CHAPTER 1

KEY CONCEPTS AND ISSUES IN PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

INTRODUCTION

In this chapter, we introduce key concepts and principles for program evaluations. We describe how program evaluation and performance measurement are complementary approaches to creating information for decision makers and stakeholders in public and nonprofit organizations. We introduce the performance management cycle and show how program evaluation and performance measurement fit results-based management systems. A typical **program evaluation** is illustrated with a **case study**, and its strengths and limitations are summarized. Although our main focus in this textbook is on understanding how to evaluate the effectiveness of programs, we introduce 10 general questions (including program effectiveness) that can underpin evaluation projects. We also summarize 10 key steps in assessing the feasibility of conducting a program evaluation, and conclude with the five key steps in doing and reporting an evaluation.

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 when we design and implement policies and programs. **Evaluation** can be viewed as a structured process that creates and synthesizes information intended to reduce the level of uncertainty for decision makers and stakeholders about a given program or policy. It is usually 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. Evaluations can also discover unintended effects of programs and policies, which can affect overall assessments of programs or policies.

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 decision makers and stakeholders as they assess whether and how a program accomplished its intended outcomes.

As you read this textbook, 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 guide 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 and practitioners of evaluation will encounter **randomized experiments (randomized controlled trials** or **RCTs)** in which some 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 RCTs are relatively rare in the practice of program evaluation, and there is controversy around making them the **benchmark** or **gold standard** for sound evaluations, they are still often considered as exemplars of "good" evaluations (Cook, Scriven, Coryn, & Evergreen, 2010).

More frequently, program evaluators do not have the resources, time, or control over program design or implementation situations to conduct experiments. In many cases, an

experimental design may not 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 readily available, and scarce resources available to collect information.

Alternatively, data may exist (program records would be a typical situation) but closer scrutiny of these data indicates 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. We will cover these kinds of evaluation settings throughout the book.

Integrating Program Evaluation and Performance Measurement

Evaluation as a field has been transformed in the past 20 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 to that end. Measuring performance is often accompanied by requirements to publicly report performance results for programs.

Performance measurement is controversial among evaluators; some advocate that the profession embrace performance measurement (Bernstein, 1999), while others are skeptical (Feller, 2002; 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 on 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 that permit managers or other stakeholders 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. However, this textbook will show how sound performance measurement, regardless of who does it, depends on an understanding of program evaluation principles and practices. Core skills that evaluators learn can be applied to performance measurement (McDavid & Huse, 2006). Managers and others who are involved in developing and implementing performance measurement systems for programs or organizations typically encounter problems similar to those encountered by 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.

Consequently, rather than seeing performance measurement as a quasi-independent enterprise, in this textbook we *integrate* performance measurement into evaluation by grounding it in the same core tools and methods that are essential to assess program processes and effectiveness. Thus, **program logic models** (Chapter 2), **research designs** (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 as a replacement for evaluations. Analysts in the evaluation field (Mayne, 2001, 2006, 2008; McDavid & Huse, 2006; Newcomer, 1997) have generally recognized this complementarity, but in some jurisdictions, efforts to embrace performance measurement have eclipsed program evaluation (McDavid, 2001; McDavid & Huse, 2006). There is growing evidence that the promises that have been made for performance measurement as an accountability and **performance management** tool have not materialized (McDavid & Huse, 2012; Moynihan, 2008). We see an important need to balance these two approaches, and our approach in this textbook is to show how they can be combined in ways that make them complementary, without overstretching their real capabilities.

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 is part of a broad movement of **new public management (NPM)** in public administration that has had significant impacts on governments worldwide since it came onto the scene in the early 1990s. NPM is premised on principles that emphasize the importance of stating clear program and policy objectives, measuring and reporting program and policy outcomes, and holding managers, executives, and politicians accountable for achieving expected results (Hood, 1991; Osborne & Gaebler, 1992). Evidence of actual accomplishments is central to performance management. Evidence-based or evidence-informed policy making has become an important feature of the administration of governments in Western countries (Campbell, Benita, Coates, Davies, & Penn, 2007; Solesbury, 2001). Evidence-based decision making depends heavily on both evaluation and performance measurement.

Increasingly, there is an expectation that managers will be able to participate in evaluating their own programs and also be involved in developing, implementing, and publicly reporting the results of performance measurement. Information from program evaluations and performance measurement systems is expected to play a role in the way managers manage their programs. Changes to improve program operations and **efficiency** 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 have widely embraced a focus on **program outcomes**. Central agencies (including the U.S. Federal Office of Management and Budget [OMB] and the General Accountability Office [GAO] and the Treasury Board of Canada Secretariat [TBS]), as well as state and provincial finance

departments and auditors, have developed policies and articulated expectations that shape the ways program managers are expected 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, successive federal administrations beginning with the Clinton administration have embraced program **goal** setting, performance measurement, and reporting as a regular feature of program accountability (Roessner, 2002). The Bush administration between 2002 and 2009 emphasized the importance of program performance in the budgeting process. The OMB introduced assessments of departments and agencies using a methodology called PART (Performance Assessment Rating Tool). Essentially, OMB analysts reviewed existing evaluations conducted by departments and agencies as well as performance measurement results and offered their own overall rating of program performance. Each year, one fifth of all federal programs were "PARTed," and the review results were included with the administration's budget request to Congress.

The Obama administration, although departing from top-down PART assessments of program performance (Joyce, 2011), continued this emphasis on performance by appointing the first Federal Chief Performance Officer, leading the "management side of OMB," which is expected to work with agencies to "encourage use and communication of performance information and to improve results and transparency" (OMB, 2012). Also evident is the emphasis on program evaluation as an approach to assessing performance. In the fiscal year 2011 budget cycle, for example, a total of 36 high-profile evaluations of programs were approved for funding for 17 departments and agencies (Joyce, 2011).

In Canada, a major update of the federal government's evaluation policy was announced in 2009 (TBS, 2009). The main plank in that policy is a requirement that federal departments and agencies evaluate all their programs on a 5-year cycle. Program evaluation is explicitly linked to assessing "program performance"—what is noteworthy is that performance includes the economy, efficiency, and effectiveness of programs. For the first time, the performance measurement function in all departments and agencies, which had been a separate management activity, is now linked to the evaluation function. Heads of departmental evaluation units are expected to take some responsibility for ensuring that program performance measures are implemented in ways that support program evaluation requirements.

Performance management is now central to public and nonprofit management. What was once an innovation in the public and nonprofit sectors in the early 1990s has since become an expectation. Fundamental to performance management is the importance of program and policy performance results being collected, analyzed, compared (often with performance targets), and then used to monitor and make decisions. Performance results are also expected to be used to increase the transparency and accountability of public and nonprofit organizations and even governments, principally through periodic public performance reporting. Many jurisdictions have embraced mandatory public performance reporting as a visible sign of their commitment to improved accountability (Hatry, 2006).

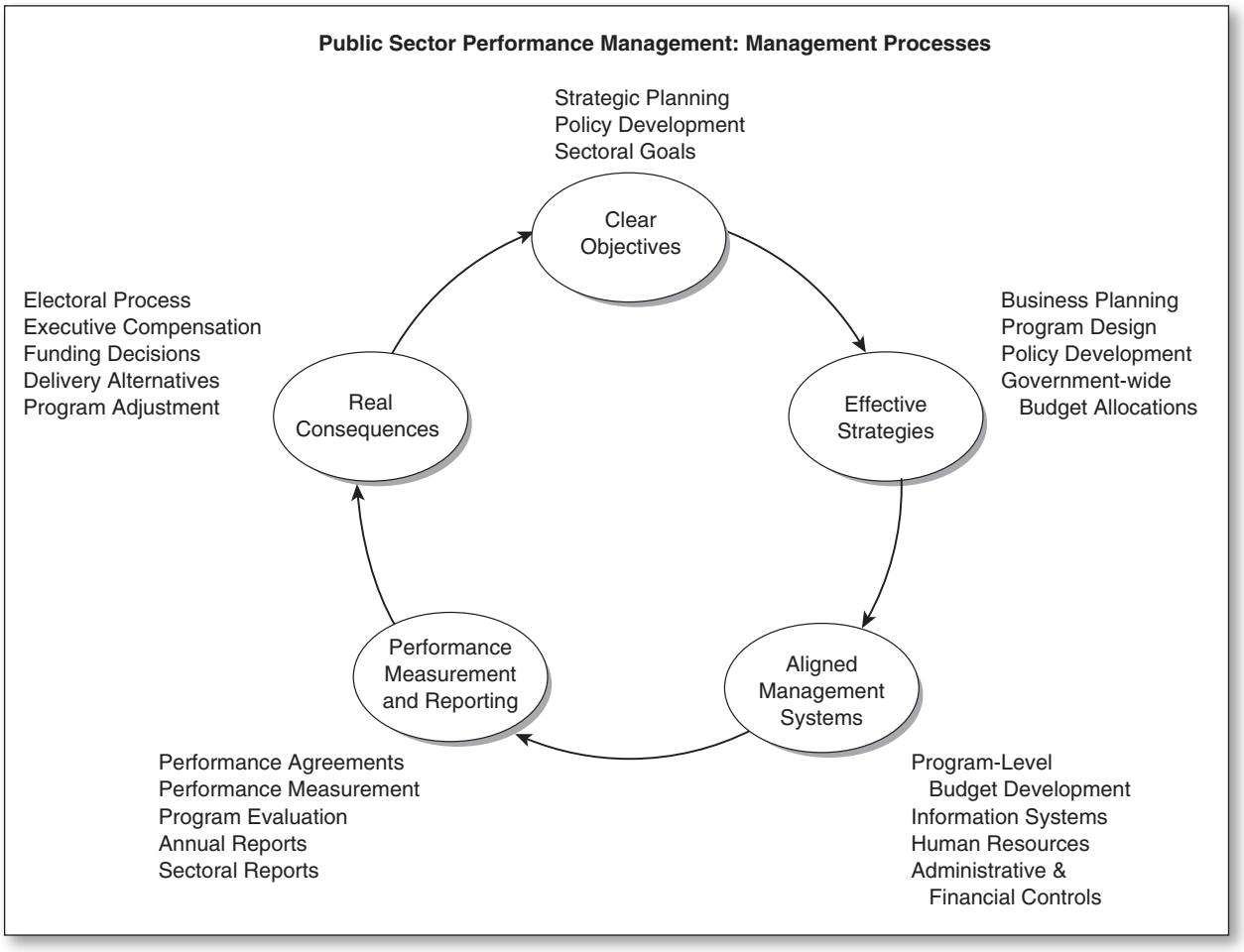
The Performance Management Cycle

Organizations typically run through an annual performance management cycle that includes budgeting, managing, and reporting their financial and nonfinancial results. Stepping back from this annual cycle, we can see a more strategic cycle that encompasses

strategic planning through to evaluating and reporting results. The **performance management cycle** is a model that includes an iterative planning–implementation–evaluation–program adjustments sequence in which program evaluation and performance measurement play important roles as ways of providing information to decision makers who are engaged in leading and 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 and other stakeholders in public sector and nonprofit organizations. Figure 1.1 shows a model of how organizations can 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 & Deputy Ministers' Council, 1996), the terminology and the look of the framework are similar to others that have been adopted by many North American, European, and Australasian jurisdictions.

