evaluation after the program has been in place for some time, in circumstances where control groups are not feasible, and resource constraints limit the kinds of data we can gather. Or, we are confronted by a situation where the evaluation design that we had developed in consultation with stakeholders, is undermined by the implementation process. Fitzgerald and Rasheed (1998) describe an evaluation of a program intended to increase paternal involvement in inner-city families where the father does not share custody of the children. The evaluation design started out as a randomized control and treatment experiment, but quickly evolved in ways that made that design unfeasible.

As we shall see, this kind of situation is not intractable. But it demands from us the exercise of professional judgment, and a self-conscious recognition that whatever conclusions and recommendations we produce, they are flavored by what we, as evaluators, bring to the project. Fitzgerald and Rasheed (1998) salvaged the evaluation by including qualitative data collection methods to develop an understanding of how the program actually worked for the participants at the three implementation sites. Although their approach did not meet the standards that they had in mind when they began, they were able to adjust their expectations, take advantage of a **mix of methods** available to them, and produce credible recommendations.

It is tempting, particularly in this latter kind of situation, to conclude that we are not really doing program evaluations, but some other form of "review." Some would argue that real program evaluations are more "pure," and that the absence of some minimum level of methodological sophistication disqualifies what we do from even being considered program evaluation.

But such a stance, although it has some appeal for those who chiefly value methodological sophistication and elegance, is difficult to defend. Drawing some line between "real" and "pseudo" program evaluations is arbitrary. Historically in our profession, there was a time when experimental methods were considered to be the *sine qua non* of evaluations. During the latter part of the 1960s and the first part of the 1970s, experimental methods were applied to evaluating social programs—often with ambiguous conclusions while still being costly (Basilevsky & Hum, 1984).

Now, there is no one dominant view of "correct" evaluation methods. Indeed, qualitative evaluation methods were borne out of a strong reaction to the insular and sometimes remote evaluations produced by social experimenters. Qualitative evaluators like Michael Patton (Patton, 1997) eschew

10 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

much of the methodological armamentarium of their predecessors, and point out that if we want our work to be used, we need to conduct evaluations in ways that encourage the users to take ownership of the conclusions and recommendations. The upshot of this diversity in how we define good evaluations is that drawing a line between real and pseudo evaluations presupposes we agree on one continuum of methods—and we simply do not.

The stance taken in this book, and reflected in the contents of the chapters, is that our program evaluation practice is rich and very diverse. Key to understanding *all* evaluation practice is accepting that no matter how sophisticated your designs, measures, and other methods are, you *will* exercise professional judgment in your work. It is true that the nature and the consequences of such judgments will differ from one evaluation to the next, but fundamentally, we are *all* in the same boat. In this book, we will see where professional judgment is exercised in the evaluation process and begin to learn how to make defensible judgments.

Some readers might have concluded by now that we are condoning an “anything goes” attitude. Readers will discover, instead, that in preparing this book, we have taken a “structured” approach to evaluations that relies on understanding the tools that have been developed in the profession, and applying them in ways that maximize (within the constraints that exist) the defensibility of what we produce.

Program evaluation clients often expect evaluators to come up with ways of telling whether the program achieved its objectives, despite the difficulties of constructing an evaluation design that meets conventional standards to assess the cause and effect relationships between the program and its outcomes. The following case summary illustrates one way that one program evaluator responded to the challenge of conducting an evaluation with limited resources, while addressing questions that we might assume would require more sophisticated research designs. It also illustrates some of the features of the practice of program evaluation.

● A TYPICAL PROGRAM EVALUATION: ASSESSING THE NEIGHBOURHOOD INTEGRATED SERVICE TEAM PROGRAM

In the summer of 1995, a west coast city implemented a Neighbourhood Integrated Service Team (NIST) **program**. NIST was intended as a way to get major city departments involved in neighborhood-level communications and problem solving. A key objective of the program was to improve

cross-department service delivery by making services more responsive to community needs. Related to this objective was a second one: to strengthen partnerships with the community and increase community involvement in problem solving.

The program was a response to concerns that city departments were not doing a good job of coordinating their work, particularly for problems that crossed department responsibilities. The existing “stovepipe” model of service delivery did not work for problems like the “Carolina Street House.”

Citizens in the Mount Pleasant area of Vancouver had spent several frustrating years trying to get the city to control a problem house on Carolina Street. Within a 1-year period alone, neighbors noted that the police had attended the house 157 times, while the fire department had been called 43 times. Property use inspectors had attended the house on a regular basis, as had environmental health officers. In total, over a 3-year period, it was estimated that the city had spent more than CAD\$300,000 responding to citizen complaints related to this property (Talarico, 1999).

The City Manager’s Office reviewed this problem in 1994 and determined that each city department involved had responded appropriately within the scope of its mandate. Where the system had broken down was its failure to facilitate effective communications and collaboration among departments. NIST was intended to address this problem, and deal with situations like Carolina Street before they became expensive and politically embarrassing.

The Neighbourhood Integrated Service Teams were committees of representatives from all eight of the major city departments. The city was divided into 16 neighborhoods, based on historical and city planning criteria, and a NIST committee was formed for each neighborhood.

The committees met on a monthly basis to share information and identify possible problems, and between meetings, members were encouraged to contact their counterparts in other departments as the need arose. With the City Manager’s Office initially pushing NIST, the program was implemented within a year of its start date.

Implementation Concerns

Although the program seemed to be the right solution to the problem that had prompted its creation, concern surfaced around how well it was actually working. Existing city departments continued to operate as separate hierarchies, in spite of the NIST committees that had been formed.

In some areas of the city, the committees did not appear to be very active, and committee members expressed frustration at the lack of continued

12 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

leadership from the City Manager's Office. Although a coordinator had been hired, the position did not carry the authority of a senior manager.

A key concern was whether the program was making a difference: Had service delivery improved and was the community more involved in problem solving? Although the city was receiving inquiries from communities elsewhere about NIST, it could not point to any systematic evidence that the program was achieving its intended objectives.

The Evaluation

In early 1998, the Deputy City Manager commissioned an evaluation of NIST. Since she had been principally responsible for designing and implementing NIST, she wanted an independent view of the program—she would be the client for the evaluation, but the study would be conducted by an independent contractor.

The terms of reference for the evaluation focused in part on whether the program was, in fact, fully implemented: How well were the 16 NIST committees actually working? A related evaluation issue was learning from the experiences of the committees that were working well, so that their practices could be transferred to other committees that needed help.

Although the evaluation did not focus primarily on whether the objectives of the program had been achieved, the Deputy City Manager wanted the contractor to look at this question, as information was being gathered on the strengths and weaknesses of the NIST committees and the work that they did.

The contractor selected to do this evaluation had a limited budget: her time, access to city employees, use of a city vehicle, and an office in city hall. She opted for a qualitative approach to do the study. She would sample and interview persons from four key stakeholder groups: NIST committee members, department staffs, city council, and the community.

