1

KEY CONCEPTS AND ISSUES IN PROGRAM EVALUATION AND PERFORMANCE MEASUREMENT

CONTENTS

INTRODUCTION

Our main focus in this textbook is on understanding how to evaluate the **effectiveness** of public-sector policies and programs. **Evaluation** is widely used in public, nonprofit, and private-sector organizations to generate information for policy and program planning, design, implementation, assessment of results, improvement/learning, accountability, and public communications. It 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 uncover unintended effects of programs and policies, which can affect overall assessments of programs or policies. On a perhaps more subtle level, the *process* of measuring performance or conducting program evaluations—that is, aside from the reports and other evaluation products—can also have impacts on the individuals and organizations involved, including attentive stakeholders and citizens.

The primary goal of this textbook is to provide a solid methodological foundation to evaluative efforts, so that both the *process* and the *information created* offer defensible contributions to political and managerial decision-making. **Program evaluation** is a rich and varied combination of theory and practice. This book will introduce a broad range of evaluation approaches and practices, reflecting the richness of the field. 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.

In the rest of this chapter, we do the following:

- Describe how program evaluation and performance measurement are complementary approaches to creating information for decision makers and stakeholders in public and nonprofit organizations.
- Introduce the concept of the performance management cycle, and show how program evaluation and performance measurement conceptually fit the performance management cycle.
- Introduce key concepts and principles for program evaluations.
- Illustrate a program evaluation with a **case study**.
- Introduce 10 general questions that can underpin evaluation projects.
- Summarize 10 key steps in assessing the feasibility of conducting a program evaluation.
- Finally, present an overview of five key steps in doing and reporting an evaluation.

Integrating Program Evaluation and Performance Measurement

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 people (or other **units of analysis**) 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. Essentially, that means determining the difference between what occurred as a result as a program and what would have occurred if the program had not been implemented. Although RCTs are not the most common method used 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 exemplars of "good" evaluations (Cook, Scriven, Coryn, & Evergreen, 2010; Donaldson, Christie, & Melvin, 2014).

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 policy or program that has already been implemented, with no real ways to create **control groups** and usually no baseline (pre-program) data to construct before–after comparisons. Often, measurement of program outcomes is challenging—there may be no data readily available, a short timeframe for the need for the information, and/or 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 or client characteristics that only partly overlap with the key questions that need to be addressed in the evaluation. We will learn about quasi-experimental designs and other quantitative and qualitative evaluation methods throughout the book.

So how does performance measurement fit into the picture? Evaluation as a field has been transformed in the past 40 years by the broad-based movement in public and nonprofit organizations to construct and implement systems that measure program and organizational performance. Advances in technology have made it easier and less expensive to create, track, and share performance measurement data. Performance measures can, in some cases, productively be incorporated into evaluations. 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.

The use of performance measures in evaluative work is, however, seldom straightforward. For example, recent analysis has shown that in the search for government efficiencies, particularly in times of fiscal restraint, governments may cut back on evaluation capacity, with expectations that performance measurement systems can substantially cover the performance management information needs (de Lancer Julnes & Steccolini, 2015). This trend to lean on performance measurement, particularly in high-stakes accountability situations, is increasingly seen as being detrimental to learning, policy and program effectiveness, and staff morale (see, for example, Arnaboldi et al., 2015; Coen & Roberts, 2012; Greiling & Halachmi, 2013; Mahler & Posner, 2014). We will explore this conundrum in more depth later in the textbook.

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. 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.

We see performance measurement approaches as *complementary* to program evaluation, and not as a replacement for evaluations. The approach of this textbook is that evaluation includes both program evaluation and performance measurement, and we build a foundation in the early chapters of the textbook that shows how program evaluation can inform measuring the performance of programs and policies. Consequently, 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. 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, but without overstretching their real capabilities. 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). Chapter 6 on **needs assessments** builds on topics covered in the earlier chapters, including Chapter 1. Needs assessments can occur in several phases of the performance management cycle: strategic planning, designing effective programs, implementation, 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 part of strategic planning, or as we design policies or programs, or as we evaluate their outcomes (the assessment and reporting phase).

Below, we introduce the relationship between organizational management and evaluation activities. We expand on this issue in Chapter 11, where we examine how evaluation theory and practice are joined with management in public and nonprofit organizations. 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.

Connecting Evaluation to the Performance Management System

Information from program evaluations and performance measurement systems is expected to play a role in the way managers operate their programs (Hunter & Nielsen, 2013; Newcomer & Brass, 2016). Performance management, which is sometimes called **results-based management**, emerged as an organizational management approach that has been part of a broad movement of **new public management (NPM)** in public administration. NPM has had significant impacts on governments worldwide since it came onto the scene in the early 1990s. It 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).

While the drive for NPM—particularly the emphasis on explicitly linking funding to targeted outcomes—has abated somewhat as paradoxes of the approach have come to light (Pollitt & Bouckaert, 2011), particularly in light of the global financial crisis (Coen & Roberts, 2012; OECD, 2015), the importance of evidence of actual accomplishments is still considered central to performance management. Performance management systems will continue to evolve; evidence-based and evidence-informed decision making depend heavily on both evaluation and performance measurement, and will respond as the political and fiscal structure and the context of public administration evolve. There is discussion recently of a transition from NPM to a more centralized but networked New Public Governance (Arnaboldi et al., 2015; Osborne, 2010; Pollitt & Bouckaert, 2011), Digital-Era Governance (Dunleavy, Margetts, Bastow, & Tinker, 2006; Lindquist & Huse, 2017), Public Value Governance (Bryson, Crosby, & Bloomberg, 2014), and potentially a more agile governance (OECD, 2015; Room, 2011). In any case, evidence-based or evidence-informed policy making will remain an important feature of public administration and public policy.

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. These efforts are part of an organizational architecture designed to pull together the components to achieve organizational goals. 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.

✓ AMERICAN GOVERNMENT FOCUS ON PROGRAM PERFORMANCE RESULTS

In the United States, successive federal administrations beginning with the Clinton administration in 1992 embraced program **goal** setting, performance measurement, and reporting as a regular feature of program accountability (Joyce, 2011; Mahler & Posner, 2014). The Bush administration, between 2002 and 2009, emphasized the importance of program performance in the budgeting process. The Office of Management and Budget (OMB) introduced assessments of programs using a methodology called PART (Performance Assessment Rating Tool) (Gilmour, 2007). 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 executive branch (presidential) budget requests to Congress.

The Obama administration, while instituting the 2010 GPRA Modernization Act (see Moynihan, 2013) and 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 was expected to work with agencies to "encourage use and communication of performance information and to improve results and transparency" (OMB archives, 2012). The GPRA Modernization Act is intended to create a more organized and publicly accessible system for posting performance information on the www.Performance.gov website, in a common format. There is also currently a clear theme of improving the efficiencies and *integration* of evaluative evidence, including making better use of existing data.

At the time of writing this book, it is too early to tell what changes the Trump administration will initiate or will keep from previous administrations, although there is intent to post performance information on the Performance.gov website, reflecting updated goals and alignment. Its current mission is "to assist the President in meeting his policy, budget, management and regulatory objectives and to fulfill the agency's statutory responsibilities" (OMB, 2018, p. 1).