Figure 1.1 Performance Management Cycle



Source: Adapted from Auditor General of British Columbia and Deputy Ministers' Council (1996).

The five stages in the performance management cycle begin and end with formulating clear (strategic) objectives for organizations and, hence, for programs and policies. Strategic objectives are translated into program and policy designs intended to achieve those objectives. These are connected with resources. ***Ex ante* evaluations** can occur at the stage when options are being considered and compared as candidates for implementation. We will look at *ex ante* evaluations shortly in this chapter, but for now, think of them as evaluations that assess program or policy options *before* any are selected for implementation.

The third phase in the cycle is about implementation. This phase involves building or adapting organizational structures and processes to facilitate implementing policies or programs. One perspective on organizations is that they can be viewed primarily as instruments—means by which policy and program objectives (ends) are achieved. Implementation-focused evaluations can occur in conjunction with the implementation phase of the cycle. In this textbook, we will look at **formative evaluations** as a type of implementation-related evaluation. Formative evaluations are discussed later in this chapter. Typically, implementation evaluations assess the extent to which intended program or policy designs are successfully implemented by the organizations that are tasked with doing so. Implementation is not the same thing as outcomes/results. Weiss (1972) and others have pointed out that assessing implementation is a necessary condition to being able to evaluate the extent to which a program has achieved its intended outcomes. Bickman (1996), in his seminal evaluation of the Fort Bragg Continuum of Care Program, makes a point of assessing how well the program was implemented, as part of his evaluation of the outcomes. It is possible to have implementation failure, in which case any observed outcomes cannot be attributed to the program. Implementation evaluations can also examine the ways that existing organizational structures, processes, and cultures either facilitate or impede **program implementation**.

The fourth phase in the cycle is about monitoring performance, assessing performance results, evaluating, and reporting. Monitoring with performance measures is an important way to tell how a program is tracking over time. Performance data can be useful for evaluations, as well as for making management-related decisions. This phase is also about **summative evaluation**, that is, evaluation that is aimed at answering questions about a program or policy achieving its intended results, with a view to making decisions about the future of the program. We will discuss summative evaluations more thoroughly later in this chapter.

"Performance measurement and reporting" is expected to contribute to "real consequences" for programs. Among these consequences are 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. Finally, strategic objectives are revisited, and the evidence from earlier phases in the cycle is among the inputs that may result in new or revised objectives—usually through another round of a strategic planning process.

In this book, the performance management cycle illustrated in Figure 1.1 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. The "performance measurement and reporting" part of the cycle is central to this textbook, but we take the view that that phase of the performance management cycle is about *evaluation and reporting.* Evaluation includes both program evaluation and performance measurement, and we build a foundation in the early chapters of the textbook that can be used for both approaches to evaluating programs and policies.

Chapter 6 on **needs assessments** builds on topics covered in Chapter 4 (measurement) and can occur in several phases of 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 policies or 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 can be 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.

WHAT ARE PROGRAMS AND POLICIES?

As you have been reading this chapter, you will have noticed that we mention both *policies* and *programs* as candidates for performance measurement and evaluation. Our view is that the methodologies that are discussed in this textbook are generally appropriate for evaluating *both* programs and policies. Some analysts use the terms interchangeably—in some countries, policy analysis and evaluation is meant to encompass program evaluation (Curristine, 2005). We will define them both so that you can see what the essential differences are.

What Is a Policy?

Policies connect means and ends. The cores of policies are statements of intended outcomes/objectives (ends) and the means by which government(s) or their agents (perhaps nonprofit organizations or even private sector companies) will go about achieving these outcomes. Policy objectives usually reflect the political objectives and values of the government of the day. These objectives can be expressed in election platforms, political speeches, government responses to questions by the media, or other announcements. Ideally, before a policy is created or announced, research and analysis has been done that establishes the feasibility, the estimated effectiveness, or even the anticipated cost-effectiveness of proposed strategies to address a problem or issue. Often, new policies are modifications of existing policies that expand, refine, or reduce existing governmental activities.

When governments make a policy, there are usually stages, beginning with assessing the need or demand for a policy (perhaps informally), through to implementation and to evaluation of the extent to which the policy was successful. These stages can be more or less formal—sometimes a problem is examined in depth before any policies are drafted. Royal Commissions (in Canada), task forces, reports by independent bodies, or even public inquiries (Congressional hearings, for example) are ways that in-depth reviews can set the stage for developing or changing public policies. In other cases, announcements by elected officials addressing a perceived problem can serve as the impetus to develop a policy.

An example of a policy that has significant planned impacts is the British Columbia Government's November 2007 legislation that committed the provincial government to reducing greenhouse gas emissions in the province by 33% by 2020. The legislation also

states that by 2050, greenhouse gas emissions will be 80% below 2007 levels. Reducing greenhouse gas emissions in British Columbia will be challenging. A Climate Action Committee of experts and other stakeholders was formed to assist the government in coming up with ways to achieve the policy outcome. The Committee suggested an array of additional policies and more specific **programs** to the government, all aimed at increasing the likelihood that the policy will be successful.

What Is a Program?

To reduce greenhouse gases in British Columbia, many different programs will be required—some targeting the government itself, others targeting industries, consumers, and other governments (e.g., British Columbia local governments). Programs to reduce greenhouse gases will be concrete expressions of the policy. Policies are usually higher level statements of intent—they need to be translated into programs of actions to achieve intended outcomes. Policies generally enable programs. In the British Columbia example, one of the programs that was implemented starting in 2008 was a tax on the carbon content of all fuels used in British Columbia by both public and private sector emitters, including all who drive vehicles in the province.

Programs are similar to policies—they are means–ends chains that are intended to achieve some agreed-on objective(s). 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 complicated—an addiction treatment program might have a broad range of activities, from public advertising, through intake and treatment, to referral, and finally to follow-up.

Increasingly, programs can involve several levels of government or governmental agencies and nonprofit organizations. These kinds of programs are challenging for evaluators and have prompted some in the field to suggest alternative ways of assessing program processes and outcomes. Michael Patton (1994, 2011) has introduced **developmental evaluation** as one approach, and John Mayne (2001, 2011) has introduced **contribution analysis** as a way of addressing attribution questions in complex program settings.

In the chapters of this textbook we will introduce multiple examples of both policies and programs, and the evaluative approaches that have been used for them. A word on our terminology—although we intend this book to be useful for both program evaluation and policy evaluation, we will refer mostly to program evaluations.

THE PRACTICE OF PROGRAM EVALUATION: THE ART AND CRAFT OF FITTING ROUND PEGS INTO SQUARE HOLES

One of the principles underlying this book is the importance of exercising professional judgment as program evaluations are designed, executed, and acted on. Michael Scriven has defined evaluation as judging the merit and worth of programs (Lincoln & Guba, 1980; Scriven, 1972), where merit is an intrinsic judgment of the absolute value of a program in

terms of general normative criteria, and worth is a judgment based on success in achieving program objectives. The methodological tools we learn, and the pluses and minuses of applying them in 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 assumptions specified for their application. As evaluators, we learn to adapt the tools we know, given our training and experience, to the uniqueness of each evaluation setting. In some situations, we find that no approach we know quite fits the circumstances, so we improvise.

Our tools are indispensable—they help us 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. Some of what we bring with us to an evaluation is **tacit knowledge**—it is knowledge based on our experience, and it is not learned or communicated except by experience.

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 will 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 in which control groups are not feasible, and when resource constraints limit the kinds of data we can gather. Or, we are confronted by a situation in which the evaluation design that we had developed in consultation with stakeholders is undermined by the evaluation 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 the 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 above-mentioned 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 the 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 they produced ambiguous conclusions while still being costly (Basilevsky & Hum, 1984).

Now, there is no one dominant view of "correct" evaluation methods, notwithstanding the continued debate between proponents of randomized controlled trials as the methodological gold standard and those who believe methodologies must be situation specific and more eclectic (Cook et al., 2010). Indeed, evaluation methods that rely on the collection and analysis of spoken and written words were born out of a strong reaction to the insular and sometimes remote evaluations produced by social experimenters. Qualitative evaluators such as Egon Guba and Yvonna Lincoln (1989) offer a deep critique of quantitative approaches and advocate for the use of methods that emphasize understanding and working with the subjectivity that is inherent in all human interactions.

Michael Patton (2008) has taken a pragmatic view of mixing qualitative and quantitative evaluation approaches. His goal, which is widely reflected in the field, is that if we want our work to be used, we need to conduct evaluations that engage users in the evaluation process in ways that encourage them 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. Evaluation as a field has moved toward a more pluralistic view of what is appropriate in particular evaluation settings.

The stance taken in this book, and reflected in the contents of the chapters, is that program evaluation practice is rich and very diverse. Key to understanding all evaluation practice is accepting that no matter how sophisticated our designs, measures, and other methods are, we will exercise professional judgment in our work. In this book, we will see where professional judgment is exercised in the evaluation process, and will begin to learn how to make defensible judgments. Chapter 12 is devoted to the nature and practice of professional judgment in evaluation.