She used a combination of individual interviews and focus groups to contact 48 NIST team members, 24 departmental staff (three from each of the eight departments involved in NIST), four members of city council, and 24 representatives from community groups that were active in city neighborhoods.

Using interview questions that were intended to get at the principal evaluation issues, she recorded (using written notes and, in some cases, tape recordings for focus groups) responses, observations, and her own reflections on the information she was gathering.

Data analysis involved content-analyzing interview notes, identifying common ideas in the material she had recorded, and organizing all the

information into themes. Four main categories of themes emerged: areas where the program was reported to be functioning well, areas where there was room for improvement, stakeholder recommendations, and “other” themes. Each of these four areas was subdivided further to assist in the analysis.

Because the evaluation had solicited the views of four key stakeholder groups, the similarities and differences among their views were important. As it turned out, there was a high level of agreement across stakeholders—most identified similar strengths and weaknesses of NIST and offered similar suggestions for making the program work better.

A total of six recommendations came from the evaluation, the key ones focused on ways of better integrating NIST into the city departments. Stakeholders generally felt that although NIST was a good thing and was making a difference, it was not clear how team members were expected to balance their accountability to NIST and to their home departments.

NIST needed to be reflected in department business plans, acknowledging its continued role in city service delivery, and NIST needed stronger leadership to advocate the program within city departments.

Since this evaluation was completed, the Deputy City Manager has been appointed as the City Manager, and her commitment to NIST is reflected in her policies and initiatives. The NIST program has since been recognized for its innovative approach to community service delivery, winning a United Nations Award for Innovation in the Public Service.¹

In addition, the city is leading a partnership with other levels of government to implement a multi-organizational **strategy** using NIST-like mechanisms to tackle the homelessness, crime, and drug problems in one neighborhood that has been the single most difficult challenge for social service agencies, the police department, and other criminal justice agencies (Bakvis & Juillet, 2004).

Connecting the NIST Evaluation to This Book

The development of this program and its evaluation are typical of many in public and nonprofit organizations. In fact, NIST came into being in response to a politically visible problem in this city—a fairly typical situation when we look at the **program rationale**. When NIST was put into place, the main concern was dealing with the problem of the Carolina Street house and others like it. Little attention was paid to how the program would be evaluated. The evaluation was grounded in specific concerns of a senior manager who wanted answers to questions about NIST that were being raised by key

14 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

stakeholders. She had a general idea of what the problems were, but wanted an independent evaluation to either confirm them or indicate where and what the real problems were.

The NIST evaluation is typical of many that evaluation practitioners will encounter. It was expected to answer several important questions for the Deputy City Manager, and do so within the guidelines established by the client. The evaluation was constrained by both time and money; it was not possible, for example, to conduct community surveys to complement other lines of data collected. Nor was it possible to compare NIST to other, non-NIST communities. Other noteworthy points are as follows:

- The evaluation relied on **triangulating** evidence from different points of view with respect to the program, and using these perspectives to help answer the questions that motivated the study.
- Data collection and analysis relied on methods that are generally well-understood and are widely used by other program evaluators. In this case, the evaluator relied on qualitative data collection and analysis methods—principally because they were the most appropriate ways to gather credible information that addressed the evaluation questions.
- The recommendations were based on the analysis and conclusions, and were intended to be used to improve the program. There was no “threat” that the evaluation results might be used to cancel the program. In fact, as mentioned, the program has since been recognized internationally for its innovative approach to community problem solving.
- The evaluation, and the circumstances prompting it, are typical. The evaluator operated in a setting where her options were constrained. She developed a methodology that was defensible, given the situation, and produced a report and recommendations that were seen to be credible and useful.
- The evaluator used her own professional judgment throughout the evaluation process. Methods decisions, data collection, interpretation of findings, conclusions, and recommendations were all informed by her judgment. There was no template or formula to design and conduct this evaluation. Instead, there were methodological tools that could be applied by an evaluator who had learned her craft and was prepared to creatively tackle this project.

Each of these (and other) points will be discussed and elaborated in the other chapters of this textbook. Fundamentally, program evaluation is about

gathering information that is intended to answer questions that program managers and other stakeholders have about a program. Program evaluations are always affected by organizational and political factors and are a balance between methods and professional judgment.

The NIST evaluation illustrates one example of how evaluations are actually done. Your own experience and practice will offer many additional examples (both positive and otherwise) of how evaluations get done. In this book, we will blend together important methodological concerns—ways of designing and conducting defensible and credible evaluations—with the practical concerns facing evaluators, managers, and other stakeholders as they balance evaluation requirements and organizational realities.

WHAT IS A PROGRAM? ●

A program can be thought of as a group of related activities that is intended to achieve one or several related objectives. Programs are means-ends relationships that are designed and implemented purposively. They can vary a great deal in scale. For example, a nonprofit agency serving seniors in the community might have a volunteer program to make periodic calls to persons who are disabled or otherwise frail and living alone. Alternatively, a department of social services might have an income assistance program serving clients across an entire province or state. Likewise, programs can be structured simply: a training program might just have classroom sessions for its clients; or be complex: an addiction treatment program might have a broad range of activities from public advertising, through intake and treatment, to referral, and follow-up. In Chapter 2, we look at the structure of programs in depth and apply an **open systems approach** to describe and model programs.

KEY CONCEPTS IN PROGRAM EVALUATION ●

One of the key questions that many program evaluations are expected to address can be worded as follows

To what extent, if any, did the program achieve its intended objectives?

Usually, we assume that the program in question is “aimed” at some intended objective(s). Figure 1.2 offers a picture of this expectation.

The program has been depicted in a “box,” which serves as a conceptual boundary between the program and the **program environment**. The

16 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT



Figure 1.2 Linking Programs and Intended Objectives

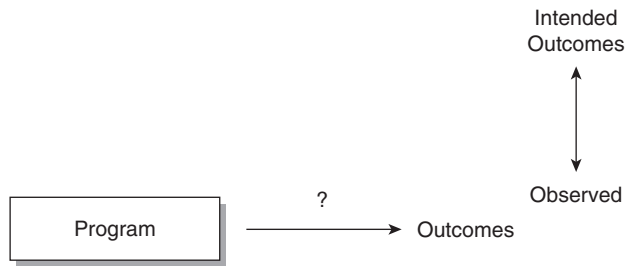


Figure 1.3 The Two Program Effectiveness Questions Involved in Most Evaluations

intended objectives, which we can think of as statements of the **program's intended outcomes**, are shown as occurring *outside* the program itself, that is, the intended outcomes are results intended to make a difference outside of the program itself.

The arrow connecting the program and its intended outcomes is a key part of most program evaluations. It shows that the program is intended to *cause* the outcomes. We can restate the “objectives achievement” question in words that are a central part of most program evaluations.

Was the program effective (in achieving its intended outcomes)?