CANADIAN GOVERNMENT EVALUATION POLICY

In Canada, there is a long history of requiring program evaluation of federal government programs, dating back to the late 1970s. More recently, a major update of the federal government's evaluation policy occurred in 2009, and again in 2016 (TBS, 2016a). The main plank in that policy is a requirement that federal departments and agencies evaluate the relevance and performance of their programs on a 5-year cycle, with some exemptions for smaller programs and contributions to international organizations (TBS, 2016a, sections 2.5 and 2.6). Performance measurement and program evaluation is explicitly linked to accountability (resource allocation [s. 3.2.3] and reporting to parliamentarians [s. 3.2.4]) as well as managing and improving departmental programs, policies, and services (s. 3.2.2). There have been reviews of Canadian provinces (e.g., Gauthier et al., 2009), American states (Melkers & Willoughby, 2004; Moynihan, 2006), and local governments (Melkers & Willoughby, 2005) on their approaches to evaluation and performance measurement. In later chapters, we will return to this issue of the challenges of using the same evaluative information for different purposes (see Kroll, 2015; Majone, 1989; Radin, 2006).

In summary, 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. Central agencies (including the U.S. Federal Office of Management and Budget [OMB], the General Accountability Office [GAO], and the Treasury Board of Canada Secretariat [TBS]), as well as state and provincial finance departments and auditors, develop policies and articulate expectations that shape the ways program managers are expected to create and use performance information to inform their administrative superiors and other stakeholders outside the organization about what they are doing and how well they are doing it. It is worthwhile following the websites of these organizations to understand the subtle and not-so-subtle shifts in expectations and performance frameworks for the design, conduct, and uses of performance measurement systems and evaluations over time, especially when there is a change in government.

Fundamental to performance management is the importance of program and policy performance results being collected, analyzed, compared (sometimes to performance targets), and then used to monitor, learn, 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 (Van de Walle & Cornelissen, 2014).

The Performance Management Cycle

Organizations typically run through an annual **performance management cycle** that includes budget negotiations, announcing budget plans, designing or modifying programs, managing programs, reporting their financial and nonfinancial results, and making informed adjustments. The performance management cycle is a useful normative model that includes an iterative planning–implementation–assessment–program adjustments sequence. The model can help us understand the various points at which program evaluation and performance measurement can play important roles as ways of providing information to decision makers who are engaged in leading and managing organizations and programs to achieve results, and reporting the results to legislators and the public.

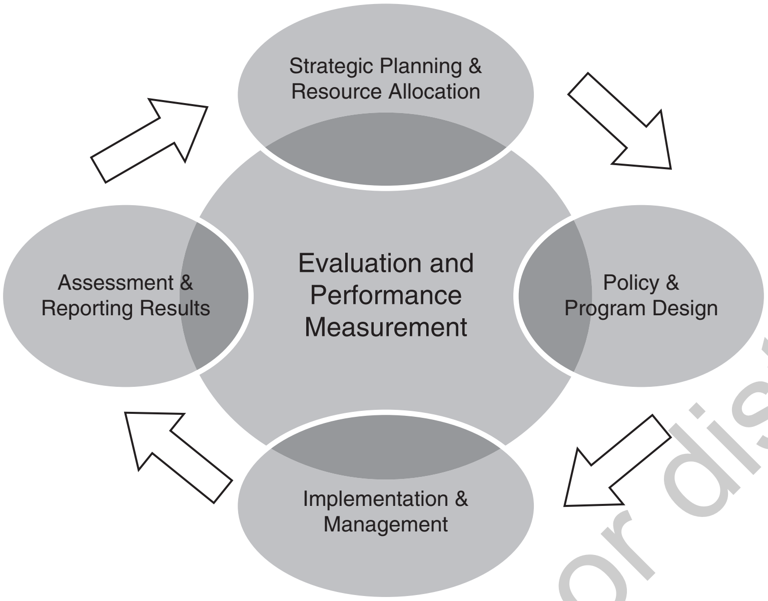
In this book, the performance management cycle illustrated in Figure 1.1 is used as a framework for organizing different evaluation topics and showing how the analytical approaches covered in key chapters map onto the performance management cycle. Figure 1.1 shows a model of how organizations can integrate strategic planning, program and policy design, implementation, and assessment of results into a cycle where evaluation and performance measures can inform *all phases of the cycle*. The assessment and reporting part of the cycle is central to this textbook, but we take the view that all phases of the performance management cycle can be informed by evaluation and performance measurement.

We will use the performance management cycle as a framework within which evaluation and performance measurement activities can be situated for managers and other stakeholders in public sector and nonprofit organizations. It is important to reiterate, however, that specific evaluations and performance measures are often designed to serve a *particular* informational purpose—that is, a certain phase of the cycle—and may not be appropriate for other uses.

The four-part performance management cycle begins with formulating and budgeting for clear (strategic) objectives for organizations and, hence, for programs and policies. Strategic objectives are then translated into program and policy designs intended to achieve those objectives. This phase involves building or adapting organizational structures and processes to facilitate implementing and managing policies or programs. ***Ex ante*** **evaluations** can occur at the stage when options are being considered and compared as candidates for design and implementation. We will look a bit more closely at *ex ante* evaluations later in the textbook. 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 policy and **program implementation** and management. In this textbook, we will look at **formative evaluations** as a type of implementation-related evaluation that typically informs managers how to improve their programs. Normally, 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

FIGURE 1.1 ● THE PERFORMANCE MANAGEMENT CYCLE


Strategic Planning & Resource Allocation
Policy & Program Design
Implementation & Management
Assessment & Reporting Results
Evaluation and Performance Measurement

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, cultures, and priorities either facilitate or impede program implementation.

The fourth phase in the cycle is about assessing performance results, and reporting to legislators, the public, and other (internal or external) stakeholders. 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 substantial program changes, or decisions about the future of the program. We will discuss formative and summative evaluations more thoroughly later in this chapter.

Performance monitoring is an important way to tell how a program is tracking over time, but, as shown in the model, performance measures can inform decisions made at any stage of the performance cycle, not just the assessment stage. Performance data can be useful for strategic planning, program design, and management-related implementation decisions. At the Assessment and Reporting Results phase, "performance measurement and reporting" is expected to contribute to accountability for programs. That is, performance measurement can lead to a number of consequences, from program adjustments to impacts on elections. In the final phase of the cycle, 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 strategic planning.

Stepping back from this cycle, we see a strategic management system that encompasses how ideas and evaluative information are gathered for policy planning and subsequent funding allocation and reallocation. Many governments have institutionalized their own performance information architecture to formalize how programs and departments are expected to provide information to be used by the managerial and political decision makers. Looking at Canada and the United States, we can see that this architecture evolves over time as the governance context changes and also becomes more complex, with networks of organizations contributing to outcomes. The respective emphasis on program evaluation and performance measurement can be altered over time. Times of change in government leadership are especially likely to spark changes in the performance information architecture. For example, in Canada, the election of the current Liberal Government in the 2015 federal election after nine years of Conservative Government leadership has resulted in a government-wide focus on implementing high-priority policies and programs and ensuring that their results are actually delivered (Barber, 2015; Barber, Moffitt, & Kihn, 2011).

POLICIES AND PROGRAMS

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 policies and programs. 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 core 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. Initially, policy objectives can be expressed in election platforms, political speeches, government responses to questions by the media, or other announcements (including social media). 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.

Royal commissions (in Canada), task forces, reports by independent bodies (including think tanks), 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—some policies are a response to a political crisis.

An example of a policy that has significant planned impacts is the British Columbia government's November 2007 Greenhouse Gas Reduction Targets Act (Government of British Columbia, 2007) that committed the provincial government to reducing greenhouse gas emissions in the province by 33% by 2020. From 2007 to 2013, British Columbia reduced its per capita consumption of petroleum products subject to the carbon tax by 16.1%, as compared with an *increase* of 3.0% in the rest of Canada (World Bank, 2014).