Some readers may have concluded by now that we are condoning an "anything goes" attitude. Readers will discover, instead, that we have taken a structured approach to evaluations that relies on understanding the range of tools that have been developed in and for the profession and are applying them in ways that improve (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—whether the intended outcomes were realized and why—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, Vancouver, British Columbia, implemented a Neighbourhood Integrated Service Team (NIST) program. The NIST program 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 Can$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. The NIST program was intended to address this problem and deal with situations like Carolina Street before they became expensive and politically embarrassing.

The NISTs 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 the NIST program, it 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 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 the NIST program, 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 the NIST program. Since she had been principally responsible for designing and implementing it, 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 limited resources: 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: (1) NIST committee members, (2) department staffs, (3) the city council, and (4) the community.

She used a combination of individual interviews and **focus groups** to gather responses from 48 NIST team members, 24 departmental staff members (three from each of the eight departments involved in the NIST program), four members of the 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 used written notes and, in some cases, tape recordings to record responses, observations, and her own reflections on the information she was gathering.

Data analysis involved doing **content analysis** of interview notes, identifying common ideas in the material she had recorded, and organizing all the information into themes. Four main categories of themes emerged: (1) areas where the program was reported to be functioning well, (2) areas where there was room for improvement, (3) stakeholder recommendations, and (4) "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 the NIST program 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 the NIST program into the city departments. Stakeholders generally

felt that although the program was a good thing and was making a difference, it was not clear how team members were expected to balance their accountability to the NIST program and to their home departments. The NIST program needed to be reflected in department business plans, acknowledging its continued role in city service delivery, and needed stronger leadership to advocate the program within city departments.

Since this evaluation was completed, the NIST program has won a United Nations Award for Innovation in the Public Service. It has been widely recognized as a model for horizontal integration of local government administrative departments.

In addition, the city has taken the lead in 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 Vancouver's Downtown Eastside neighborhood, a neighborhood that some argue has been the single most difficult challenge for regional 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, the NIST program 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 the program was put into place, the main concern was dealing with the problem of the Carolina Street house in Mt. Pleasant and others like it. Little attention was paid to how the program would be evaluated. The evaluation was grounded in the specific concerns of a senior manager who wanted answers to questions about the NIST program that were being raised by key 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 also typical in that it 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 with other, non-NIST communities. Other noteworthy points are as follows:

- The evaluation relied on **multiple independent lines of evidence** from different points of view with respect to the program, and used these perspectives to help answer the questions that motivated the study. The evaluator has taken a pragmatic stance about combining qualitative and quantitative lines of evidence; that is, if a particular approach works in the evaluation, use it.
- 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 and continues to exist today.

- 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.

KEY CONCEPTS IN PROGRAM EVALUATION

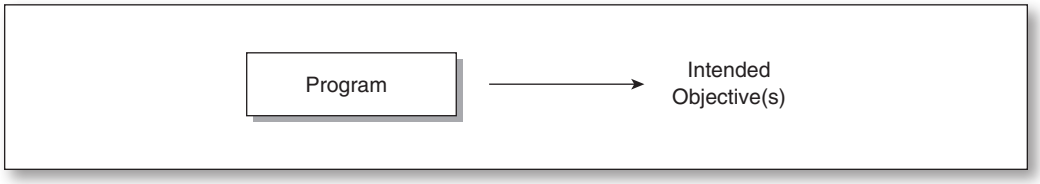
One of the key questions that many program evaluations are expected to address can be worded as follows:

- To what extent, if any, were the intended objectives met?

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 intended objectives, which we

Figure 1.2 Linking Programs and Intended Objectives



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 horizontal causal link between the program and its outcomes has been modified in two ways: (1) intended outcomes has been replaced by the **observed outcomes** (what we actually observe when we do the evaluation), and (2) a question mark (?) has been placed over that causal arrow.

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

- To what extent, if at all, was the program responsible for 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 *the program* was *the cause* or *a 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?

We examine the attribution question in some depth in Chapter 3, 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 The Two Program Effectiveness Questions Involved in Most Evaluations

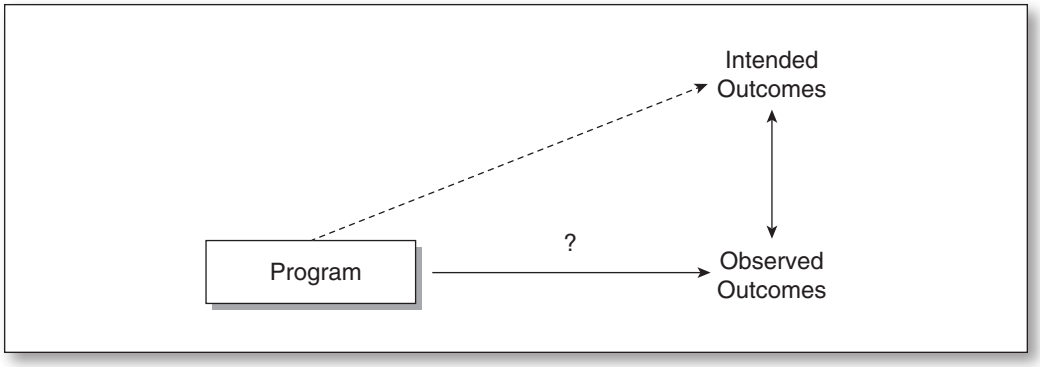


Figure 1.3 also raises a second evaluation question:

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

Here, we are comparing what we actually find with 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 "achievement of intended outcomes" 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 actual results are consistent with intended outcomes. In Figure 1.3, the dashed arrow connects the program to the intended outcomes, and assessments of that link are often a focus of performance measurement systems. Where benchmarks or performance targets have been specified, comparisons between actual outcomes and intended outcomes can also be made, but what is missing from such comparisons is an assessment of the extent to which observed and intended outcomes are attributable to the program (McDavid & Huse, 2006).

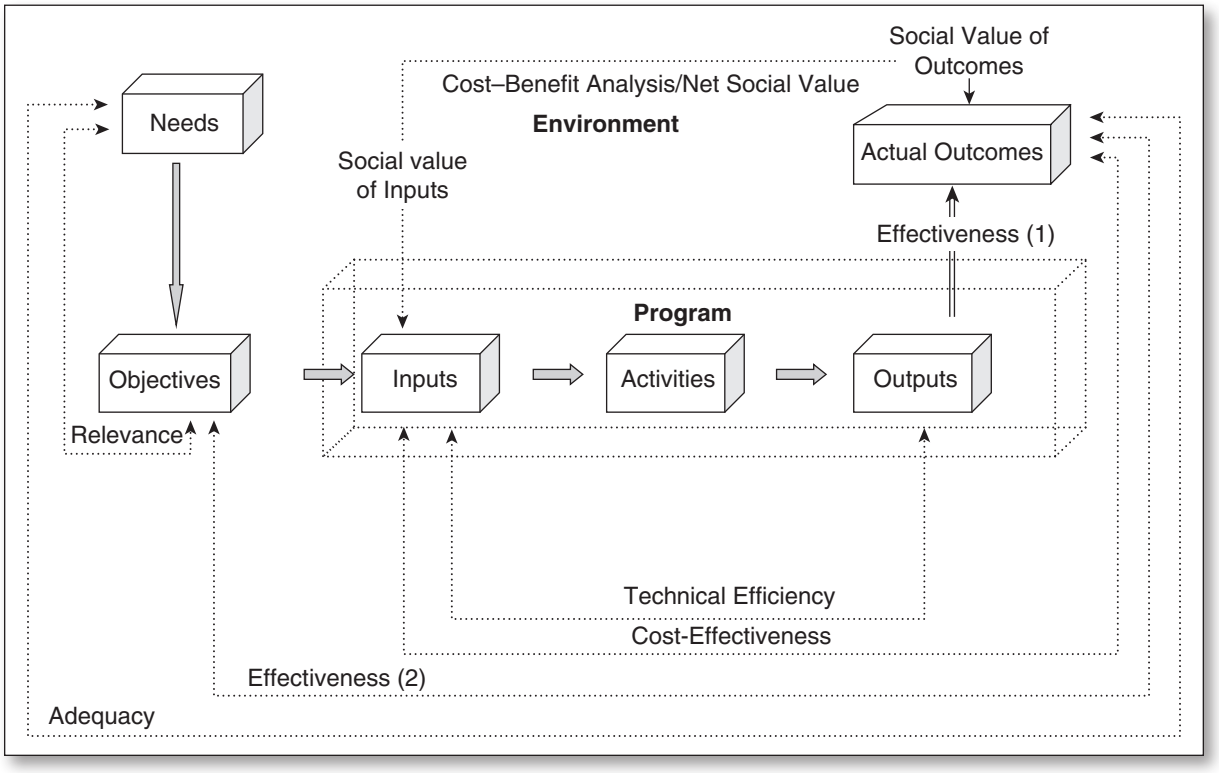
TEN KEY EVALUATION QUESTIONS

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 can address. They are summarized in Table 1.1. To help us make sense of these 10 questions, we have included an open systems model (Figure 1.4) of a typical program that shows how objectives, resources (inputs), outputs, and outcomes are linked. You can review that model, locate the key words that are highlighted in Table 1.1, and see how the questions are related to each other.

Table 1.1 Ten Possible Evaluation Questions

1. What is the **need** for a program?
2. Is the program **relevant**?
3. Was the structure/logic of the program **appropriate**?
4. Was the program implemented as intended?
5. Was the program **technically efficient**?
6. Was the program responsible for the outcomes that actually occurred (**effectiveness 1**)?
7. Did the program achieve its intended objectives (**effectiveness 2**)?
8. Was the program **cost-effective**?
9. Was the program **cost beneficial**?
10. Was the program **adequate**?

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



Source: Adapted from Nagarajan and Vanheukelen (1997, p. 25).

1. *What is the need for a program?* A needs assessment for a program can occur either before options are developed (an *ex ante* needs assessment) or during its implemented lifetime (*ex post* needs assessment). Typically, needs assessments gather information using either or both qualitative and quantitative methodologies, and compare existing programs or services with levels and types of needs that are indicated by the data. These comparisons can suggest gaps that might be addressed by developing or modifying programs, and allocating resources to reduce or eliminate these gaps.