Assessing **program effectiveness** is the most common reason we conduct program evaluations. We want to know whether, and to what extent, the program's actual results are consistent with the outcomes we expected. In fact, there are *two* evaluation issues related to program effectiveness. Figure 1.3 separates these two issues, so it is clear what each means.

The causal link between the program and its outcomes has been modified in two ways: intended outcomes has been replaced by the **program's observed outcomes** (what we actually observe when we do the evaluation), and a question mark (?) has been placed over the causal arrow.

We need to restate our original question about achieving *intended* objectives to say instead

To what extent, if at all, did the program achieve the observed outcomes?

Notice that we have focused the question on what we *actually observe* in conducting the evaluation, and that the (?) above the causal arrow now raises the key question of whether the program (or possibly something else) caused the outcomes we observe. In other words, we have introduced the **attribution** question, that is, the extent to which *our program* was the *cause* of the outcomes we observed in doing the evaluation. Alternatively, were there factors in the *environment* of the program that caused the observed outcomes?

In Chapter 3 we examine the attribution question in some depth and refer to it repeatedly throughout this book. As we will see, it is often challenging to address this question convincingly, given the constraints within which program evaluators work.

Figure 1.3 also raises a second evaluation question.

To what extent, if at all, are the observed outcomes consistent with the intended outcomes?

Here, we are comparing what we find to what the program was expected to accomplish. Notice that answering that question *does not* tell us whether *the program* was responsible for the *observed* or *intended* outcomes.

Sometimes, evaluators or persons in organizations doing performance measurement do not distinguish the attribution question from the “objectives achievement” question. In implementing performance measures, for example, managers or analysts spend a lot of effort developing measures of intended outcomes. When performance data are analyzed, the key issue is often whether the measures are consistent with intended outcomes. In other words, do the patterns in observed outcomes correspond to the trends or levels that were predicted for that program for that year (or whatever time-frame is specified)? If benchmarks were specified, did the observed outcomes meet the benchmarks? The attribution question (Were the observed outcomes caused by the program?) is not usually addressed, although some analysts with experience in both program evaluation and performance measurement are beginning to look at this issue (Mayne, 2001).

KEY EVALUATION QUESTIONS •

The previous discussion focused on one of the key questions that program evaluations are expected to answer, namely, whether the program was successful in achieving its intended outcomes. Aside from the question of program effectiveness, there are a number of other questions that evaluations

18 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

can address. They are listed below, together with a short explanation of each question.

Was the program efficient? This question needs to be unpacked—it is really two questions expressed in the same words, since it relates to both **technical efficiency** and **economic efficiency**. The question of whether the program was technically efficient refers to whether the ratios of program inputs (usually money) to program outputs (work done) were acceptable in comparison to specified **benchmarks**. Typically, evaluations that focus on **technical efficiency** or **productivity** are concerned with costs per unit of work done or service delivered (an example might be the cost per client served in a meals-on-wheels program). Sometimes, externally developed benchmarks are available to compare observed efficiency ratios to established “standards.” Alternatively, technical efficiency ratios might be compared over time for one program, creating a program baseline against which future ratios could be compared.

In research conducted by the Local Government Institute in the School of Public Administration at the University of Victoria, the technical efficiency of local government services is measured and compared using national surveys. An example of this research is comparing the cost per household served for residential solid waste collectors in Canadian local governments (McDavid, 2001).

Economic efficiency, discussed in Chapter 7, focuses on the benefits of a program compared to its economic costs. Costs and benefits are expressed in monetary terms, and a number of issues having to do with attribution, measurement, and valuing benefits and costs need to be addressed if cost-benefit analysis is used.

Was the program cost-effective? We examine this question, as well, in Chapter 7. Briefly, **cost-effectiveness** is determined by comparing program costs to program outcomes. Ratios of cost per unit of outcome are among the features of this approach to evaluating programs. An example of a cost-effectiveness ratio in a burglary reduction program would be the cost per percent of burglaries reduced in a neighborhood or a community.

Was the program appropriate? The question addresses the technologies and structures incorporated into a program. Was the program constructed logically, that is, did the structure “make sense,” given the intended outcomes? Given the circumstances, did the program designers (often, the program managers) do the best job possible in selecting and organizing the “knowledge base” available to achieve the intended outcomes?

What is the rationale for the program? Given the intended objectives of the program, is the program still relevant to the mission, goals, and objectives of the government or agency in which it is embedded? How well does the program “fit” current and emerging priorities and policies? Questions about program rationale can sometimes relate to needs assessments. Asking whether a program is still relevant can be answered in part by conducting a needs assessment.

Was the program adequate? Given the scale of the problem or condition that the program was expected to address, is the program large enough to do the job? Program adequacy is a question of the “fit” between program resources (and, hence, outcomes) and the scale of the problem embedded in or implicit in the **program objectives**.

Effective and efficient programs may or may not be adequate. An example might be a healthy babies program operating out of a neighborhood family center and does a good job of improving the birth weights and post-natal health of newborns in that neighborhood, but does not visibly affect the overall rate of underweight births in the whole city.

Anticipating the adequacy of a program is also connected with assessing the *need* for a program: Is there a (continuing/growing/diminishing) need for a program? *Needs assessments* are an important part of the program management cycle, and although they present methodological challenges, can be very useful in planning or revising programs. We discuss needs assessments in Chapter 6.

Figure 1.4 offers a visual model and summary of key evaluation concepts (Nagarajan & Vanheukelen, 1997). The model relies on an open systems model of a program: program inputs are converted to program activities/processes that in turn produce outcomes. The program operates in an environment into which outcomes are delivered, and in turn, offers opportunities or constraints with which the program must work. In this figure we show how objectives drive the program process: they are connected with actual outcomes via **Effectiveness** (2): Are the observed outcomes consistent with the intended objectives? **Effectiveness** (1): Is our attribution question: did the program cause the observed outcomes? **Cost-effectiveness** connects inputs and actual outcomes, and **cost-benefit analysis** connects the economic value of the benefits to the economic value of the costs. **Adequacy** connects actual outcome to the needs: Were the program outcomes sufficient to meet the need? **Relevance** connects need with objectives: Are the program objectives addressing the need that motivated the program? Although **appropriateness** is not included in the model, it focuses on the logic of the program—the connections among inputs, activities, outputs, and outcomes.