The legislation states that by 2050, greenhouse gas emissions will be 80% below 2007 levels. Reducing greenhouse gas emissions in British Columbia will be challenging, particularly given the more recent provincial priority placed on developing liquefied natural gas facilities to export LNG to Asian countries. In 2014, the BC government passed a Greenhouse Gas Industrial Reporting and Control Act (Government of British Columbia, 2014) that includes a baseline-and-credit system for which there is no fixed limit on emissions, but instead, polluters that reduce their emissions by more than specified targets (which can change over time) can earn credits that they can sell to other emitters who need them to meet their own targets. The World Bank annually tracks international carbon emission data (World Bank, 2017).

✓ WHAT IS A PROGRAM?

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 and scope. 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 range of activities, from public advertising, through intake and treatment, to referral, and finally to follow-up—or be complex—a multijurisdictional program to reduce homelessness that involves both governments and nonprofit organizations.

To reduce greenhouse gases in British Columbia, many different programs have been implemented—some targeting the government itself, others targeting industries, citizens, and other governments (e.g., British Columbia local governments). Programs to reduce greenhouse gases are 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, a key program that was implemented starting in 2008 was a broad-based 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. That is, there is a carbon tax component added to vehicle per liter fuel costs.

Increasingly, programs can involve several levels of government, governmental agencies, and/or nonprofit organizations. A good example is Canada's federal government initiatives, starting in 2016, to bring all provinces on board with GHG reduction initiatives. 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.

KEY CONCEPTS IN PROGRAM EVALUATION

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 view 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 (Cook & Campbell, 1979).

In the evaluation field, different approaches to assessing causal relationships have been proposed, and the debate around using experimental designs continues (Cook et al., 2010; Cresswell & Cresswell, 2017; Donaldson et al., 2014). 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/infer causal relationships therein.

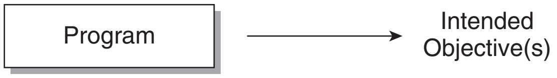
For evaluators, it may seem “second best” to have to rely on their own judgment, but realistically, *all* program evaluations entail a substantial number of judgment calls, even when valid and reliable data and appropriate comparisons are available. As Daniel Krause (1996) has pointed out, “A program 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 the responsibility of exercising professional judgment as they do their work.

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.

FIGURE 1.2 ⬢ LINKING PROGRAMS AND INTENDED OBJECTIVES



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 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 activities of the program itself.

The arrow connecting the program and its intended outcomes is a key part of most program evaluations and performance measurement systems. 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)?

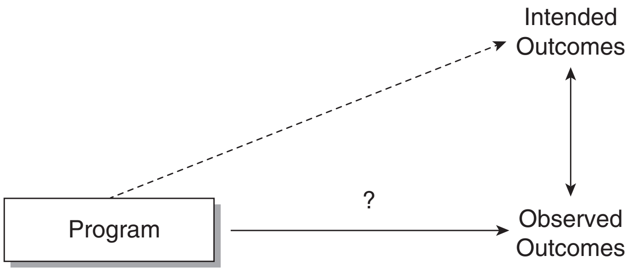
Assessing **program effectiveness** is the most common reason we conduct program evaluations and create performance measurement systems. We want to know whether, and to what extent, the program's actual results are consistent with the outcomes we expected. In fact, there are *two* evaluation issues related 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 have 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?

FIGURE 1.3 ⬢ THE TWO PROGRAM EFFECTIVENESS QUESTIONS INVOLVED IN MOST EVALUATIONS



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 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).

Formative and Summative Evaluations

Michael Scriven (1967) introduced the distinction between formative and summative evaluations (Weiss, 1998a). Since then, he has come back to this issue several more times (e.g., Scriven, 1991, 1996, 2008). Scriven's definitions reflected his distinction between implementation issues and evaluating program effectiveness. He associated formative evaluations primarily with analysis of program design and 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, stated objectives (the worth of a program). Summative evaluations could, for example, be used for accountability purposes or for budget reallocations.

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—and 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.

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 (1998a) 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, in which the continued 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 deeper 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, which differs in some ways from the formative and summative approaches generally adopted for this textbook. Patton sees developmental evaluations as preceding formative or summative evaluations (Patton, 2011). As we shall see, however, there can be pressures to use evaluations (and performance measures) that were originally intended for formative purposes, to be repurposed and "used" summatively. This is a challenge particularly in times of fiscal stress, where cutbacks in budget are occurring and can result in evaluations being seen to be inadequate for the (new) uses at hand (Shaw, 2016).

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 address at the design stage whether a policy or program (or one option from among several alternatives) is cost-beneficial. Assumptions about implementation and the existence and timing of outcomes, as well as costs, are required to facilitate such analyses. We discuss economic evaluation in Chapter 7.

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. Body-worn cameras for police officers are often introduced on a pilot basis, accompanied by an evaluation of their effectiveness.

One other possibility is to plan a program so that before it is implemented, **baseline measures** of outcomes 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.

Program evaluation clients often expect evaluators to come up with ways of telling whether the program achieved its objectives—that is, 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 IMPORTANCE OF PROFESSIONAL JUDGMENT IN EVALUATIONS

One of the principles underlying this book is the importance of exercising professional judgment as program evaluations are designed, executed, and acted on. Our view is that although sound and defensible methodologies are necessary foundations for credible evaluations, each evaluation process and the associated evaluation context necessitates making decisions that are grounded in professional judgment. Values, ethics, political awareness, and social/cultural perspectives are important, beyond technical expertise (Donaldson & Picciotto, 2016; House, 2016; Schwandt, 2015). There are growing expectations that stakeholders, including beneficiaries, be considered equitably in evaluations, and expectations to integrate evaluative information across networked organizations (Stockmann & Meyer, 2016; Szanyi, Azzam, & Galen, 2013).

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 at the time with what our own experiences, beliefs, values, and expectations furnish and display. Some of what we bring with us to an evaluation is **tacit knowledge**—that is, knowledge based on our experience—and it is not learned or communicated except by experience.

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.

The following case summary illustrates many of the facets of program evaluation, performance measurement, and performance management that are discussed in this textbook. We will outline the case in this chapter, and will return to it and other examples in later chapters of the book.

EXAMPLE: EVALUATING A POLICE BODY-WORN CAMERA PROGRAM IN RIALTO, CALIFORNIA

The Context: Growing Concerns With Police Use of Force and Community Relationship

Police forces in many Canadian and American cities and towns—as part of a global trend—have begun using body-worn cameras (BWCs) or are considering doing so (Lum et al., 2015). Aside from the technological advances that have made these small, portable cameras and their systems available and more affordable, there are a number of reasons to explain their growing use. In some communities, relationships between police and citizens are strained, and video evidence holds the promise of reducing police use of force, or complaints against the police. Recordings might also facilitate resolution of complaints. Just the presence of BWCs might modify police and citizen behaviors, and de-escalate potentially violent encounters (Jennings, Fridell, & Lynch, 2014). Recent high-profile incidents of excessive police use of force, particularly related to minority groups, have served as critical sparks for immediate political action, and BWCs are seen as a partial solution (Cubitt, Lesic, Myers, & Corry, 2017; Lum et al., 2015; Maskaly et al., 2017). Recordings could also be used in officer training. Aside from the intent to improve transparency and accountability, the use of BWCs holds the potential to provide more objective evidence in crime situations, thereby increasing the likelihood and speed of convictions.