Needs assessment done before a program is developed can inform the way that the objectives are stated, and suggest targets that would reduce needs gaps. If a needs assessment is done during the time a program is implemented, it can be a part of an evaluation of the program's effectiveness—is the program achieving its intended outcomes, *and* does the program meet the needs of the stakeholder groups at which it was targeted? Such an evaluation might suggest ways of improving the existing program, including refocusing the program to better meet client needs. We will be discussing needs assessments in Chapter 6 of this textbook.

2. *Is the program relevant?* Programs are aimed at objectives that are intended to reflect priorities of government, boards of directors, or other stakeholders. These priorities can change. Governments change, and differing views on social, economic, or political issues

emerge that suggest a need to reassess priorities and either adjust direction or embark on a new course. Programs that were consistent with government or other stakeholder priorities at one point can become less relevant over time.

Assessing the **relevance** of a program typically involves examining documents that outline the original (and current) directions of the program, on the one hand, and comparing those with statements of current and future priorities, on the other. Interviews with key stakeholders are usually an important part of relevance assessments. Assessing the relevance of a program is different from assessing the need for a program or measuring its effectiveness—assessments of relevance are almost always qualitative and rely substantially on the experience and judgment of the evaluators as well as of stakeholders.

3. *Was the structure/logic of the program appropriate?* Typically, programs address a problem or issue that has arisen in the public sector. The scope and reach of programs can vary a great deal, depending on the complexity of the problem. When programs are being developed, researching options is useful. This often involves comparisons among jurisdictions to see whether/how they have tackled similar problems and whether they have information about the success of their strategies.

Selecting a strategy to address a problem is constrained by time, available resources, and prevailing political views. Proposed solutions (programs) can be a compromise of competing organizational/stakeholder views, but this may not be the most appropriate means to achieving a desired objective.

Assessing the appropriateness of a program focuses on the structure that is intended to transform resources into results. Related questions include the following:

- Does the logic of the program reflect evidence-based theories of change that are relevant for this situation?
- Does the logic of the program reflect smart or promising practices in other jurisdictions?
- Is the logic of the program internally consistent?
- Are all the essential components there, or are there one or more components that should be added to increase the likelihood of success?
- Overall, is the logic/design the best feasible means to achieve the objectives?

We discuss program theories and program logics in Chapter 2.

4. *Was the program implemented as intended?* This is the one question in Table 1.1 that is not reflected in Figure 1.4. Assessing implementation involves an examination of the **program components**, their **activities**, and the **outputs** from those activities. Programs or policies are implemented in environments that are affected by, and can affect, the program. Program objectives drive the design and implementation process; inputs (typically budgetary resources, human resources, and technologies) are converted into activities that in turn produce outputs. These are explained in greater detail in Chapter 2.

Programs can consist of several components, and each is associated with a stream of activities and outputs. For example, a program that is focused on training unemployed persons so that they can find permanent jobs may have a component that markets the program to prospective clients, a component wherein the actual training is offered, a component that features activities intended to connect trained persons with prospective

employers, and a component that follows up with clients and employers to solve problems and increase the likelihood that job placements are successful.

Assessing such a program to see whether it has been fully implemented would involve looking at each component, assessing the way that it had been implemented, identifying and describing any bottlenecks in the process, and summarizing the outputs that had been produced. Since the outputs of most programs are necessary (but not sufficient) to produced outcomes, tracking outputs as part of measuring program performance monitors program implementation and provides information that is an essential part of an implementation evaluation.

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, two 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.

5. *Was the program technically efficient?* **Technical efficiency** involves comparing inputs with outputs, usually to assess the productivity of the program or to calculate the costs per unit of output. For example, most hospitals calculate their cost per patient day. This measure of technical efficiency compares the costs of serving patients (clients) with the numbers of clients and the time that they (collectively) spend in the hospital. If a hospital has 100 beds, it can provide a maximum of 36,500 (100 × 365) patient days of care in a year. Administrative and resource-related constraints would typically reduce such a maximum to some fraction of that number.

Knowing the expenditures on patient care (calculating this cost can be challenging in a complex organization like a hospital), and knowing the actual number of patient days of care provided, it is possible to calculate the cost of providing a unit of service (cost per patient day). An additional indicator of technical efficiency is the comparison of the actual cost per patient day with an ideal cost per patient day if the hospital were fully utilized.

6. *Was the program responsible for the outcomes that actually occurred?* **Effectiveness (1)** in Figure 1.4 focuses on the linkage between the program and the *outcomes that actually happened.* The question is whether the observed outcomes were due to the program or, instead, were due to some combination of environmental factors other than the program. In other words, can the observed outcomes be attributed to the program? We discuss the attribution issue in Chapter 3.

7. *Did the program achieve its intended objectives?* **Effectiveness (2)** in Figure 1.4 compares the program objectives with the outcomes that actually occurred. Attaining the intended outcomes is *not* equivalent to saying that the program caused these outcomes. It is possible that shifts in environmental factors accounted for the apparent success (or lack of it) of the program. An example of environmental factors interfering with the evaluation of a program in British Columbia occurred in a province-wide program to target drinking drivers in the mid-1970s. The Counterattack Program involved public advertising, roadblocks,

vehicle checks, and 24-hour license suspensions for persons caught with alcohol levels above the legal blood alcohol limit. A key measure of success was the number of fatal and injury accidents on British Columbia provincial highways per 100 million vehicle miles driven—the expectation being that the upward trend prior to the program would be reversed after the program was implemented. Within five months of the beginning of that program, British Columbia also adopted a mandatory seat belt law, making it impossible to tell whether Counterattack was responsible (at a province-wide level) for the observed downward trend in accidents that happened.

Performance measures are often intended to track whether policies and programs achieve their intended objectives (usually, yearly targets are specified). Measuring performance is not equivalent to evaluating the effectiveness of a program or policy. Achieving intended outcomes does not tell us whether the program or policy in question caused those outcomes. If the outcomes were caused by factors other than the program, the resources that were expended were not used cost-effectively.

8. *Was the program cost-effective?* Cost-effectiveness involves comparing the costs of a program with the outcomes. *Ex post* (after the program has been implemented) cost-effectiveness analysis compares actual costs with actual outcomes. *Ex ante* (before implementation) cost-effectiveness analysis compares expected costs with expected outcomes. The validity of *ex ante* cost-effectiveness analysis depends on how well costs and outcomes can be forecasted. Cost-effectiveness analyses can be conducted as part of assessing the effectiveness of the policy or program. Ratios of costs per unit of outcome offer a way to evaluate a program's performance over time, or compare a program with other similar programs elsewhere, or compare program performance with some benchmark (Yeh, 2007).

Key to conducting a cost-effectiveness evaluation is identifying an outcome that represents the program well (validly) and can be compared with costs quantitatively to create a measure of unit costs. An example of a cost-effectiveness ratio for a program intended to place unemployed persons in permanent jobs would be cost per permanent placement.

There is an important difference between technical efficiency and cost-effectiveness. Technical efficiency compares the cost of inputs with units of outputs, whereas cost-effectiveness compares the cost of inputs with units of outcomes. For example, if one of the components of the employment placement program is training for prospective workers, a measure of the technical efficiency (comparing costs with units of output) would be the cost per worker trained. Training could be linked to permanent placements, so that more trained workers would presumably lead to more permanent placements (an outcome). Cost-effectiveness is discussed in Chapter 7.

9. *Was the program cost-beneficial?* Cost–benefit analysis compares the costs and the benefits of a program. Unlike technical efficiency or cost-effectiveness analysis, cost–benefit analysis converts all the outcomes of a program into monetary units (e.g., dollars), so that costs and benefits can be compared directly. Typically, a program or a project will be implemented over several years, and expected outcomes may occur over a longer period of time. For example, when a cost–benefit analysis of a hydroelectric dam is being conducted, the costs and the benefits would be spread out over a long period of time, making it necessary to take into account when the expected costs and benefits occur, in any calculations of total costs and total benefits.

In many public sector projects, particularly those that have important social dimensions, converting outcomes into monetary benefits is difficult and often necessitates assumptions that can be challenged.

Cost–benefit analyses can be done *ex ante* or *ex post*, that is, before a program is implemented or afterward. *Ex ante* cost–benefit analysis can indicate whether it is worthwhile going ahead with a proposed option, but to do so, a stream of costs and outcomes must be assumed. If implementation problems arise, or the expected outcomes do not materialize or unintended impacts occur, the actual costs and benefits can diverge substantially from those estimated before a program is implemented. Cost–benefit analysis is the subject of Chapter 7.

10. *Was the program adequate?* Even if a program was technically efficient, cost-effective, and even cost-beneficial, it is still possible that the program will not resolve the problem for which it was intended. An evaluation may conclude that the program was efficient and effective, but the magnitude of the problem was such that the program was not **adequate** to achieve the overall objective.

Changes in the environment can affect the adequacy of a program. A program that was implemented to train unemployed persons in resource-based communities might well have been adequate in an expanding economy, but if macroeconomic trends reverse, resulting in the closure of mills or mines, the program may no longer be sufficient to address the problem at hand.

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, they can be very useful in planning or revising programs. We discuss needs assessments in Chapter 6.

Formative and Summative Evaluations

Michael Scriven (1967) introduced the original distinction between formative and summative evaluations (Weiss, 1998). Since then, he has come back to this issue several more times (e.g., Scriven, 1991, 1996). Scriven's definitions reflected his distinction between implementation issues and evaluating program effectiveness. He 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 (the worth of a program). Scriven also emphasized assessments of the merit of programs—their intrinsic value, given our social, and even democratic, values.

Although Scriven's (1967) distinction between formative and summative evaluations has become a part of any evaluator's vocabulary, it has been both elaborated and challenged by others in the field. Chen (1996) introduced a framework that featured two evaluation purposes—improvement and assessment—and two program stages—process and outcomes. His view was that many evaluations are mixed, that is, evaluations can be both formative and summative, making Scriven's original dichotomy incomplete. For Chen (1996), improvement was formative and assessment was summative—an evaluation that is looking

to improve a program can be focused on both implementation and objectives achievement. The same is true for evaluations that are aimed at assessing programs.