20 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

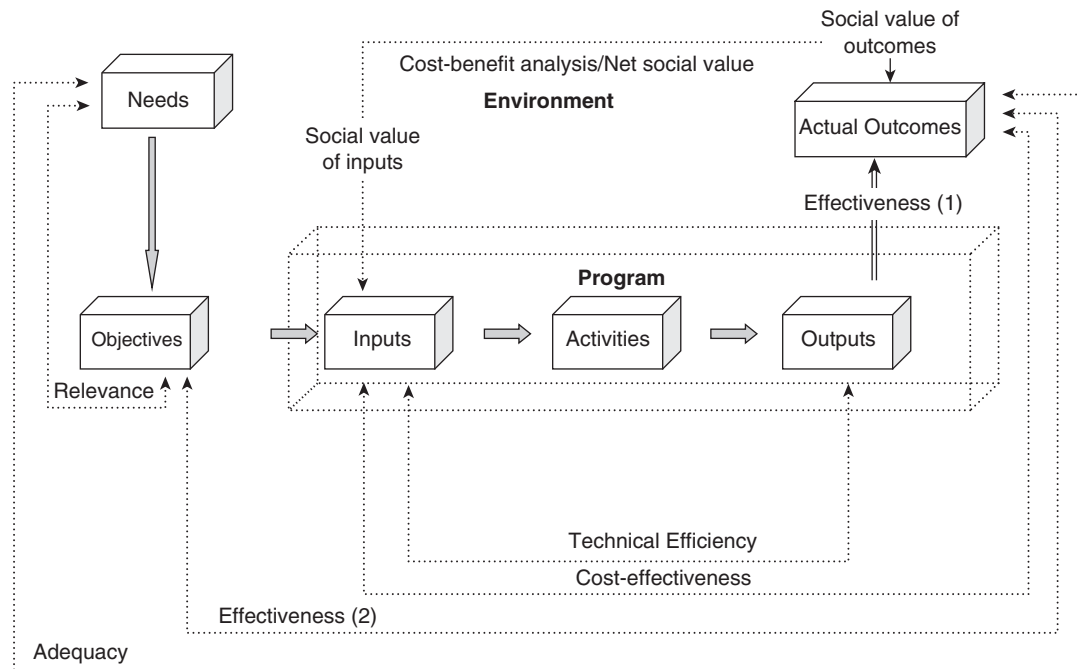


Figure 1.4 An Open Systems Model of Programs and Key Evaluation Issues

SOURCE: Adapted from Nagarajan, N. & Vanheukelen, M. (1997). *Evaluating EU expenditure programmes: A guide* (p. 25).

Each program evaluation will also generate questions that reflect the interests and concerns of the stakeholders involved in particular evaluation processes. These questions may be unique, prompted by circumstances that need to be addressed for that particular evaluation.

Another evaluation question is implicit in most program evaluations: *How well (if at all) was the program implemented?* Program implementation is often assumed as other evaluation questions are addressed, but it is obvious that unless there is evidence of the program having been implemented, it is not meaningful to ask most other evaluation questions (Weiss, 1998).

Assessing program implementation is sometimes done in the first stages of an evaluation process, when considering evaluation questions, clarifying the program objectives, understanding the program structure, and putting together a history of the program. Where programs are “new” (say, 2 years old or less), it is quite possible that gaps will emerge between descriptions of intended program activities and what is *actually* getting done. Indeed, if the gaps are substantial, a program evaluator may elect to recommend an analysis

that focuses on implementation issues, setting aside other results-focused questions for a future time.

In Chapter 2, we look at program logic models that include implementation objectives. By distinguishing between program objectives (what the program is intended to accomplish) and implementation objectives (what has to happen to get the program “on the ground” so that it can produce outputs) we can examine how well the program has been implemented.

FORMATIVE AND SUMMATIVE PROGRAM EVALUATIONS •

Michael Scriven in 1967 introduced the distinction between formative and summative evaluations (Weiss, 1998). Scriven’s definitions reflected his distinction between implementation issues and program effectiveness. Scriven associated **formative evaluations** primarily with analysis of program implementation, with a view to providing program managers and other stakeholders with advice intended to improve the program “on the ground.” For Scriven, **summative evaluations** dealt with whether the program had achieved intended objectives.

Although Scriven’s distinction between formative and summative evaluations has become a part of any evaluator’s vocabulary, it does not generally reflect the way program evaluation is practiced. In program evaluation practice, it is common to see terms of reference that include questions about how well the program was implemented; how (technically) efficient was the program; and how effective was the program. A focus on **program processes** is combined with concerns about whether the program was achieving its intended objectives.

In this book, formative and summative evaluations will be defined in terms of their *intended uses*. This is similar to the distinction offered in Weiss (1998). Formative evaluations are intended to provide feedback and advice with the intention of *improving* the program. Formative evaluations in this book *include* those that examine program effectiveness, but are intended to offer advice aimed at improving the effectiveness of the program. One can think of formative evaluations as manager-focused evaluations, wherein the existence of the program is not questioned.

Summative evaluations are intended to ask “tough questions”: Should we be spending less money on this program; should we be reallocating the money to other uses; or should the program continue to operate? Summative evaluations focus on the “bottom line” with issues of value for money (costs in relation to observed outcomes) as alternative analytical approaches.

22 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

Eleanor Chelimsky (1997) makes a similar distinction between the two primary types of evaluation, which she calls evaluation for development (i.e., the provision of evaluative help to strengthen institutions and to improve organizational performance) and evaluation for accountability (i.e., the measurement of results or efficiency to provide information to decision makers). She adds to the discussion a third general purpose for doing evaluations: evaluation for knowledge (i.e., the acquisition of a more profound understanding about the factors underlying public problems and about the “fit” between these factors and the programs designed to address them).

Program evaluations can, of course, strive to meet all these objectives and be both formative (advice/feedback) and summative (future of the program). But program evaluators will discover that wearing a formative “hat” is a very different experience from being expected to conduct a summative evaluation. Understandably, program manager reactions to these two types of evaluation are quite different. We examine these differences in Chapter 11 (management and evaluation).

● *EX ANTE* AND *EX POST* PROGRAM EVALUATIONS

Typically, program evaluators are expected to conduct evaluations of ongoing programs. Usually, the program has been in place for some time, and the evaluator’s tasks include assessing the program up to the present and offering advice for the future. These *ex post* evaluations are challenging: they necessitate relying on information sources that may or may not be ideal for the evaluation questions at hand. Rarely are baselines or comparison groups available, and if they are, they are only roughly appropriate. In Chapter 3 and Chapter 5 we will learn about the research design options and qualitative evaluation alternatives that are available for such situations.

Ex ante (before implementation) program evaluations are less frequent. Cost-benefit analyses can be conducted *ex ante*, to prospectively assess whether a program at the design stage (or one option from among several alternatives) is cost-beneficial. Assumptions about implementation and the existence and timing of outcomes are required to permit such analyses.

In some situations, it may be possible to implement a program in stages, beginning with a pilot project. The pilot can then be evaluated (and compared to the existing “no program” status quo), and the evaluation results used as a kind of *ex ante* evaluation of a broader implementation.

One other possibility is to plan a program so that before it is implemented, **baseline measures** are constructed and appropriate data are gathered. The “before” situation can be documented and included in any future

program evaluation or performance measurement system. In Chapter 3, we discuss the strengths and limitations of before-and-after research designs. They clearly offer us an opportunity to assess the incremental impacts of the program, but in environments where there are other factors that could also plausibly account for the observed outcomes, this design, by itself, may not be adequate.