On the other hand, implementation efforts can be hampered by police occupational cultures and their responses to the BWC use policies. Also, because the causal mechanisms are not well understood, BWCs may have unanticipated and unintended negative consequences on the interactions between police and citizens. There are also privacy concerns for both police and citizens. Thus, police BWC programs and policies raise a number of causality questions that have just begun to be explored (see Ariel et al., 2016; Ariel et al., 2018a, 2018b; Cubitt et al., 2017; Hedberg, Katz, & Choate, 2017; Lum et al., 2015; Maskaly et al., 2017). The Center for Evidence-Based Crime Policy at George Mason University (2016) notes, "This rapid adoption of BWCs is occurring within a low information environment; researchers are only beginning to develop knowledge about the effects, both intentional and unintentional, of this technology" (p. 1 of website). Some of the evaluations are RCTs (including our example that follows).

The U.S. Bureau of Justice Assistance (2018) provides a website (Body-Worn Camera Toolkit: https://www.bja.gov/bwc/resources.html) that now holds over 700 articles and additional resources about BWCs. About half of these are examples of local governments' policies and procedures. Public Safety Canada (2018) has approximately 20 similar resources. The seminal study by Ariel, Farrar, and Sutherland, *The Effect of Body-Worn Cameras on Use of Force and Citizens' Complaints Against the Police: A Randomized Controlled Trial* (Ariel et al., 2015) will be used in this chapter to highlight the importance of evaluating the implementation and outcomes of this high-stakes program. Related studies will also be mentioned throughout this textbook, where relevant.

Implementing and Evaluating the Effects of Body-Worn Cameras in the Rialto Police Department

The City of Rialto Police Department was one of the first in the United States to implement body-worn cameras and systematically evaluate their effects on citizen–police interactions (Ariel, Farrar, & Sutherland, 2015). The study itself took place over 12 months, beginning in 2012. Rialto Police Department was nearly disbanded in 2007 when the city considered contracting for police services with the Los Angeles County Sherriff's Department. Beset by a series of incidents involving questionable police officer behaviors including use-of-force incidents, the city hired Chief Tony Farrar in 2012. He decided to address the problems in the department by investing in body-worn cameras for his patrol officers and systematically evaluating their effectiveness. The evaluation addressed this question: "Do body-worn cameras reduce the prevalence of use-of-force and/or citizens' complaints against the police?" (Ariel et al., 2015, p. 509). More specifically, the evaluation was focused on this hypothesis: Police body-worn cameras will lead to increases in socially desirable behaviors of the officers who wear them and reductions in police use-of-force incidents and citizen complaints.

To test this hypothesis, a randomized controlled trial was conducted that became known internationally as the "Rialto Experiment"—the first such study of BWCs (Ariel et al., 2015). Over the year in which this program was implemented, officer shifts (a total of 988 shifts) were randomly assigned to either "treatment-shifts" (489), where patrol officers would wear a BWC that recorded all incidents of contact with the public, or to "control-shifts" (499), where they did not wear a BWC. Each week entailed 19 shifts, and each shift was 12 hours in duration and involved approximately 10 officers patrolling in Rialto. Each of the 54 patrol officers had multiple shifts where they did wear a camera, and shifts where they did not.

The study defined a use-of-force incident as an encounter with "physical force that is greater than basic control or 'compliance holds'—including the use of (a) OC spray [pepper spray], (b) baton (c) Taser, (d) canine bite or (e) firearm" (Ariel et al., 2015, p. 521). Incidents were measured using four variables:

1. Total incidents that occurred during experiment shifts, as recorded by officers using a standardized police tracking system;
2. Total citizen complaints filed against officers (as a proxy of incidents), using a copyrighted software tool;
3. Rate of incidents per 1,000 police–public contacts, where total number of police–public contacts was recorded using the department's computer-aided dispatch system; and
4. Qualitative incident analysis, using videotaped content.

Key Findings

Ariel et al. (2015) concluded that the findings supported the overall hypothesis that wearing cameras increased police officers' compliance with rules of conduct around use of force, due to increased self-consciousness of being watched.

A feature of the evaluation was comparisons not only of the BWC shifts and the non-BWC shifts (the experimental design) but comparisons with data from months and years *before* the initiation of the study, as well as after implementation. Thus, the evaluation design included two complementary approaches. The data from the before–after component of the study showed that complaints by citizens for the whole department dropped from 28 in the year before the study, to just three during the year it was implemented; almost a 90% drop. Use-of-force incidents dropped from 61 in the year before implementation to 25 during implementation, a 60% drop.

When comparing the BWC shifts with the non-BWC (control) shifts, there were about half as many use-of-force incidents for the BWC shifts (eight as compared with 17 respectively). There was not a significant difference in number of citizen complaints, given how few there were during the year of the experiment.

The qualitative findings supported the main hypothesis in this evaluation.

Tying the findings back to the key questions of the study, the results indicated that wearing cameras did appear to increase the degree of self-awareness that the police officers had of their behavior and thereby could be used as a social control mechanism to promote socially desirable behavior.

More generally, the significance of the problem of police uses of force in their encounters with citizens is international is scope. Since the Rialto evaluation, there have been a large number of evaluations of similar programs in other U.S. cities, as well as cities in other countries (Cubitt et al., 2017; Maskaly et al., 2017). The widespread interest in this technology as an approach to managing use-of-force incidents has resulted in a large number of variations in how body-worn cameras have been deployed (for example, whether they must be turned on for all citizen encounters—that was true in Rialto—or whether officers can exercise discretion on whether to turn on the cameras), what is being measured as program outcomes, and what research designs/comparisons are conducted (U.S. Bureau of Justice, 2018; Cubitt et al., 2017).

Program Success Versus Understanding the Cause-and-Effect Linkages: The Challenge of Unpacking the Body-Worn Police Cameras "Black Box"

Even though the Rialto Police Department program was evaluated with a randomized controlled design, it presents us with a puzzle. It has been recognized that it may not have simply been the wearing of cameras that modified behaviors but an additional "treatment" wherein officers informed citizens (in an encounter) that the interaction was being recorded (Ariel et al., 2018a, 2018b; White, Todak, & Gaub, 2017). In fact, at least four different causal mechanisms can be distinguished:

1. One in which the cameras being on all the time changed police behavior.
2. A second in which the cameras being on all the time changed citizen behavior.
3. A third in which the cameras being on all the time changed police behavior and that, in turn, changed citizen behavior.
4. A fourth in which the body-worn cameras affect citizen behavior and that, in turn, affects police behavior.

Collectively, they create a challenge in interpreting the extent to which the cameras themselves affect officer behaviors and citizen behaviors. This challenge goes well beyond the Rialto experiment. By 2016, Barak Ariel and his colleagues had found, after 10 studies, that "in some cases they [BWCs] help, in some they don't appear to change police behavior, and in other situations they actually backfire, seemingly increasing the use of force" (Ariel, 2016, p. 36). This conundrum highlights the importance of working to determine the underlying mechanisms that cause a policy or program to change people's behavior.

Ariel et al. (2017), Hedberg et al. (2017), and Gaub et al. (2016) are three of the most recent studies to explore the contradictory findings from BWC research. The root of the problem is that we do not yet know *what the BWC mechanisms are that modify the behaviors of police or citizens when BWCs are in use.* Are the mechanisms situational, psychological, or organizational/institutional? If a theory of deterrence (see Ariel et al., 2018b; Hedberg et al., 2017) cannot adequately explain police and citizen behavioral outcomes of the use of BWCs, do other behavioral organizational justice theories (Hedberg et al., 2017; Nix & Wolfe, 2016) also have a role to play in our understanding? Deterrence theory relates to individual reactions to the possibility of being under surveillance, whereas organizational justice concepts, in the case of policing, relate to perceptions of procedural fairness in the organization. Nix and Wolfe (2016) take a closer look at organizational justice in the policing context and explain,

> The third, and most important, element of organizational justice is procedural fairness. Over and above outcome-based equity, employees look for supervisory decisions and organizational processes to be handled in procedurally just manners—decisions are clearly explained, unbiased, and allow for employee input. (p. 14)

So what mechanisms and theories might explain police and citizen changes in behavior when body-worn cameras are introduced into the justice system? As Ariel (2016) noted as the subtitle of his recent paper, *Body-worn cameras give mixed results, and we don't know why.*

Connecting Body-Worn Camera Evaluations to This Book

Although this textbook will use a variety of evaluations from different fields to illustrate points about evaluation theory and practice, body-worn-camera-related programs and their evaluations give us an opportunity to explore a timely, critical policy issue with international reach. We will pick up on the ways that evaluations of body-worn cameras intersect with different topics in our book: logic models, research designs, measurement issues, implementation issues, and the uses of mixed methods to evaluate programs.