Scriven's (1967) original definitions do not generally reflect the way program evaluation is practiced today. 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 the program was, and how effective the program was. A focus on **program processes** is combined with concerns about whether the program was achieving its intended objectives.

In this book, we will refer to formative and summative evaluations but will define them in terms of their *intended uses.* This is similar to the distinction offered in Weiss (1998) and Chen (1996). Formative evaluations are *intended* to provide feedback and advice with the goal 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? 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.

In addition to formative and summative evaluations, others have introduced several other classifications for evaluations. Eleanor Chelimsky (1997), for example, makes a similar distinction to the one we make between the two primary types of evaluation, which she calls (1) evaluation for development (i.e., the provision of evaluative help to strengthen institutions and to improve organizational performance) and (2) 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). Patton's (1994, 2011) "developmental evaluation" is another approach, related to ongoing organizational learning in complex settings, that differs in some ways from the formative and summative approaches generally adopted for this textbook.

EX ANTE AND *EX POST* EVALUATIONS

Typically, 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 Chapters 3 and 5, we will learn about the research design options and qualitative evaluation alternatives that are available for such situations. Chapter 5 also looks at mixed-methods designs for evaluations.

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 facilitate 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 with the existing "no program" status quo) and the evaluation results used as a kind of *ex ante* evaluation of a broader implementation or scaling up of the program.

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 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.

CAUSALITY IN PROGRAM EVALUATIONS

In this textbook, a key theme is the evaluation of the effectiveness of programs. One aspect of that issue is whether the program caused the observed outcomes. Our belief is that program effectiveness and in particular attribution of observed outcomes are the core issues in evaluations. In fact, that is what distinguishes program evaluation from other, related, professions such as auditing and management consulting. Picciotto (2011) points to the centrality of program effectiveness as a core issue for evaluation as a discipline/profession:

> What distinguishes evaluation from neighboring disciplines is its unique role in bridging social science theory and policy practice. By focusing on whether a policy, a program or project is working or not (and unearthing the reasons why by attributing outcomes) evaluation acts as a transmission belt between the academy and the policy-making. (p. 175)

In Chapter 3, we will describe the logic of research designs and how they can be used to examine causes and effects in evaluations. Briefly, there are three conditions that are widely accepted as being jointly necessary to establish a causal relationship between a program and an observed outcome: (1) the program has to precede the observed outcome, (2) the presence or absence of the program has to be correlated with the presence or absence of the observed outcome, and (3) there cannot be any plausible rival explanatory factors that could account for the **correlation** between the program and the outcome.

In the evaluation field, how to assess causality is controversial. One view is that to really pin down the three causal conditions, we need to construct and implement evaluation designs that incorporate random assignment of prospective program recipients to program and control groups and compare the outcome variables for the two groups, once the program has been implemented. This view dominated the field in the 1960s and the 1970s but has been challenged principally around the "one-size-fits-all" approach that is implied by privileging experimental designs. Different approaches to assessing causal relationships have been proposed, and the debate around using experimental designs continues (Cook et al., 2010). Our view is that the *logic* of causes and effects (the three necessary conditions) is important to understand, if you are going to do program evaluations. Looking for plausible

rival explanations for observed outcomes is important for any evaluation that claims to be evaluating program effectiveness. But that does not mean that we have to have experimental designs for every evaluation.

Program evaluations are often conducted under conditions in which data appropriate for ascertaining or even systematically addressing the attribution question are hard to come by. 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 (2008) suggests that our capacity to observe and detect **causal relationships** is built into us. We are hardwired 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 STEPS IN CONDUCTING A PROGRAM EVALUATION

Our approach to presenting the key topics in this book is that an understanding of program evaluation concepts and principles is important *before* designing and implementing performance measurement systems. When performance measurement expanded across government jurisdictions in the 1990s, expectations were high for this new approach (McDavid & Huse, 2012). In many organizations, performance measurement was viewed as a replacement for program evaluation (McDavid, 2001; McDavid & Huse, 2006). Two decades of experience with actual performance measurement systems suggests that initial expectations were unrealistic. Relying on performance measurement alone to evaluate programs does not get at why observed results occurred. Performance measurement systems monitor—evaluations are intended to answer the "why" question.

In this chapter, we will outline how program evaluations in general are done, and once we have covered the core evaluation-related knowledge and skills in Chapters 2, 3, 4, and 5, we will turn to performance measurement in Chapters, 8, 9, and 10. In Chapter 9, we will outline the key steps involved in designing and implementing performance measurement systems.

Even though each evaluation is different, it is useful to outline the steps that are generally typical, keeping in mind that for each evaluation, there will be departures from these steps. Our experience with evaluations is that as each evaluation is designed and conducted, the steps in the process are revisited in an iterative fashion. For example, the process of constructing a logic model of the program may result in clarifying or revising the program objectives and even prompt revisiting the purposes of the evaluation, as additional consultations with stakeholders take place.

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. It is worth mentioning that the steps outlined below imply that a typical program evaluation is a project, with a beginning and an end point. This is still the mainstream view of evaluation as a profession, but there are others who argue that evaluation should be more than "studies." Mayne and Rist (2006), for example, suggest that evaluators should be prepared to do more than evaluation projects. Instead, they need to be more engaged in organizational management: leading the development of results-based management systems (including performance measurement and performance management systems), and using all kinds of evaluative information, including performance measurement, to strengthen the evaluative capacity in organizations. They maintain that creating and using evaluative information has to become more real-time, and that managers and evaluators need to think of each other as partners in constructing knowledge management systems and practices. Patton (2011) takes this vision even further—for him, developmental evaluators in complex settings need to be engaged in organizational change, using their evaluation knowledge and skills to provide real-time advice that is aimed at organizational innovation and development.

Table 1.2 Checklist of Key Questions and Steps in Conducting Evaluation Feasibility Assessments and Evaluation Studies

Steps in assessing the feasibility of an evaluation

1. Who are the clients for the evaluation?
2. What are the questions and issues driving the evaluation?
3. What resources are available to do the evaluation?
4. Given the evaluation questions, what do we already know?
5. What is the logic of the program?
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 feasible?
8. What data sources are available and 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 to 8, which evaluation strategy is most feasible, and defensible?
10. Should the evaluation be undertaken?

Steps in conducting and reporting an evaluation

1. Develop the data collection instruments and pretest them.
2. Collect data/lines of evidence that are appropriate for answering the evaluation questions.
3. Analyze the data, focusing on answering the evaluation questions.
4. Write, review, and finalize the report.
5. Disseminate the report.

Table 1.2 (page 27) summarizes 10 questions that are important as part of evaluation feasibility assessments. Assessing the feasibility of a proposed evaluation project, and making a decision about whether to go ahead with it, is a strategy that permits several decision points before the budget for an evaluation is fully committed. A sound feasibility assessment will yield products that are integral to a defensible evaluation product.

The end product of the feasibility assessment phase entails the aggregation of enough information and detail that it should be straightforward to implement the evaluation project. In Chapter 6, when we discuss needs assessments, we will see that there is a similar assessment phase for planning needs assessments.

Five additional steps are also outlined in Table 1.2 for conducting and reporting evaluations. Each of the questions and steps is elaborated in the discussion that follows.

Assessing the Feasibility of the Evaluation

1. *Who are the clients for the evaluation?* Program evaluations are substantially *user driven.* Michael Patton (2008) makes **utilization** a key criterion in the design and execution of program evaluations. Intended users must be identified early in the process and must be involved in the evaluation feasibility assessment. The extent of their involvement will depend on whether the evaluation is intended to make incremental changes to the program, or instead is intended to provide information that affects the existence of the program. Possible clients could include, but are not limited to,

- program/policy managers,
- agency/ministry executives,
- external agencies (including central agencies),
- program recipients,
- funders of the program,
- political decision makers/members of governing bodies (including boards of directors), and
- community leaders.

All evaluations are affected by the interests of stakeholders. Options for selecting what to evaluate, who to report the results to, how to collect the information, and even how to interpret the data generally take into account the interests of key stakeholders. In most evaluations, the evaluation's clients will have some influence over 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 negotiation process that surrounds the evaluation itself. Indeed, as 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).

An evaluation plan, outlining items such as the purpose of the evaluation, the key evaluation questions, and the intended audience(s), worked out and agreed to by the evaluators and the clients prior to the start of the evaluation, is 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?* 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 policy or program evaluated?
- What are the main evaluation issues that they want addressed (combinations of the 10 evaluation questions summarized in Table 1.1 are usually in play)?
- Is the evaluation intended to be for incremental adjustments/improvements or major decisions about the future of the program, or both?

Answering these 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 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 issues of all program stakeholders within one evaluation. For this reason, the evaluator must reach a firm agreement with the evaluation clients about the questions to be answered by the evaluation. This process will involve working with the clients to help narrow the list of questions they are interested in, a procedure that 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?* Typically, resources to design and complete evaluations are scarce. Greater sophistication in evaluation designs almost always entails larger expenditures. For example, achieving the necessary control over the program and its environment to conduct experimental or quasi-experimental evaluations generally entails modifying existing administrative procedures and perhaps even temporarily changing or suspending policies (e.g., to create no-program comparison groups).

It is useful to distinguish among several kinds of resources needed for 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

It is possible to construct and implement evaluations with very modest resources. One of the authors was involved in the evaluation of a homeless shelter pilot program in Victoria, British Columbia, that had a total program budget of less than Can$70,000 for the six weeks it operated. The total budget for the evaluation was Can$5,000. Bamberger, Rugh, Church, and Fort (2004) have suggested strategies for designing impact evaluations with very modest resources—they call their approach **shoestring evaluation**. Agreements reached about all resource requirements should form part of the written evaluation plan.

4. *What evaluation work has been done previously?* Evaluators should take advantage of work that has already been done. There may be previous evaluations of the current program or evaluations of similar ones in other jurisdictions. Internet resources are very useful as you are planning an evaluation, although many program evaluations are unpublished and may be available only through direct inquiries.