ANALYZING CAUSE AND EFFECT LINKAGES IN PROGRAM EVALUATIONS

The attribution question—*To what extent, if any, did the program cause the observed outcomes?*—is at the heart of most program evaluations. Figure 1.4 depicts the program being “connected” to the observed outcomes, with a question mark above the causal arrow to suggest that this connection is a key subject of most evaluations.

To say that the program caused the observed outcomes is really to assert three different things.

- The program occurred before the observed outcomes.
- The program co-varied with the observed outcomes, that is, when (or where) the program occurred, the outcomes tended to occur.
- There were no other plausible rival explanations for the observed outcomes, other than the program.

These are “conditions of causality,” which we examine in Chapter 3, and they are a key challenge for program evaluators. Of the three, the last one is most problematic in that program evaluations are often conducted in circumstances where it is not possible to rigorously rule out **rival hypotheses**.

This model of cause and effect relationships is intended to work in situations where we typically do not have a “full” explanation for what we observe in an evaluation. For example, if we were evaluating the effectiveness of a neighborhood watch program in reducing the number of burglaries in a community, we would ideally design an evaluation that would permit us to see what effect the program had on burglaries, controlling for other factors that could also affect burglary levels. We would not expect the program to eliminate burglaries; instead, we would be looking for a reduction that we judged to be significant.

In program evaluations, we almost never expect the implementation of a program to account for all the changes we observe in the outcome in

24 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

question. The relationship between the program and the observed outcome is probabilistic; there is almost always variance in the outcome variable(s) that is not accounted for by the presence or absence of the program.

Some program evaluations are conducted under conditions where data appropriate for ascertaining or even systematically addressing the attribution question are scarce. In these situations, the evaluator or members of the evaluation team may end up relying, to some extent, on their professional judgment. Indeed, such judgment calls are familiar to program managers, who rely on their own observations, experiences, and interactions to detect patterns and make choices on a daily basis. Scriven (in progress, unpublished) suggest that our capacity to observe and detect **causal relationships** is built into us. We are hard-wired to be able to organize our observations into patterns and detect causal relationships therein.

For evaluators, it may seem “second best” to have to rely on their own judgment, but realistically *all* program evaluations entail a substantial number of judgment calls, even when valid and reliable data and appropriate comparisons are available. As Daniel Krause (1996) has pointed out, “a program evaluation involves human beings and human interactions. This means that explanations will rarely be simple, and interpretations cannot often be conclusive” (p. xviii). Clearly then, systematically gathered evidence is a key part of any good program evaluation, but evaluators need to be prepared for, and accept the responsibility of, exercising professional judgment as they do their work.

● THE PROCESS OF CONDUCTING A PROGRAM EVALUATION

A key assumption in this book is that designing a sound performance measurement system benefits from an understanding of the principles and practices of program evaluation. Although some practitioners appear to take the view that performance measurement can replace program evaluations in organizations, that view overstates the benefits of performance measurement and understates the benefits of program evaluations. The two approaches share a common **goal** of illuminating program processes and outcomes to improve the quality of decisions made about programs and policies. But, as we shall see, they are not interchangeable. In Chapter 8, we compare program evaluation and performance measurement—pointing out the differences between these two evaluation approaches.

Because the textbook is structured so that program evaluation methods are introduced and explained first, it is useful, in this introductory chapter,

to outline the sequence of activities in a typical program evaluation. In Chapter 9, we do the same thing for designing and implementing a performance measurement system.

The 15 steps in conducting a program evaluation (10 steps for the evaluability assessment phase and 5 steps for the evaluation study itself) are offered in the spirit of the Checklists Project led by Daniel Stufflebeam and Michael Scriven (The Evaluation Center, 2001).

Scriven (2000) advocates the use of checklists.

There are many different types of checklist, although they have at least one nondefinitional function in common—that of being a mnemonic device. This function alone makes them useful in evaluation, since the nature of evaluation calls for a systematic approach to determining the merit, worth, etc., of what are often complex entities. Hence, a list of the many components or dimensions of performance of such entities is frequently valuable. (p. 1)

It is important to remember that each program evaluation situation is different and that the “steps” outlined below may not reflect the actual train of events as one designs and conducts a given evaluation. It is possible, for example, that clarifying the purposes of the evaluation may need to be revisited as the evaluator has discussions with program managers and other stakeholders about the structure and the objectives of the program.

General Steps in Conducting a Program Evaluation

Rutman (1984) distinguished between planning for an evaluation and actually conducting the evaluation. The **evaluation assessment** process can be separated from the **evaluation study** itself, so that managers and other stakeholders can see whether the results of the evaluation assessment support a decision to proceed with the evaluation.

Table 1.1 summarizes 10 questions that are important to answer as part of most evaluation assessments. Also are five additional steps that are included in most evaluation studies. Each of the questions and steps is elaborated in the discussion that follows Table 1.1.

The Evaluation Assessment

1. Who are the client(s) for the evaluation?

The general perspective taken in this outline of the evaluation assessment process is that program evaluations are substantially *user-driven*. Like

26 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

Table 1.1 Summary of Key Questions and Steps in Conducting Evaluation Assessments and Evaluation Studies

Questions to answer as part of an evaluation assessment:

1. Who are the client(s) for the evaluation?
2. What are the questions and issues driving the evaluation?
3. What resources are available to do the evaluation?
4. What has been done previously?
5. What is the program all about?
6. What kind of environment does the program operate in and how does that affect the comparisons available to an evaluator?
7. Which research design alternatives are desirable and appropriate?
8. What information sources are available/appropriate, given the evaluation issues, the program structure and the environment in which the program operates?
9. Given all the issues raised in points 1–8, which evaluation strategy is least problematical?
10. Should the program evaluation be undertaken?

Steps in conducting an evaluation study:

1. Develop the measures and collect the data.
2. Analyze the data.
3. Write the report.
4. Disseminate the report.
5. Make changes, based on the evaluation.

Michael Patton (1997), who makes **utilization** a key criterion in the design and execution of program evaluations, the view taken here is that the intended users *must* be identified early in the process and *must* be involved in the evaluation assessment. That does not mean that intended users should determine the way the evaluation should be conducted, but their information needs are a key part of the assessment process.

Possible clients of the evaluation include, but are not limited to

- Program managers
- Agency/department executives
- External agencies (including funding agencies)

- Clients of the program
- Political decision makers/members of governing bodies (including boards of directors)
- Community leaders

All program evaluations are political in that selecting what to evaluate, who to report the results to, how to collect the information, even how to interpret the data, are affected by the interests and values of key stakeholders. The evaluation's client(s) will also likely affect how the goals, objectives, activities, and intended outcomes of the program are defined for the purpose of the evaluation (Boulmetis & Dutwin, 2000). Generally, the more diverse the clients for the evaluation results, the more complex the political process that surrounds the evaluation itself. Indeed, as Ian Shaw (2000) comments, "many of the issues in evaluation research are influenced as much, if not more, by political as they are by methodological considerations" (p. 3).