The BWC studies offer us timely examples that can help evaluators to understand the on-the-ground implications of conducting defensible evaluations. Briefly, they are as follows:

- Body-worn camera programs for police forces have come into being in response to high-stakes sociopolitical problems—clearly there is **rationale** for such programs.
- Evaluation of BWC initiatives fit into varying components of the **performance management cycle**, including strategic planning and resource allocation, program and policy design, implementation and management, and assessing and reporting results.
- ***Ex ante*** studies have been conducted in some jurisdictions to examine police perceptions about the possibility of initiating BWC programs, before a BWC system is purchased and implemented.
- "Gold standard" **randomized controlled trials** have been conducted and have produced compelling evidence, yet the results of multiple studies are contradictory.
- Much can be learned from the **internal validity** and **construct validity** problems for BWC studies. For example, even in randomized settings, it is difficult to keep the "experimental" and the "control" group completely separate (in Rialto, the same officers were part of both the experimental and control groups suggesting **diffusion effects**—a construct validity problem).
- Local and organizational culture seems to be at the root of puzzling and sometimes contradictory evaluation results (an **external validity** issue).
- Existing data and **performance measures** are inconsistently defined and collected across communities, creating a challenge for evaluators wanting to synthesize existing studies as one of their lines of evidence.
- Many evaluations of BWCs include quantitative and qualitative **lines of evidence**.
- **Implementation** issues are as much a concern as the outcomes of BWC programs. There is so much variability in the way the BWCs are instituted, the policies (or not) on their uses, and the contexts in which they are introduced that it is difficult to pin down what this program is fundamentally about. (What is the core technology?) This is both an implementation problem and a construct validity problem.
- Governments and police forces are concerned with **cost-based analyses** and other types of economic evaluations but face challenges in quantitatively estimating actual costs and benefits of BWCs.

- BWC evaluators operate in settings where their options are constrained. They are challenged to develop a methodology that is defensible and to produce reports and recommendations that are seen to be credible and useful, even where, for example, there is resistance to the mandatory use of BWCs for the "experimental" police (as compared with the control group).
- The evaluators use their professional judgment as they design and implement their studies. Methods decisions, data collection decisions, interpretations of findings, conclusions, and recommendations *are all informed by judgment*. There is no template or formula to design and conduct such evaluations in particular settings. Instead, there are methodological approaches and tools that are applied by evaluators who have learned their craft and, of necessity, tackle each project as a craftsperson.

These points will be discussed and elaborated in 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.

Your own experience and practice will offer 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.

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 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.

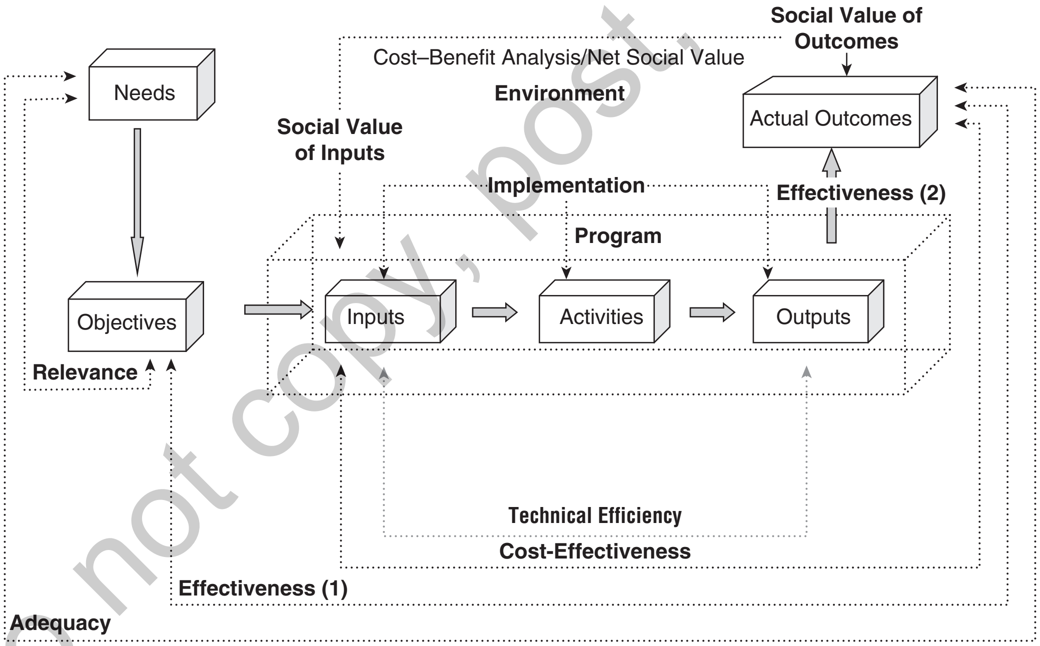
1. *What is the need for a program?*

A needs assessment can occur either before program options are developed (an *ex ante* needs assessment) or during their 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.

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. 20).

Needs assessment done before a program is developed can inform the way that the objectives are stated, and suggest performance measures and 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 governments, 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. Programs often elaborate policies. 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 (if there are such theories of change)?
- 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 means to achieve the objectives, given the context in which the program will be implemented?

We discuss program theories and program logics in Chapter 2.

4. *Was the program implemented as intended?*

Assessing implementation involves an examination of the **program inputs, program 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 (components are typically clusters of activities), 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 in which 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 (what activities have happened), identifying and describing any bottlenecks in the processes, and seeing whether outputs have been produced for different activities. Since the outputs of most programs are necessary (but not sufficient) to produce 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, 2 years old or less), it is quite possible that gaps will emerge between *descriptions* of intended program activities and what is *actually* getting done. One way to assess implementation is to examine the fidelity between intended and actual program components, activities, and even outputs (Century, Rudnick, & Freeman, 2010). Indeed, if the gaps are substantial, a program evaluator may elect to recommend an analysis that focuses on just 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 would be the comparison of the actual cost per patient day with a benchmark cost per patient day if the hospital were fully utilized. Economic evaluation issues are examined in Chapter 7.

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 5 months of the beginning of that program, British Columbia also adopted a mandatory seatbelt law, making it impossible to tell whether Counterattack was responsible (at a province-wide level) for the observed downward trend in accidents that happened. In effect, the seatbelt law was a **rival hypothesis** that could plausibly explain the outcomes of the Counterattack Program.

Performance measures are often intended to track whether policies and programs achieve their intended objectives (usually, yearly outcome targets are specified). Measuring performance is not equivalent to evaluating the effectiveness (1) 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, 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 job 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 and operate 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 a 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.

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). Three 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 (Effectiveness [1]). Performance measurement systems monitor and can tell us whether a program "achieved" its intended outcomes (Effectiveness [2]). Program evaluations are intended to answer "why" questions.