Aside from literature reviews, which have been a staple of researchers for as long as empirical work has been done, there is growing emphasis on approaches that take advantage of the availability of reports, articles, and other documents on the Internet. An example of a **systematic review** was the study done by Anderson, Fielding, Fullilove, Scrimshaw, and Carande-Kulis (2003) that focused on cognitive outcomes for early childhood programs in the United States. Anderson and her colleagues began with 2,100 possible publications and, through a series of filters, narrowed those down to 12 studies that included comparison group research designs, were robust in terms of their internal validity, and measured cognitive outcomes for the programs being evaluated.

The Cochrane Collaboration (Higgins & Green, 2011) is an international project begun in 1993 that is aimed at conducting systematic reviews of health-related interventions. These reviews can be useful inputs for governments and organizations that want to know the aggregate effect sizes for interventions using randomized controlled trials that have been grouped and collectively assessed.

The Campbell Collaboration (2010) is an organization that is focused on the social sciences and education. Founded in 1999, its mission is to help people "make well-informed decisions by preparing, maintaining and disseminating systematic reviews in education, crime and justice, and social welfare."

The Government Social Research Unit in the British government has published a series of guides, including *The Magenta Book: Guidance Notes for Policy Evaluation and Analysis* (Government Social Science Research Unit, 2007). Chapter 2 in *The Magenta Book*, "What Do We Already Know?" focuses on using existing research in policy evaluations. Literature reviews and quantitative and qualitative systematic reviews are covered. The main point is that research is costly, and being able to take advantage of what has already been done can be a cost-effective way to construct lines of evidence in an evaluation.

An important issue in synthesizing previous work is how comparable the studies are. Variations in research designs, the ways that studies have been conducted (the precise research questions that have been addressed), the sizes of samples used, and the measures that have been selected will all influence the comparability of previous studies and the validity of any aggregate estimates of policy or program effects.

5. *What is the structure of the program?* Programs are **means–ends relationships**. Their intended objectives, which are usually a product of organizational/political negotiations, are intended to address problems or respond to social/economic/political issues or needs that emerge from governments, interest groups, and other stakeholders. Program structures are the means by which objectives are expected to be achieved.

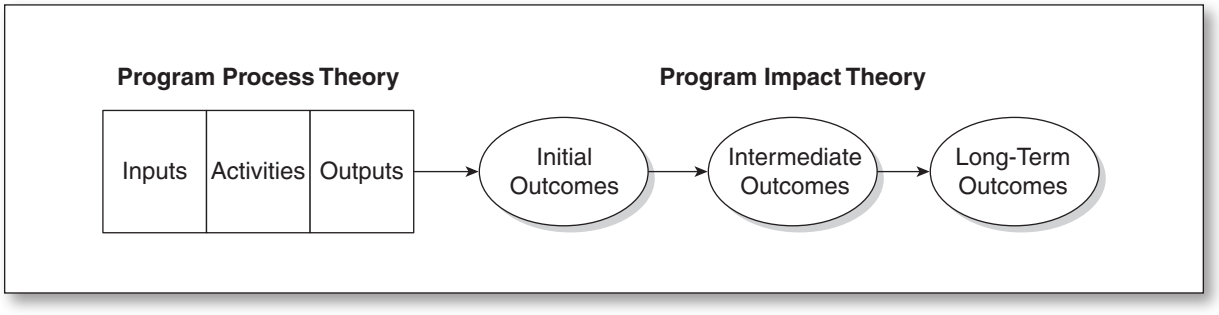
Logic models are useful for visually summarizing the structure of a program. They are a part of a broader movement in evaluation to develop and test program theories when doing evaluations (Coryn, Schröter, Noakes, & Westine, 2011). **Program logic** models are widely used to show the intended **causal linkages** in a program. There are many different styles of logic models (Funnell & Rogers, 2011) but what they have in common is identifying the major sets of activities in the program, their intended outputs, and the outcomes (often short, medium, and longer term) that are expected to flow from the outputs (Knowlton & Phillips, 2009).

An example of a basic schema for a logic model is illustrated in Figure 1.5. The model shows the stages in a typical logic model: program process (including outputs) and outcomes. We will be discussing logic models in some detail in Chapter 2 of this textbook.

Logic models are usually about *intended* results—they outline how a program is expected to work, if it is implemented and works as planned. Key to constructing a logic model is a clear understanding of the program objectives. One challenge for evaluators is working with stakeholders, including program managers and executives, to refine the program objectives. Ideally, program objectives should have five characteristics:

1. An expected direction of change for the outcome is specified.
2. An expected magnitude of change is specified.
3. An expected time frame is specified.
4. A target **population** is specified.
5. The outcome is measurable.

Figure 1.5 Linear Program Logic Model



Source: Coryn, Schröter, Noakes, and Westine (2011), as adapted from Donaldson (2007, p. 25).

The Government's stated objective of reducing greenhouse gas emissions in British Columbia by 33% by the year 2020 is a good example of a clearly stated policy objective. From an evaluation standpoint, having an objective that is clearly stated simplifies the task of determining whether the policy has achieved its intended outcome. Political decision makers often prefer more general language in program or policy objectives so that there is "room" to interpret results in ways that suggest some success.

6. *What kind of environment does the program operate in and how does that affect the comparisons available to an evaluator?* Key to evaluating the effectiveness of a program are comparisons that allow us to estimate the incremental impacts of the program; this is the attribution question. In most evaluations, it is not feasible to conduct a randomized experiment—in fact, it is often not feasible to find a control group. Under these conditions, if we want to assess program effectiveness, it is still necessary to construct comparisons (e.g., among subgroups of program recipients who differ in their exposure to the program) that permit some ways of estimating whether the program made a difference (over what would have happened if there had been no intervention).

For evaluators, there are many issues that affect the evaluation design choices available. Among them are the following:

- Have any baseline data been kept?
- Is it possible to identify one or more comparison groups that are either not affected by the program or would be affected at a later time?
- How large is the client base for the program? (This affects **sampling** and statistical options.)
- Is the organization in which the program is embedded stable, or in a period of change? (This can affect the feasibility of proceeding with the evaluation.)

Programs are always embedded in an environment. The ways that the environment, other programs, organizational leaders, other departments in the government, central agencies, funders, boards of directors, and clients and other stakeholders affect and are affected by a program are typically dynamic. Even if a program is well established and the organization in which it is embedded is stable, these and other external influences can affect how the program is implemented as well as what it accomplishes. Many evaluators do not have sufficient control in evaluation engagements to partial out environmental factors, so qualitative assessments, direct observation, experience, and judgment often play key roles in estimating (a) which factors, if any, are in play for a program at the time it is evaluated and (b) how those factors affect the program process and results.

7. *Which research design alternatives are desirable and appropriate?* Typically, evaluations involve constructing multiple comparisons using multiple research designs; it is unusual, for example, for an evaluator to construct a design that relies on measuring just one outcome variable using one research design. Instead, evaluations will identify a set of outcome (and output) variables. Usually, each outcome variable will come with its own implicit research design. For example, a policy of reducing alcohol-related fatal crashes on British Columbia highways might focus on using coordinated police roadblocks and breathalyzer tests to affect the likelihood that motorists will drink and drive. A key outcome variable would be a time series of (monthly)

totals of alcohol-related fatal crashes—data collected by the Insurance Corporation of British Columbia. An additional measure of the success might be the cross-sectional survey-based perceptions of motorists in jurisdictions in which the policy has been implemented. The two research designs—a **single time series** and a case study **design**—have some complementary features that can strengthen the overall evaluation design.

When we look at evaluation practice, many evaluations rely on research design options that do not have the benefit of baselines or no-program comparison groups. These evaluations rely instead on a combination of independent lines of evidence to construct a multifaceted picture of program operations and results. **Triangulating** those results becomes a key part of assessing program effectiveness. An important consideration for practitioners is knowing the strengths and weaknesses of different designs so that combinations of designs can be chosen that complement each other (offsetting each other's weaknesses where possible).

8. *What information/data sources are available and appropriate, given the evaluation issues, the program structure, and the environment in which the program operates?* In most 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 the constructs in key evaluation questions would be measured, in conjunction with decisions about research designs. Research design considerations (specifically, **internal validity**) can be used as a rationale for prioritizing additional data collection.

Specific questions include the following:

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

The availability and quality of program performance data have the potential to assist evaluators in scoping an evaluation project. Performance measurement systems that have been constructed for programs, policies, or organizations are usually intended to periodically measure outputs and outcomes. For monitoring purposes, these data are often arrayed in a time series format, so that managers can monitor the trends and estimate whether performance results are tracking in ways that suggest program effectiveness. Where performance targets have been specified, the data can be compared periodically with the targets to see what the gaps are, if any.

Some jurisdictions, including the federal government in Canada (TBS, 2009), have linked performance data to program evaluations, with the goal of making performance results information—which is usually intended for program managers—more useful for evaluations of program efficiency and effectiveness.

There is one more important point to make with respect to potential data sources. Evaluations that focus on a set of questions on, for example, program effectiveness, program relevance, or program appropriateness, will usually break these questions down further, so that an evaluation question will yield several more specific subquestions that are tailored to that evaluation. Collectively, answering these questions and subquestions is the agenda for the whole evaluation project.

What can be very helpful is to construct a matrix that displays the evaluation questions and subquestions as rows, and the prospective data sources that will be used to address each question as columns. In one table, then, stakeholders can see how the evaluation will address each question and subquestion. Given that typical evaluations are about gathering and analyzing multiple lines of evidence, a useful practice is to make sure that each evaluation subquestion is addressed *by at least two lines of evidence.* Lines of evidence typically include administrative records, surveys, focus groups, stakeholder interviews, literature reviews/syntheses, and case studies.

9. *Given all the issues raised in Points 1 to 8, which evaluation strategy is most feasible and defensible?* No evaluation design is unassailable. The important thing for evaluators is to be able to understand the underlying logic of assessing the cause and effect linkages in an intended program structure, *anticipate* the key criticisms that could be made, and have a response (quantitative, qualitative, or both) to each criticism.