Because of the political nature of program evaluations, an evaluation plan (outlining such items as the purpose of the evaluation, the key evaluation questions, and the intended audience), worked out and agreed to by both the evaluator and the client prior to the start of the evaluation, can be very useful. Owen and Rogers (1999) discuss the development of evaluation plans in some detail. In the absence of such a written plan, they argue, "there is a high likelihood that the remainder of the evaluation effort is likely to be unsatisfactory to all parties" (p. 71), and they suggest the process should take up to 15% of the total evaluation budget.

2. What are the questions and issues driving the evaluation?

Program evaluators, particularly as they are learning their craft, are well advised to seek explicit answers to the following questions:

- Who wants the evaluation done?
- Why do they want it done?
- Are there hidden agendas or covert reasons for wanting the program evaluated?
- What are the main evaluation issues that they want addressed (effectiveness, efficiency, adequacy, appropriateness, rationale, need)?
- Is the evaluation intended to be formative or summative, or both?

Answering these sorts of questions prior to agreeing to conduct an evaluation is essential because, as Owen and Rogers (1999) point out, "there is often a diversity of views among program stakeholders about the purpose of an evaluation. Different interest groups associated with a given program

28 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

often have different agendas, and it is essential for the evaluator to be aware of these groups and know about their agendas in the negotiation stage” (p. 66).

Given time and resource constraints, an evaluator cannot hope to address all the interests of all program stakeholders within the context of one evaluation. For this reason, the evaluator must reach a firm agreement with the evaluation client(s) about the questions to be answered by the evaluation. This process will often involve working with the client to help them narrow the list of questions they are interested in, a procedure which may necessitate “educating them about the realities of working within a budget, challenging them as to the relative importance of each issue, and identifying those questions which are not amenable to answers through evaluation” (Owen & Rogers, 1999, p. 69).

3. What resources are available to do the evaluation?

A general issue for almost all program evaluations is the scarcity of resources available to design and complete the work. Greater sophistication in evaluation research designs almost always entails larger expenditures of resources. For example, achieving the necessary control over the program and its implementation environment to conduct experimental or quasi-experimental evaluations generally entails modifying existing administrative procedures and perhaps even temporarily changing or suspending policies (to create no-program comparison groups, for example).

Although most resources can be converted to money, it is useful to distinguish among several kinds of resources needed for program evaluations.

- Time
- Human resources, including persons with necessary skills and experience
- Organizational support, including written authorizations for other resources needed to conduct the evaluation
- Money

Agreements reached about all resource requirements should form part of the evaluation plan.

4. What has been done previously?

Program evaluators should take advantage of work that has already been done. In a given situation, there may be previous evaluations or evaluations of similar programs in other settings.

In the field of program evaluation, a sub-discipline has arisen that focuses on the **meta-evaluation** of programs. Meta-evaluations are syntheses

of existing studies in a given area, and are intended to summarize what we know about, for example, head start programs.

Although we will not be covering meta-evaluation as a separate topic in this book, readers will find references to this area in other program evaluation texts (see Cook, 1994; Patton, 1997; Rossi, Freeman, & Lipsey, 1999; Shadish, Cook, & Campbell, 2002).

Questions to keep in mind as you review previous work are as follows:

- If there are other evaluations, how comparable are the program(s) to the one you are proposing to evaluate?
- Is there published research that is relevant?
- Are there unpublished assessments or opinions about the program or related programs?
- Who did the previous work, and how credible is it?

5. *What is the program all about?*

In Chapter 2 we will learn ways of constructing program logic models. Briefly, **program logics** are models of programs that depict the key activities in the program and the flow/conversion of resources to outcomes.

Related questions include:

- What are the program objectives?
- What is the history of the program?
- Is the program growing, remaining stable, or declining?
- What is the structure of the program (the key causal linkages among the main parts of the program)?

Although central agencies, executive managers, and even program managers all claim a commitment to stating clear objectives for programs, the nature of the political processes that create programs often means that objectives are stated in very general terms. This presents a challenge for an evaluator who wants to interpret language, which is often intended for stakeholders whose interests in a program may compete with each other, into words that lend themselves to measurement. In the absence of clearly and consistently stated goals, evaluating a program's effectiveness at achieving its goals becomes extremely difficult, if not impossible (Berk & Rossi, 1999).

Another important issue to consider is whether the program has been implemented and is being administered as the program logic intends. As Berk and Rossi (1999) point out, "questions about how a program is functioning logically precede questions about program impact. An impact assessment is a waste of time unless the intervention is known and understood"

30 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

(p. 71). In other words, a program's success or failure at accomplishing its stated objectives may have nothing to do with the design of the program or the theory behind it. If the program has not been implemented the way it was intended, its impact (or lack thereof) may be the result of its administration rather than of its design (Weiss, 1998).

6. What kind of environment does the program operate in and how does that affect the comparisons available to an evaluator?

In this book, we conceptualize programs as open systems. Although we look at what that means in Chapter 2, one implication of an open systems approach is that programs interact with the environments in which they are embedded.

For program evaluators, there are issues that affect the design choices available.

- Have any baseline data been kept?
- How large is the client base for the program?
- Is the organization in which the program is embedded stable or in a period of change?

7. Which research design alternatives are desirable and appropriate?

We discuss research designs in Chapter 3, and a key point underlying that chapter is the fact that *all* program evaluations involve making comparisons. The kinds of comparisons we make depend on the evaluation questions we are addressing *and* on the resources available.

Typically, program evaluations involve multiple research designs—it is unusual for an evaluator to construct a design that relies on, for instance, a single time series alone. An important consideration for practitioners is to know the strengths and weaknesses of different designs so that combinations of designs can be chosen that complement each other.

8. What information sources are available/appropriate, given the evaluation issues, the program structure, and the environment in which the program operates?

In most program evaluations, resources to collect data are quite limited, and many research design options that would be desirable are simply not feasible. Given that, it is important to ask what data are available, and how key constructs in the program logic would be measured, in conjunction with decisions about research designs. Research design considerations can be used as a rationale for prioritizing additional data collection.

Specific questions include the following:

- What data are currently available?
- Are currently available data reliable and complete?
- How can currently available data be used to validly measure constructs in the program logic?
- Are data available that allow us to measure key environmental factors that will affect the program and its outcomes?
- Will it be necessary for the evaluator to collect additional information to measure key constructs in the program logic?
- Given research design considerations, what are the highest priorities for collecting additional data?

In Chapter 4 we discuss the process of measuring program outputs, outcomes, and environmental factors, taking into account validity and reliability issues for different measurement processes.