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.

DESIGNING AND CONDUCTING AN EVALUATION IS NOT A LINEAR PROCESS

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 next imply that a typical program evaluation is a project, with a beginning and an end point. This is still the mainstream view of evaluation practice, but others have argued 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 engaged with 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 summarizes 10 questions that are important as part of evaluation 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 that it should be straightforward to implement the evaluation project, should it proceed. 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.

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, and who are the stakeholders?
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 and structure of the program?
6. Which research design alternatives are desirable and feasible?
7. What kind of environment does the program operate in, and how does that affect the comparisons available to an evaluator?
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 which is defensible?
10. Should the evaluation be undertaken?

Steps in conducting and reporting an evaluation

1. Develop the data collection instruments, and pre-test 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.

Assessing the Feasibility of the Evaluation

1. *Who are the clients for the evaluation, and who are the other stakeholders?*

Program evaluations are substantially *user driven*. Michael Patton (2008) makes a **utilization focus** 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 will have access to the results, 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 clients (those commissioning the evaluation) 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 and audience 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:

- Why do the clients want it done?
- What are the main evaluation issues that the clients want addressed? (Combinations of the 10 evaluation questions summarized in Table 1.1 are usually in play).

- Are there hidden agendas or covert reasons for wanting the policy or program evaluated? For example, how might the program organization or the beneficiaries be affected?
- Is the evaluation intended to be for incremental adjustments/improvements, 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 organizational expenditures and greater degrees of control by the evaluator. 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). This can have ethical implications—withholding a program from vulnerable persons or families can cause harm (Rolston, Geyer and Locke, 2013). We discuss the ethics of evaluations in Chapter 12.

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

- Time
- Human resources, including persons with necessary knowledge, 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. Bamberger, Rugh, Church, and Fort (2004) have suggested strategies for designing impact evaluations with very modest resources—they call their approach **shoestring evaluation**. Another recently introduced approach is rapid impact evaluation (Government of Canada, 2018; Rowe, 2014). 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 theoretical and empirical work have been done, there is growing emphasis on approaches that take advantage of the availability of consolidations 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 (2018) is an international project begun in 1993 that is aimed at conducting systematic reviews of health-related interventions. They also produce the *Cochrane Handbook for Systematic Reviews of 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 (2018) is an organization that is focused on the social sciences and education. Founded in 1999, its mission is to promote "positive social and economic change through the production and use of systematic reviews and other evidence synthesis for evidence-based policy and practice."

The Government Social Research Unit in the British government has published a series of guides, including *The Magenta Book: Guidance for Evaluation* (HM Treasury, 2011). Chapter 6 in *The Magenta Book*, "Setting Out the Evaluation Framework," includes advice on using existing research in policy evaluations. Literature reviews and quantitative and qualitative systematic reviews are covered. The main point here 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/comparisons, 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 and logic 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 (Funnell & Rogers, 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).

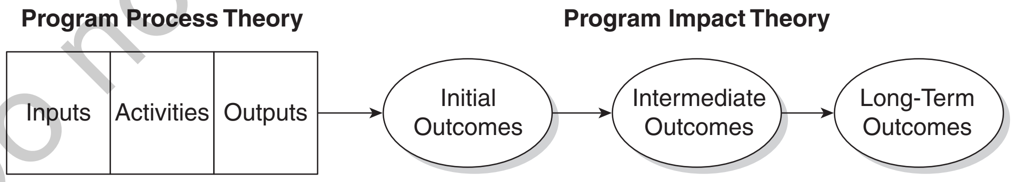
An example of a basic schema for a logic model is illustrated in Figure 1.5. The model shows the stages in a typical logic model: program process (including outputs) and outcomes. We will be discussing logic models in some detail in Chapter 2 of this textbook.

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.

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 that 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. As well, many public-sector policy objectives are challenging to measure.

FIGURE 1.5 LINEAR PROGRAM LOGIC MODEL



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

6. *Which research design alternatives are desirable and appropriate?*

Key to evaluating the effectiveness of a program are comparisons that allow us to estimate the incremental impacts of the program, ideally over what would have happened if there had been no intervention. 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.

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

- 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.)
- How is the environment of this program different from other locales where a similar program has been initiated?

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 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 (ICBC). 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). We look at the strengths and weaknesses of different research designs in Chapter 3.

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

Programs, as **open systems**, are always embedded in an environment. The ways that the environmental factors—other programs, organizational leaders, other departments in the government, central agencies, funders, as well as the economic, political, and social context—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 all 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. In sum, identifying appropriate comparisons to answer the evaluation questions are typically conditioned by the contexts in which a program (and the evaluation) are embedded.

8. *What information/data sources are available and appropriate, given the evaluation questions, the program structure, the comparisons that would be appropriate, 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? (e.g., baseline data, other studies)
- 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 would plausibly 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, 2016a; 2016b), have linked performance data to program evaluations, with the stated 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 point to make with respect to potential data sources. Evaluations that focus 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/table that displays the evaluation questions and subquestions as rows, and the prospective data sources or lines of evidence 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 (which may involve direct observations).

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; Donaldson, Christie, & Mark, 2014; 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, particularly for high-stakes summative evaluations.

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 proceed with the evaluation and if the methodology has been determined during the feasibility stage, there are five more steps that are common to most evaluations.

1. *Develop the measures, and pre-test 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), **pre-testing** 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. Pre-testing 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 pre-tested. 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/lines of evidence 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 narrowly avoided litigation.

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. Often, 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 generalize from survey samples to populations, lower response rates weaken any generalizations. A significant problem now is that people increasingly feel they are oversurveyed. 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, 2017).

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 an 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 will 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.
- Structure the report so that it reflects the evaluation questions and subquestions that are driving the evaluation—once the executive summary, table of contents, lists of figures and tables, the introductory section of the report, and the methodology section of the report have been written, turn to the evaluation questions, and for each one, discuss the findings from the relevant lines of evidence.
- Conclusions should synthesize the findings for each evaluation question and form the basis for any recommendations that are written.
- 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 postreporting, 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. We discuss these issues in Chapters 11 and 12 of this textbook.

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, 1998a).

The whole issue of whether and to what extent evaluations are used continues to be an important topic in the field. Although there is clearly a view that the quality of an evaluation rests on its methodological defensibility (Fitzpatrick, 2002), many evaluators have taken the view that evaluation use is a more central objective 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 design of an existing policy or program
- Making incremental changes to the way the existing policy or program is implemented
- 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

These changes would reflect **instrumental uses** of evaluations (direct uses of evaluation products). In addition, there are **conceptual uses** (the knowledge from the evaluation becomes part of the background in the organization and influences other programs at other times) and **symbolic uses** (the evaluation is used to rationalize or legitimate decisions made for political reasons) (Kirkhart, 2000; Højlund, 2014; Weiss, 1998b). More recently, uses have been broadened to include **process uses** (effects of the process of doing an evaluation) and misuses of evaluations (Alkin & King, 2016; Alkin & King, 2017).

Some jurisdictions build in a required management response to program evaluations. The federal government of Canada, for example, requires the program being evaluated to respond to the report with a management response that addresses each recommendation, indicates whether the program agrees with the recommendation, if not why not, and if so, the actions that will be taken to implement each recommendation (Treasury Board of Canada, 2016a; 2016b). This process is intended to ensure that there is instrumental use of each evaluation report.

Evaluations are one source of information in policy and program 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 (Alkin & King, 2017).