Most of the work that we do as evaluators is not going to involve randomized controlled experiments or even quasi-experiments, although some consider those to be the "gold standard" of rigorous social scientific research (see, e.g., Cook et al., 2010; 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 raise criticisms of your work. It is essential to understand the principles of rigorous evaluations to be able to proactively acknowledge limitations in an evaluation strategy. In Chapter 3, we will introduce the four kinds of validity that have been associated with a structured, quantitative approach to evaluation that focuses on discerning the key cause-and-effect relationships in a policy or program. Ultimately, evaluators must make some hard choices and be prepared to accept the fact that their work can, and probably will, be criticized.

10. *Should the evaluation be undertaken?* The final question in an assessment of evaluation feasibility is whether to proceed with the actual evaluation. It is possible that after having looked at the mix of

- evaluation issues,
- resource constraints,
- organizational and political issues (including the stability of the program), and
- research design options and measurement constraints,

the 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 it can save an organization considerable time and effort that probably would not have yielded a credible product.

Evaluator experience is key to being able to negotiate a path that permits designing a credible evaluation project. Evaluator judgment is an essential part of considering the requirements for a defensible study, and making a recommendation to either proceed or not.

Doing the Evaluation

Up to this point, we have outlined a planning and assessment process for conducting program evaluations. That process entails enough effort to be able to make an informed decision about proceeding or not with an evaluation. The work also serves as a substantial foundation for the evaluation, if it goes ahead. If a decision is made to go ahead with the evaluation, there are five more steps that are common to most evaluations.

1. *Develop the measures and pretest them.* Evaluations typically rely on a mix of existing and evaluation-generated data sources. If performance data are available, it is essential to assess how accurate and complete they are before committing to using them. As well, relying on administrative databases can be an advantage or a cost, depending on how complete and accessible those data are.

For data collection conducted by the evaluator or other stakeholders (sometimes, the client will collect some of the data and the evaluators will collect other lines of evidence), instruments will need to be designed. Surveys are a common means of collecting new data, and we will include information on designing and implementing surveys in Chapter 4 of this textbook.

For data collection instruments that are developed by the evaluators (or are adapted from some other application), **pretesting** is important. As an evaluation team, you usually have one shot at collecting key lines of evidence. To have one or more data collection instruments that are flawed (e.g., questions are ambiguous, questions are not ordered appropriately, some key questions are missing, some questions are redundant, or the instrument is too long) undermines the whole evaluation. Pretesting need not be elaborate; usually asking several persons to complete an instrument and then debriefing them will reveal most problems.

Some methodologists advocate an additional step: piloting the data collection instruments once they are pretested. This usually involves taking a small sample of persons who would actually be included in the evaluation as participants, and asking them to complete the instruments. This step is most useful in situations in which survey instruments have been designed to include **open-ended questions**—these questions can generate very useful data but are time-consuming to code later on. A pilot test can generate a range of open-ended responses that can be used to develop semi-structured response frames for those questions. Although some respondents in the full survey will offer open-ended comments that are outside the range of those in the pilot test, the pre-coded options will capture enough to make the coding process less time-consuming.

2. *Collect the data that are appropriate for answering the evaluation questions.* Collecting data from existing data sources requires both patience and thoroughness. Existing records, files, spreadsheets, or other sources of secondary (existing) data can be well organized or not. In some evaluations the consultants discover, after having signed a contract that made some assumptions about the condition of existing data sources, that

there are unexpected problems with the data files. Missing records, incomplete records, or inconsistent information can increase data collection time and even limit the usefulness of whole lines of evidence.

One of the authors was involved in an evaluation of a regional (Canadian) federal-provincial economic development program in which the consulting company that won the contract counted on project records being complete and easily accessible. When they were not, the project methodology had to be adjusted, and costs to the consultants increased. A disagreement developed around who should absorb the costs, and the evaluation process was a less positive experience than it should have been.

Collecting data through the efforts of the evaluation team or their subcontractors also requires a high level of organization and attention to detail. Surveying is a principal means of collecting evaluation-related data from stakeholders. Good survey techniques (in addition to having a defensible way to sample from populations) involve sufficient follow-up to help ensure that response rates are acceptable. Routinely, surveys do not achieve response rates higher than 50% (companies that specialize in doing surveys usually get better response rates than that). If inferential statistics are being used to infer from survey samples to populations, lower response rates weaken any generalizations. A general problem now is that people increasingly feel they are over-surveyed. This can mean that response rates will be lower than they have been historically. In evaluations where resources are tight, it may be that evaluators have to accept lower response rates, and they compensate for that (to some extent) by having multiple lines of evidence to offer opportunities to triangulate findings.

3. *Analyze the data, focusing on answering the evaluation questions.* Data analysis can be **quantitative** (involves working with variables that are represented numerically) or **qualitative** (involves analysis of words, documents, text, and other non-numerical representations of information including direct observations). Most evaluations use combinations of qualitative and quantitative data. **Mixed methods** have become the dominant approach for doing evaluations, following the trend in social science research more generally (Creswell & Plano Clark, 2007).

Quantitative data facilitate numerical comparisons and are important for estimates of technical efficiency, cost-effectiveness, and the costs and benefits of a program. In many governmental settings, performance measures tend to be quantitative, facilitating comparisons between annual targets and actual results. Qualitative data are valuable as a way of describing policy or program processes and impacts, using cases or narratives to offer in-depth understanding of how the program operates and how it affects stakeholders and clients. Open-ended questions can provide the opportunity for clients to offer information that researchers may not have thought to ask for in the evaluation.

A general rule that should guide 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 models to analyze program evaluation data. Although that strategy addressed possible criticisms by scholars, it often produced reports that were inaccessible, or perceived as untrustworthy from a user's perspective because they could not be understood. More recently, program evaluators have adopted mixed strategies for analyzing data, which rely on statistical tools where necessary, but also incorporate visual/graphic representations of findings.

In this book, we will not cover data analysis methods in detail. References to statistical methods are 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 as well as mixed-methods evaluations; and in Chapter 6, in connection with needs assessments, we introduce some basics of sampling and generalizing from sample findings to populations.

4. *Write, review, and finalize the report.* Evaluations are often conducted in situations in which stakeholders have different views of the effectiveness of the program. Where the main purpose for the evaluation is to make judgments about the merit or worth of the program, evaluations can be contentious.

A steering committee that serves as a sounding board/advisory body for the evaluation is an important part of guiding the evaluation. This is particularly valuable when evaluation reports are being drafted. Assuming that defensible decisions have been made around methodologies, data collection, and analysis strategies, the first draft of an evaluation report will represent a synthesis of lines of evidence and an overall interpretation of the information that is gathered. It is essential that the synthesis of evidence address the evaluation questions that motivated the project. In addressing the evaluation questions, evaluators will be exercising their judgment. Professional judgment is conditioned by knowledge, values, beliefs, and experience and can mean that members of the evaluation team will have different views on how the evaluation report should be drafted.

Working in a team makes it possible for evaluators to share perspectives, including the responsibility for writing the report. Equally important is some kind of challenge process that occurs as the draft report is completed and reviewed. Challenge functions can vary in formality, but the basic idea is that the draft report is critically reviewed by persons who have not been involved in conducting the evaluation. In the audit community, for example, it is common for draft audit reports to be discussed in depth by a committee of peers in the audit organization who have not been involved in the audit. The idea is to anticipate criticisms of the report and make changes that are needed, producing a product behind which the audit office will stand. Credibility is a key asset for individuals and organizations in the audit community, generally.

In the evaluation community, the challenge function is often played by the evaluation steering committee. Membership of the committee can vary but will typically include external expertise, as well as persons who have a stake in the program or policy. Canadian federal departments and agencies use blind peer review of evaluation-related products (draft final reports, methodologies, and draft technical reports) to obtain independent assessments of the quality of evaluation work. Depending on the purposes of the evaluation, reviews of the draft report by members of the steering committee can be contentious. One issue for executives who are overseeing the evaluation of policies is to anticipate possible conflicts of interest by members of steering committees.

In preparing an evaluation report, a key part is the recommendations that are made. Here again, professional judgment plays a key role; recommendations must not only be backed up by evidence but also be appropriate, given the context for the evaluation. Making

recommendations that reflect key evaluation conclusions *and* are feasible is a skill that is among the most valuable that an evaluator can develop.

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 and conclusions where possible.
- Use clear, simple language in the report.
- Use more headings and subheadings, rather than fewer, in the report.
- Prepare a clear, concise executive summary.
- Be prepared to edit or even seek professional assistance to edit the penultimate draft of the report before finalizing it.

5. *Disseminate the report.* Evaluators have an obligation to produce a report *and* make a series of presentations of the findings, conclusions, and recommendations to key stakeholders, including the clients of the evaluation. There are different views of how much interaction is appropriate between evaluators and clients. One view, articulated by Michael Scriven (1997), is that program evaluators should be very careful about getting involved with their clients; interaction at *any* stage in an evaluation, including post-reporting, can compromise their **objectivity**. Michael Patton (2008), by contrast, argues that *unless* program evaluators get involved with their clients, evaluations are not likely to be used.

The degree and types of interactions between evaluators and clients/managers will depend on the purposes of the evaluation. For evaluations that are intended to recommend incremental changes to a policy or program, manager involvement will generally not compromise the validity of the evaluation products. But for evaluations in which major decisions that could affect the existence of the program are in the offing, it is important to assure evaluator independence, ensuring that they do not directly report to the managers of the policy or program being evaluated.

Making Changes Based on the Evaluation

Evaluations can, and hopefully do, become part of the process of making changes in the programs or the organization in which they operate. Where they are used, 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 evaluations are often intended to raise this question (Weiss, 1998).

The whole issue of whether, and to what extent, evaluations are used continues to be an important topic in the literature. Although there is clearly a view that the quality of an evaluation rests on its methodological defensibility, and that short term, specific uses are not that important (Fitzpatrick, 2002), many evaluators have taken the view that use is central to the reasons for doing evaluations (Amo & Cousins, 2007; Fleischer & Christie, 2009; Leviton, 2003; Mark & Henry, 2004; Patton, 2008). The following are possible changes based on evaluations:

- Making incremental changes to the existing policy or program
- Increasing the scale of the policy or program
- Increasing the scope of the policy or program
- Downsizing the policy or program

- Replacing the policy or program
- Eliminating the policy or program

Evaluations are one source of information in policy decision making. Depending on the context, evaluation evidence may be a key part of decision making, or may be one of a number of factors that are taken into account.