9. Given all the issues raised in questions 1 to 8, which evaluation strategy is least problematical?

No evaluation design is unassailable. The important thing for evaluators is to be able to understand the underlying logic of causality and the ways we can work with that logic to *anticipate* key criticisms that could be made, and have at least a well thought out verbal response to those criticisms.

Much of the work that we do as evaluators is not going to involve randomized controlled experiments, although some consider it to be the “gold standard” of rigorous social scientific research (see, for example, Lipsey, 2000). Although there is far more diversity in views of what is sound evaluation practice, it can become an issue for a particular evaluation, given the background or interests of persons or organizations who might mount criticisms of your work. The U.S. Office of Management and Budget (2004) guidelines on assessing program effectiveness, for example, emphasize the importance of randomized experimental designs, a view also held by the Council for Excellence in Government (Coalition for Evidence-Based Policy, n.d.). This standard harks back to a time in the 1960s when the evaluation profession was dominated by the belief that social experimentation was key to understanding whether programs worked. Whether randomized experiments should be the “gold standard” for the evaluation profession is still sometimes the subject of vigorous discussion in professional forums (Scriven, in progress, unpublished).

There is value in understanding the canons of rigorous research to be able to proactively acknowledge weaknesses in an evaluation strategy. But ultimately, evaluators must make some hard choices and be prepared to accept the fact that their work can and probably will be criticized.

32 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

10. Should the program evaluation be undertaken?

The final question in the evaluation assessment process is whether to proceed with the actual evaluation study. It is possible that after having looked at the mix of evaluation issues, resource constraints, organizational and political issues, research design, and measurement constraints, the program evaluator preparing the assessment recommends that no evaluation be done at this time. Although a rare outcome of the evaluation assessment phase, it does happen, and can save an organization considerable time and effort that probably would not have yielded a credible product.

Choosing the combination of methodologies, and ultimately making a decision whether to go ahead with an evaluation, involves substantial amounts of professional judgment. Evaluator experience is key in being able to negotiate a path that permits designing a credible evaluation project. Evaluator judgment is an essential part of putting together the requirements for a defensible study and making a recommendation to either proceed or not.

The Evaluation Study

Up to this point, we have outlined the planning process for conducting program evaluations. If a decision is made to go ahead with the evaluation, there are several steps common to most program evaluations.

1. Develop the measures and collect the data.

Many program evaluations rely on a mix of existing and study-specific data. Data-collection instruments may need to be designed, pretested, and then implemented. Surveys are a common means of collecting data, and we review key issues involved in doing surveys in Chapter 4.

Relevant questions include:

- Are existing data valid measures of the constructs they are intended to measure?
- Are all the evaluation questions and sub-questions addressed by at least one data-collection effort?
- Are there ways of building in triangulation of data sources, that is, two or more independent measures of a given construct?

2. Analyze the data.

Data analysis can be simple or complex, depending on the evaluation questions, the types of data that address those questions, and the comparisons needed to sort out and reduce the threats from rival hypotheses.

Data analysis can be quantitative (involve working with variables that are represented numerically) or qualitative (involve analysis of words, documents, text, narratives, and other nonnumerical representations of data).

A general rule that should govern all data analysis is to employ the *least* complex method that will fit the situation. One of the features of early evaluations based on models of social experimentation was the reliance on sophisticated, multivariate, statistical tools to analyze program evaluation data. Although that strategy addressed possible criticisms by scholars, it often produced reports that were inaccessible or untrustworthy from a user's perspective. More recently, program evaluators have adopted mixed strategies for analyzing data, which rely on statistical tools where they are necessary, but also incorporate visual/graphic representations of findings.

In this book we will not cover data-analysis methods in great detail. Reference to statistical methods is in Chapter 3 (research designs) and in Chapter 4 (measurement). In Chapter 3, key findings from examples of actual program evaluations are displayed and interpreted. In the appendix to Chapter 3, we summarize basic statistical tools and the conditions under which they are normally used. In Chapter 5 (qualitative evaluation methods), we cover the fundamentals of qualitative data analysis; and in Chapter 6, in connection with needs assessments, we introduce some basics of sampling, and generalizing from sample findings to populations.

3. Write the report.

In preparing an evaluation report, the key part is usually the recommendations that are made. Here, sound professional judgment plays a key role in that recommendations must not only be backed up by evidence, but must also be appropriate, given the organizational and political context. Making recommendations that reflect key evaluation conclusions *and* are feasible is a skill that is among the most valuable that an evaluator can possess.

Although each program evaluation report will have unique requirements, there are some general guidelines that assist in making reports readable, understandable, and useful.

- Rely on visual representations of findings where possible
- Use clear, simple language in the report
- Use more headings and subheadings rather than fewer to generate a table of contents that is complete and explicit
- Prepare a clear, concise executive summary
- Solicit feedback on drafts of the report before finalizing it

34 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

4. Disseminate the report.

Program evaluators generally have an obligation to produce a report *and* make one or more presentations of the findings, conclusions, and recommendations to key stakeholders, including the clients of the study.

Program evaluators differ in their views of how much interaction there should be between evaluators and clients, at all stages in the process. One view, articulated by Scriven (1997), is that program evaluators should be very careful about getting involved with their clients—interaction at *any* stage in an evaluation, including postreporting, can compromise their **objectivity**.

Patton (1997) argues that unless program evaluators get involved with their clients, the evaluation is unlikely to be used. We look at this issue in Chapter 11, when we consider whether program evaluations can be objective.

5. Make changes, based on the evaluation.

Program evaluations are one means by which stakeholders acquire information that becomes part of the rationale for changes in the program or the organization in which it operates. Evaluations tend to result in *incremental* changes, if any changes can be attributed to the evaluation. It is quite rare for an evaluation to result in the elimination of a program, even though summative program evaluations are often intended to raise this question (Weiss, 1998).

The following are possible changes based on program evaluations:

- Improving the existing program
- Increasing the size of the program
- Increasing the scope of the program
- Downsizing the program
- Replacing or eliminating the program

● SUMMARY

This book is intended for persons who want to learn the principles and the essentials of the practice of program evaluation and performance measurement. Given the diversity of the field, it is not practical to cover all the approaches and issues that have been raised by scholars and practitioners in the last 30-plus years. Instead, this book adopts a stance with respect to several key issues that continue to be debated in the field.

First, we approach program evaluation and performance measurement as two complementary ways of creating information that is intended to reduce uncertainties for stakeholders who are involved in making decisions

about programs or policies. We have structured the textbook so that methods and practices of program evaluation are introduced first and then are adapted to performance measurement—we believe that sound performance measurement practice depends on an understanding of program evaluation essentials.

Second, a key emphasis in this textbook is on assessing the effectiveness of programs, that is, the extent to which a program has accomplished its intended outcomes. Understanding the logic of causes and effects as it is applied to evaluating the effectiveness of programs is assisted by learning key features of experimental and quasi-experimental research designs; we discuss this in Chapter 3.