✓ EVALUATION AS PIECEWORK: WORKING COLLABORATIVELY WITH CLIENTS AND PEERS

In this chapter, we have outlined a process for designing and conducting evaluations, front to back. But evaluation engagements with clients can divide up projects so that the work is distributed. For example, in-house evaluators may do the overall design for the project, including specifying the evaluation questions, the lines of evidence, and perhaps even the methodologies for gathering the evidence. The actual data collection, analysis, and report writing may be contracted out to external evaluators. Working collaboratively in such settings where one or more stages in a project are shared, needs to be balanced with evaluator independence. Competent execution of specific tasks is part of what is expected in today's evaluation practice, particularly where clients have their own in-house evaluation capacity. In Chapter 12, we talk about the importance of teamwork in evaluation—teams can include coworkers and people from other organizations (including client organizations).

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. The core of this book is our focus on evaluating the effectiveness of policies and programs. This includes an emphasis on understanding the difference between outcomes that occur due to a program and outcomes that may have changed over time due to factors other than the program (that is, the counterfactual). We believe that is what distinguishes evaluation from other related fields. 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 40-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 those 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, our focus on program effectiveness is systematic. 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 values, ethics, and judgment calls. Programs are embedded in values and are driven by values. Program objectives are value statements—they state what programs should do. 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 values, experiences, beliefs, 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 research applications. In many program evaluations, resource and contextual constraints mean that the tools we apply are not ideal for the situation at hand. Also, more and more, evaluators must consider issues such as organizational culture, political culture, social context, and the growing recognition of the importance of “voice” for groups of people who have been marginalized.

That is, there is more to evaluation that simply determining whether a program or policy is “effective.” Effective for whom? There is growing recognition that as a profession, evaluators have an influence in making sure voices are equitably heard. 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). Professional judgment is both about disciplining our own role in evaluation practice as well as becoming more self-aware (and ethical) as practitioners.

Fourth, the importance of program evaluation and performance measurement in contemporary public and nonprofit organizations is related to a continuing, broad international movement to manage for results. Performance management depends on having credible information about how well programs 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 were 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 comfortable combining qualitative and quantitative information?
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?
7. What do you think it means to be objective? Do you think it is possible to be objective in the work we do as evaluators? In anything we do? Offer some examples of reasons why you think it is possible to be objective (or not).

REFERENCES

Alkin, M. C., & King, J. A. (2017). Definitions of evaluation use and misuse, evaluation influence, and factors affecting use. *American Journal of Evaluation*, *38*(3), 434–450.

Alkin, M. C., & King, J. A. (2016). The historical development of evaluation use. *American Journal of Evaluation*, *37*(4), 568–579.

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.

Ariel, B. (2016). The puzzle of police body cams: Body-worn cameras give mixed results, and we don't know why. *IEEE Spectrum*, *53*(7), 32–37.

Ariel, B., Farrar, W. A., & Sutherland, A. (2015). The effect of police body-worn cameras on use of force and citizens' complaints against the police: A randomized controlled trial. *Journal of Quantitative Criminology*, *31*(3), 509–535.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J.,. . . & Henderson, R. (2016). Wearing body cameras increases assaults against officers and does not reduce police use of force: Results from a global multi-site experiment. *European Journal of Criminology*, *13*(6), 744–755.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J.,. . . & Henderson, R. (2017). "Contagious accountability": A global multisite randomized controlled trial on the effect of police body-worn cameras on citizens' complaints against the police. *Criminal Justice and Behavior*, *44*(2), 293–316.

Ariel, B., Sutherland, A., Henstock, D., Young, J., Drover, P., Sykes, J.,. . . & Henderson, R. (2018a). Paradoxical effects of self-awareness of being observed: Testing the effect of police body-worn cameras on assaults and aggression against officers. *Journal of Experimental Criminology*, *14*(1), 19–47.

Ariel, B., Sutherland, A., Henstock, D., Young, J., & Sosinski, G. (2018b). The deterrence spectrum: Explaining why police body-worn cameras 'work' or 'backfire' in aggressive police–public encounters. *Policing: A Journal of Policy and Practice*, *12*(1), 6–26.

Arnaboldi, M., Lapsley, I., & Steccolini, I. (2015). Performance management in the public sector: The ultimate challenge. *Financial Accountability & Management*, *31*(1), 1–22.

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.

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.

Barber, M. (2015). *How to run a government: So that citizens benefit and taxpayers don't go crazy*. London, UK: Penguin.

Barber, M., Moffit, A., & Kihn, P. (2011). *Deliverology 101: A field guide for educational leaders*. Thousand Oaks, CA: Corwin.

Bickman, L. (1996). A continuum of care. *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.

Bryson, J. M., Crosby, B. C., & Bloomberg, L. (2014). Public value governance: Moving beyond traditional public administration and the new public management. *Public Administration Review*, *74*(4), 445–456.

Campbell Collaboration. (2018). Our vision, mission and key principles. Retrieved from https://www.campbellcollaboration.org/about-campbell/vision-mission-and-principle.html

Center for Evidence-Based Crime Policy at George Mason University (2016). Retrieved from http://cebcp.org/technology/body-cameras

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*(2), 199–218.

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.

Cochrane Collaboration. (2018). About us. Retrieved from www.cochrane.org/about-us. Also: *Cochrane handbook for systematic reviews of interventions*, retrieved from http://training.cochrane.org/handbook

Coen, D., & Roberts, A. (2012). A new age of uncertainty. *Governance*, *25*(1), 5–9.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand-McNally.

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., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks: Sage.

Creswell, J. W., & Plano Clark, V. (2017). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: Sage.

Cubitt, T. I., Lesic, R., Myers, G. L., & Corry, R. (2017). Body-worn video: A systematic review of literature. *Australian & New Zealand Journal of Criminology*, *50*(3), 379–396.

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

de Lancer Julnes, P., & Steccolini, I. (2015). Introduction to symposium: Performance and accountability in complex settings—Metrics, methods, and politics. *International Review of Public Administration*, *20*(4), 329–334.

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

Donaldson, S. I., Christie, C. A., & Mark, M. M. (Eds.). (2014). *Credible and actionable evidence: The foundation for rigorous and influential evaluations*. Los Angeles, CA: Sage.

Donaldson, S. I., & Picciotto, R. (Eds.). (2016). *Evaluation for an equitable society*. Charlotte, NC: Information Age Publishing.

Dunleavy, P., Margetts, H., Bestow, S., & Tinkler, J. (2006). New public management is dead—Long live digital-era governance. *Journal of Public Administration Research and Theory*, *16*(3), 467–494.

Farrar, W. (2013). *Self-awareness to being watched and socially-desirable behavior: A field experiment on the effect of body-worn cameras and police use-of-force*. Washington, DC: Police Foundation.

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.

Gaub, J. E., Choate, D. E., Todak, N., Katz, C. M., & White, M. D. (2016). Officer perceptions of body-worn cameras before and after deployment: A study of three departments. *Police Quarterly*, *19*(3), 275–302.

Gauthier, B., Barrington, G. V., Bozzo, S. L., Chaytor, K., Dignard, A., Lahey, R.,. . . Roy, S. (2009). The lay of the land: Evaluation practice in Canada in 2009. *The Canadian Journal of Program Evaluation*, *24*(1), 1–49.

Gilmour, J. B. (2007). Implementing OMB's Program Assessment Rating Tool (PART): Meeting the challenges of integrating budget and performance. *OECD Journal on Budgeting*, *7*(1), 1C.