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 past 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 are 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 core knowledge and skills.

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 important and involves 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 and contextual constraints 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 program 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.

2. Read the table of contents for this textbook and, based on your own background and experience, explain what you anticipate will be the easiest parts of this book for you to understand. Why?

3. Again, having looked over 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 "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 program managers." 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 halfway 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

Amo, C., & Cousins, J. B. (2007). Going through the process: An examination of the operationalization of process use in empirical research on evaluation. *New Directions for Evaluation, 116*, 5–26.

Anderson, L. M., Fielding, J. E., Fullilove, M. T., Scrimshaw, S. C., & Carande-Kulis, V. G. (2003). Methods for conducting systematic reviews of the evidence of effectiveness and economic efficiency of interventions to promote healthy social environments. *American Journal of Preventive Medicine, 24*(3 Suppl.), 25–31.

Auditor General of British Columbia and Deputy Ministers' Council. (1996). *Enhancing accountability for performance: A framework and an implementation plan—Second joint report*. Victoria, British Columbia, Canada: Queen's Printer for British Columbia.

Bakvis, H., & Juillet, L. (2004). *The horizontal challenge: Line departments, central agencies and leadership*. Ottawa, Ontario, Canada: Canada School of Public Service.

Bamberger, M., Rugh, J., Church, M., & Fort, L. (2004). Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation, 25*(1), 5–37.

Basilevsky, A., & Hum, D. (1984). *Experimental social programs and analytic methods: An evaluation of the U.S. income maintenance projects*. Orlando, FL: Academic Press.

Bernstein, D. J. (1999). Comments on Perrin's "effective use and misuse of performance measurement." *American Journal of Evaluation, 20*(1), 85–93.

Bickman, L. (1996). A continuum of care. *The American Psychologist, 51*(7), 689–701.

Boulmetis, J., & Dutwin, P. (2000). *The ABC's of evaluation: Timeless techniques for program and project managers*. San Francisco, CA: Jossey-Bass.

Campbell, S., Benita, S., Coates, E., Davies, P., & Penn, G. (2007). *Analysis for policy: Evidence-based policy in practice*. London, England: Government Social Research Unit, HM Treasury.

Campbell Collaboration. (2010). *About us*. Retrieved from http://www.campbellcollaboration.org/about_us/index.php

Chelimsky, E. (1997). The coming transformations in evaluation. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. ix–xii). Thousand Oaks, CA: Sage.

Chen, H.-T. (1996). A comprehensive typology for program evaluation. *Evaluation Practice, 17*(2), 121–130.

Cook, T. D., Scriven, M., Coryn, C. L., & Evergreen, S. D. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31*(1), 105–117.

Coryn, C. L., Schröter, D. C., Noakes, L. A., & Westine, C. D. (2011). A systematic review of theory-driven evaluation practice from 1990 to 2009. *American Journal of Evaluation, 32*(2), 199–226.

Creswell, J. W., & Plano Clark, V. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Curristine, T. (2005). Government performance: Lessons and challenges. *OECD Journal on Budgeting, 5*(1), 127–151.

Donaldson, S. I. (2007). *Program theory-driven evaluation science: Strategies and applications*. New York: Lawrence Erlbaum.

Feller, I. (2002). Performance measurement redux. *American Journal of Evaluation, 23*(4), 435–452.

Fitzgerald, J., & Rasheed, J. M. (1998). Salvaging an evaluation from the swampy lowland. *Evaluation and Program Planning, 21*(2), 199–209.

Fitzpatrick, J. (2002). Dialogue with Stewart Donaldson. *American Journal of Evaluation, 23*(3), 347–365.

Fleischer, D., & Christie, C. (2009). Evaluation use: Results from a survey of U.S. American Evaluation Association members. *American Journal of Evaluation, 30*(2), 158–175.

Funnell, S., & Rogers, P. (2011). *Purposeful program theory: Effective use of theories of change and logic models*. San Francisco, CA: Jossey-Bass.

Government Social Science Research Unit. (2007). *Magenta book: Guidance notes on policy evaluation*. Retrieved from http://www.civilservice.gov.uk/wp-content/uploads/2011/09/magenta_book_complete_sep_2007.pdf

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hatry, H. P. (2006). *Performance measurement: Getting results* (2nd ed.). Washington, DC: Urban Institute Press.

Higgins, J., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions. Version 5.0.2 [updated March 2011]*. Retrieved from www.cochrane-handbook.org

Hood, C. (1991). A public management for all seasons? *Public Administration, 69*(1), 3–19.

Joyce, P. G. (2011). The Obama administration and PBB: Building on the legacy of federal performance-informed budgeting? *Public Administration Review, 71*(3), 356–367.

Knowlton, L. W., & Phillips, C. C. (2009). *The logic model guidebook*. Thousand Oaks, CA: Sage.

Krause, D. R. (1996). *Effective program evaluation: An introduction*. Chicago, IL: NelsonHall.

Leviton, L. C. (2003). Evaluation use: Advances, challenges and applications. *American Journal of Evaluation, 24*(4), 525–535.

Lincoln, Y. S., & Guba, E. G. (1980). The distinction between merit and worth in evaluation. *Educational Evaluation and Policy Analysis, 2*(4), 61–71.

Lipsey, M. W. (2000). Method and rationality are not social diseases. *American Journal of Evaluation, 21*(2), 221–223.

Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation, 10*(1), 35–57.

Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. *Canadian Journal of Program Evaluation, 16*(1), 1–24.

Mayne, J. (2006). Audit and evaluation in public management: Challenges, reforms and different roles. *Canadian Journal of Program Evaluation, 21*(1), 11–45.

Mayne, J. (2008). *Building an evaluative culture for effective evaluation and results management*. Retrieved from http://www.cgiar-ilac.org/files/publications/briefs/ILAC_Brief20_Evaluative_Culture.pdf

Mayne, J. (2011). Contribution analysis: Addressing cause and effect. In K. Forss, M. Marra, & R. Schwartz (Eds.), *Evaluating the complex: Attribution, contribution, and beyond: Comparative policy evaluation* (Vol. 18, pp. 53–96). New Brunswick, NJ: Transaction.

Mayne, J., & Rist, R. C. (2006). Studies are not enough: The necessary transformation of evaluation. *Canadian Journal of Program Evaluation, 21*(3), 93–120.

McDavid, J. C. (2001). Program evaluation in British Columbia in a time of transition: 1995–2000. *Canadian Journal of Program Evaluation, 16*(Special Issue), 3–28.

McDavid, J. C., & Huse, I. (2006). Will evaluation prosper in the future? *Canadian Journal of Program Evaluation, 21*(3), 47–72.

McDavid, J. C., & Huse, I. (2012). Legislator uses of public performance reports: Findings from a five-year study. *American Journal of Evaluation, 33*(1), 7–25.

Moynihan, D. P. (2008). *The dynamics of performance management: Constructing information and reform*. Washington, DC: Georgetown University Press.

Nagarajan, N., & Vanheukelen, M. (1997). *Evaluating EU expenditure programs: A guide*. Luxembourg: Publications Office of the European Union.

Newcomer, K. E. (1997). Using performance measurement to improve public and nonprofit programs. In K. E. Newcomer (Ed.), *New directions for evaluation,* (Vol. 75, pp. 5–14). San Francisco, CA: Jossey-Bass.

Office of Management and Budget. (2012). *The mission and structure of the Office of Management and Budget*. Retrieved from http://www.whitehouse.gov/omb/organization_mission/

Osborne, D., & Gaebler, T. (1992). *Reinventing government: How the entrepreneurial spirit is transforming the public sector*. Reading, MA: Addison-Wesley.

Owen, J. M., & Rogers, P. J. (1999). *Program evaluation: Forms and approaches* (International ed.). London, England: Sage.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice, 15*(3), 311–319.

Patton, M. Q. (2008). *Utilization focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity to enhance innovation and use*. New York: Guilford Press.

Perrin, B. (1998). Effective use and misuse of performance measurement. *American Journal of Evaluation, 19*(3), 367–379.

Picciotto, R. (2011). The logic of evaluation professionalism. *Evaluation, 17*(2), 165–180.

Roessner, J. D. (2002). Outcome measurement in the USA: State of the art. *Research Evaluation, 11*(2), 85–93.

Rutman, L. (1984). Introduction. In L. Rutman (Ed.), *Evaluation research methods: A basic guide* (Sage Focus Editions Series, Vol. 3, 2nd ed., pp. 9–38). Beverly Hills, CA: Sage.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (AERA Monograph Series—Curriculum Evaluation, pp. 39–83). Chicago, IL: Rand McNally.

Scriven, M. (1972). The exact role of value judgments in science. In R. S. Cohen & K. Schaffner (Eds.), *Proceedings of the biennial meeting of the Philosophy of Science Association* (pp. 219–247). Dordrecht, Holland: Reidel.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin & D. C. Phillips (Eds.), *Evaluation and education: At quarter century* (pp. 18–64). Chicago, IL: University of Chicago Press.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice, 17*(2), 151–161.

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. (2008). A summative evaluation of RCT methodology & an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5*(9), 11–24.

Shaw, I. (2000). *Evaluating public programmes: Contexts and issues*. Burlington, VT: Ashgate.

Solesbury, W. (2001). *Evidence-based policy: Whence it came and where it's going* (ESRC Working Paper No. 1). Swindon, England: Centre for Evidence-based Policy and Practice. Retrieved from http://www.kcl.ac.uk/content/1/c6/03/45/84/wp1.pdf

Talarico, T. (1999). *An evaluation of the Neighbourhood Integrated Service Team program* (Unpublished master's report). University of Victoria, British Columbia, Canada.

Treasury Board of Canada Secretariat. (2009). *Policy on evaluation*. Retrieved from http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=15024

Weiss, C. H. (1972). *Evaluation research: Methods for assessing program effectiveness*. Englewood Cliffs, NJ: Prentice Hall.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Yeh, S. S. (2007). The cost-effectiveness of five policies for improving student achievement. *American Journal of Evaluation, 28*(4), 416–436.