Third, the nature of evaluation practice is such that all of us who have participated in program evaluations understand the importance of judgment calls. The evaluation process, from the initial step of deciding to proceed with an evaluation assessment to framing and reporting the recommendations, is informed by our own experiences, beliefs, values, and expectations. Methodological tools provide us with ways of disciplining our judgment and rendering key steps in ways that are transparent to others, but many of these tools are designed for social science applications. In many program evaluations, resource constraints usually mean that the tools we apply are not ideal for the situation at hand. Learning some of the ways in which we can cultivate good professional judgment is a principal topic in Chapter 12 (the nature and practice of professional judgment).

Fourth, the importance of program evaluation and performance measurement in contemporary public and nonprofit organizations is related to a broad movement in North America, Europe, and Australasia to manage for results. Performance management depends on having high-quality information about how well program and policies have been implemented and how effectively and efficiently they have performed. Understanding how program evaluation and performance measurement fit into the performance management cycle and how evaluation and management work together in organizations is a theme that runs through this textbook.

DISCUSSION QUESTIONS ●

1. As you are reading Chapter 1, what five ideas about the practice of program evaluation were most important for you? Summarize each idea in a couple of sentences, and keep them so that you can check on your initial impressions of the textbook, as you cover other chapters in the book.

36 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

2. Read the entire Table of Contents for this textbook, and based on your own background and experience, what do you anticipate will be the easiest parts of this book for you to understand? Why?
3. Again, having read the Table of Contents, which parts of the book do you think will be most challenging for you to learn? Why?
4. Do you consider yourself to be a “words” person, that is, you are most comfortable with written and spoken language; a “numbers” person, that is, you are most comfortable with numerical ways of understanding and presenting information; or you “both,” that is, you are equally comfortable with words and numbers?
5. Find a classmate who is willing to discuss Question 4 with you. Find out from each other whether you share a “words,” “numbers,” or a “both” preference. Ask each other why you seem to have the preferences you do? What is it about your background and experiences that may have influenced you?
6. What do you expect to get out of this textbook for yourself? List four or five goals or objectives for yourself as you work with the contents of this textbook. An example might be, “I want to learn how to conduct evaluations that will get used by decision makers.” Keep them so that you can refer to them as you read and work with the contents of the book. If you are using this textbook as part of a course, take your list of goals out at about the half-way point in the course and review them. Are they still relevant, or do they need to be revised? If so, revise them so that you can review them once more as the course ends. For each of your own objectives, how well do you think you have accomplished that objective?

● REFERENCES

-
- Auditor General of British Columbia and Deputy Ministers' Council. (1996). *Enhancing accountability for performance: A framework and an implementation plan—Second joint report*. Victoria: Queen's Printer for British Columbia.
- Bakvis, H., & Juillet, L. (2004). *The horizontal challenge: Line departments, central agencies and leadership*. Ottawa: Canada School of Public Service.
- Basilevsky, A., & Hum, D. (1984). *Experimental social programs and analytic methods: An evaluation of the U.S. income maintenance projects*. Orlando, FL: Academic Press.

- Berk, R. A., & Rossi, P. H. (1999). *Thinking about program evaluation* (2nd ed.). Thousand Oaks, CA: Sage
- Boulmetis, J., & Dutwin, P. (2000). *The ABC's of evaluation: Timeless techniques for program and project managers*. San Francisco: Jossey-Bass.
- Bernstein, D. (1999). Comments on Perrin's effective use and misuse of performance measurement. *American Journal of Evaluation*, 20(1), 85–93.
- Chelimsky, E. (1997). The coming transformations in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: a handbook* (pp. 1-26). Thousand Oaks, CA: Sage.
- Coalition for Evidence-Based Policy. (n.d.). Home page. Retrieved July 13, 2004, from <http://www.excelgov.org/displayContent.asp?Keyword=prppcHomePage>
- Cook, T. D. (Ed.). (1994). *Meta-analysis for explanation: A casebook* (paper ed.). New York: Russell Sage Foundation.
- Fitzgerald, J., & Rasheed, J. M. (1998). Salvaging an evaluation from the swampy lowland. *Evaluation and Program Planning*, 21(2), 199–209.
- Krause, D. R. (1996). *Effective program evaluation: An introduction*. Chicago: Nelson-Hall.
- Lipsey, M. W. (2000). Method and rationality are not social diseases. *The American Journal of Evaluation*, 21(2), 221–223.
- Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. *Canadian Journal of Program Evaluation*, 16(1), 1–24.
- McDavid, J. C. (2001). Solid-waste contracting-out, competition, and bidding practices among Canadian local governments. *Canadian Public Administration*, 44(1), 1–25.
- Nagarajan, N., & Vanheukelen, M. (1997). *Evaluating EU expenditure programmes: A guide*. Luxembourg: Office for Official Publications of the European Communities.
- Newcomer, K. E. (1997). Using performance measurement to improve public and nonprofit programs. In K. E. Newcomer (Ed.), *New directions for evaluation*, 75, 5–14.
- Office of Management and Budget. (2004). *What constitutes strong evidence of a program's effectiveness?* Retrieved July 17, 2004, from http://www.whitehouse.gov/omb/part/2004_program_eval.pdf
- Owen, J. M., & Rogers, P. J. (1999). *Program evaluation: Forms and approaches* (international ed.). London, Thousand Oaks: Sage.
- Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.
- Perrin, B. (1998). Effective use and misuse of performance measurement. *American Journal of Evaluation*, 19(3), 367–379.
- Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.
- Rutman, L. (1984). Introduction. In L. Rutman (Ed.), *Evaluation research methods: A basic guide: Vol. 3. Sage focus editions* (2nd ed., p. 239). Beverly Hills, CA: Sage.

38 ● PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

- Scriven, M. (1997). Truth and objectivity in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 477–500). Thousand Oaks, CA: Sage.
- Scriven, M. (2000). *The logic and methodology of checklists*. Retrieved July 22, 2004, from http://www.wmich.edu/evalctr/checklists/papers/logic_methodology.pdf
- Scriven, M. (In progress, unpublished). *Causation*. New Zealand: University of Auckland.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shaw, I. (2000). *Evaluating public programmes: Contexts and issues*. Aldershot, UK, Burlington, VT: Ashgate.
- Talarico, T. (1999). *An evaluation of the Neighbourhood Integrated Service Team program*. Unpublished Master's thesis, University of Victoria, British Columbia, Canada.
- The Evaluation Center. (2001). *The evaluation checklist project*. Retrieved July 13, 2004, from <http://www.wmich.edu/evalctr/checklists/>
- Treasury Board of Canada Secretariat. (2001). *Evaluation Policy*. Retrieved March 24, 2004 from http://www.tbs-sct.gc.ca/pubs_pol/dcgpubs/TBM_161/ep-pe_e.asp
- Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

● NOTE

-
1. (http://www.city.vancouver.bc.ca/parks/board/2004/040322/accomplishments_2003.PDF)