Government of British Columbia. (2007). Greenhouse Gas Reduction Targets Act. British Columbia: Queen's Printer. Retrieved from http://www.bclaws.ca/EPLibraries/bclaws_new/document/ID/freeside/00_07042_01#section12

Government of British Columbia. (2014). Greenhouse Gas Industrial Reporting and Control Act. Retrieved from http://www.bclaws.ca/civix/document/id/lc/statreg/14029_01

Government of British Columbia. (2016). Climate leadership. Victoria, BC: Government of British Columbia. Retrieved from http://climate.gov.bc.ca

Government of Canada. (2018). Guide to rapid impact evaluation. Retrieved from https://www.canada.ca/en/treasury-board-secretariat/services/audit-evaluation/centre-excellence-evaluation/guide-rapid-impact-evaluation.html

Greiling, D., & Halachmi, A. (2013). Accountability and organizational learning in the public sector. *Public Performance & Management Review*, *36*(3), 380–406.

Hedberg, E., Katz, C. M., & Choate, D. E. (2017). Body-worn cameras and citizen interactions with police officers: Estimating plausible effects given varying compliance levels. *Justice Quarterly*, *34*(4), 627–651.

HM Treasury, Government of the United Kingdom. (2011). *Magenta book: Guidance for evaluation*. Retrieved from https://www.gov.uk/government/publications/the-magenta-book

Højlund, S. (2014). Evaluation use in the organizational context–changing focus to improve theory. *Evaluation*, *20*(1), 26–43.

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

House, E. R. (2016). The role of values and evaluation in thinking. *American Journal of Evaluation*, *37*(1), 104–108.

Hunter, D., & Nielsen, S. (2013). Performance management and evaluation: Exploring complementarities. *New Directions in Evaluation*, *137*, 7–17.

Jennings, W. G., Fridell, L. A., & Lynch, M. D. (2014). Cops and cameras: Officer perceptions of the use of body-worn cameras in law enforcement. *Journal of Criminal Justice*, *42*(6), 549–556.

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.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation*, *88*, 5–23.

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.

Kroll, A. (2015). Drivers of performance information use: Systematic literature review and directions for future research. *Public Performance & Management Review*, *38*(3), 459–486.

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

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

Lindquist, E. A., & Huse, I. (2017). Accountability and monitoring government in the digital era: Promise, realism and research for digital-era governance. *Canadian Public Administration*, *60*(4), 627–656.

Lum, C., Koper, C., Merola, L., Scherer, A., & Reioux, A. (2015). *Existing and ongoing body worn camera research: Knowledge gaps and opportunities*. Fairfax, VA: George Mason University.

Mahler, J., & Paul Posner, P. (2014). Performance movement at a crossroads: Information, accountability and learning. *International Review of Public Administration*, *19*(2), 179–192.

Majone, G. (1989). *Evidence, argument, and persuasion in the policy process*. London, UK: Yale University Press.

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

Maskaly, J., Donner, C., Jennings, W. G., Ariel, B., & Sutherland, A. (2017). The effects of body-worn cameras (BWCs) on police and citizen outcomes: A state-of-the-art review. *Policing: An International Journal of Police Strategies & Management*, *40*(4), 672–688.

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

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.

Melkers, J., & Willoughby, K. (2004). *Staying the course: The use of performance measurement in state governments*. Washington, DC: IBM Center for the Business of Government.

Melkers, J., & Willoughby, K. (2005). Models of performance-measurement use in local government: Understanding budgeting, communication, and lasting effects. *Public Administration Review*, *65*(2), 180–190.

Moynihan, D. P. (2006). Managing for results in state government: Evaluating a decade of reform. *Public Administration Review*, *66*(1), 77–89.

Moynihan, D. P. (2013). *The new federal performances system: Implementing the new GPRA Modernization Act*. Washington, DC: IBM Center for the Business of Government.

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

Newcomer, K., & Brass, C. T. (2016). Forging a strategic and comprehensive approach to evaluation within public and nonprofit organizations: Integrating measurement and analytics within evaluation. *American Journal of Evaluation*, *37*(1), 80–99.

Nix, J., & Wolfe, S. E. (2016). Sensitivity to the Ferguson effect: The role of managerial organizational justice. *Journal of Criminal Justice*, *47*, 12–20.

OECD. (2015). *Achieving public sector agility at times of fiscal consolidation*, OECD Public Governance Reviews. Paris, France: OECD Publishing.

Office of Management and Budget. (2012). *Office of Management and Budget [Obama archives]*. Retrieved from https://obamawhitehouse.archives.gov/omb/organization_mission/

Office of Management and Budget. (2018). *Office of Management and Budget*. Retrieved from https://www.whitehouse.gov/omb

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

Osborne, S. P. (Ed.). (2010). *The new public governance: Emerging perspectives on the theory and practice of public governance*. London, UK: Routledge.

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, NY: Guilford Press.

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

Pollitt, C., & Bouckaert, G. (2011). *Public management reform* (2nd and 3rd ed.). Oxford, UK: Oxford University Press.

Public Safety Canada (2018). Searchable website: https://www.publicsafety.gc.ca/

Radin, B. (2006). *Challenging the performance movement: Accountability, complexity, and democratic values*. Washington, DC: Georgetown University Press.

Rolston, H., Geyer, J., & Locke, G. (2013). *Final report: Evaluation of the Homebase Community Prevention Program*. New York, NY: ABT Associates. Retrieved from http://www.abtassociates.com/AbtAssociates/files/cf/cf819ade-6613–4664–9ac1–2344225c24d7.pdf

Room, G. (2011). *Complexity, institutions and public policy: Agile decision making in a turbulent world*. Cheltenham, UK: Edward Elgar.

Rowe, A. (2014). *Introducing Rapid Impact Evaluation (RIE): Expert lecture*. Retrieved from https://evaluationcanada.ca/distribution/20130618_rowe_andy.pdf

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.

Schwandt, T. (2015). *Evaluation foundations revisited: Cultivating a life of the mind for practice*. Stanford, CA: Stanford University Press.

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. (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.

Shaw, T. (2016). Performance budgeting practices and procedures. *OECD Journal on Budgeting*, *15*(3), 1–73.

Stockmann, R., & Meyer, W. (Eds.). (2016). *The future of evaluation: Global trends, new challenges and shared perspectives*. London, UK: Palgrave Macmillan.

Szanyi, M., Azzam, T., & Galen, M. (2013). Research on evaluation: A needs assessment. *Canadian Journal of Program Evaluation*, *27*(1), 39–64.

Treasury Board of Canada Secretariat. (2016a). Policy on results. Retrieved from http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=31300&section=html

Treasury Board of Canada Secretariat. (2016b). Directive on results. Retrieved from https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=31306&section=html

U. S. Bureau of Justice Assistance (2018). Body-worn camera toolkit, U.S. department of justice: Bureau of justice assistance. Retrieved from https://www.bja.gov/bwc/resources.html

Van de Walle, S., & Cornelissen, F. (2014). Performance reporting. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford handbook on public accountability* (pp. 441–455). Oxford, UK: Oxford University Press.

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

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

Weiss, C. H. (1998b). Have we learned anything new about the use of evaluation? *American Journal of Evaluation*, *19*(1), 21–33.

White, M. D., Todak, N., & Gaub, J. E. (2017). Assessing citizen perceptions of body-worn cameras after encounters with police. *Policing: An International Journal of Police Strategies & Management*, *40*(4), 689–703.

World Bank. (2014). *Developing in a changing climate. British Columbia's carbon tax shift: An environmental and economic success* (Blog: Submitted by Stewart Elgie). from http://blogs.worldbank.org/climatechange/print/british-columbia-s-carbon-tax-shift-environmental-and-economic-success

World Bank. (2017). State and Trends of Carbon Pricing 2017. Washington, DC: World Bank. © World Bank. https://www.openknowledge.worldbank.org/handle/10986/28510 License: CC BY 3.0 IGO.

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