2

FIRST CLARIFY THE PURPOSE

Scoping the Evaluation

In this chapter, we introduce Step 1 of the RealWorld Evaluation (RWE) approach—"scoping the evaluation." We begin by considering the widely different expectations that **clients** can have about the purpose and nature of evaluation and what they understand by and expect from an impact evaluation. It is important to understand client information needs, their assumptions on how the evaluation should be conducted, and how clients expect to use the information produced by the evaluation. We also point to the need to identify other **stakeholders** and the nature and degree to which they should be involved in an evaluation.

In recent years, and particularly with the increased attention to systems analysis and complexity-responsive evaluation, the importance of a clear definition of the boundaries of the evaluation is recognized. "Boundaries determine what is *in* and what is *out* of any endeavor" (Williams & van't Hof, 2014). Boundaries determine what is valued in any program, and also in the evaluation. For example, is the goal of the program to maximize employment and economic growth, or to ensure that all sectors of the target population have equitable access to program benefits and a voice in how the program is designed and evaluated? Different values and perspectives require different evaluation designs. Different stakeholders have different perceptions on the underlying values the program is seeking to achieve and consequently how boundaries should be defined. Using a boundary framework presents various challenges for the evaluator. First, often values are implicit in the program design and not clearly stated, so the evaluator must elicit the values held by different stakeholders. Second, how can values of different stakeholders be incorporated into the evaluation? Third, how should this understanding of values be reflected in how the boundaries of the evaluation are defined? Boundaries determine the kinds of questions that are asked and who is covered in the evaluation. Boundaries also affect the evaluation design by determining the size of the group affected by the program, the range of outcomes to be assessed, and the period over which outcomes are measured. Decisions on each of these have an important influence on the operational utility of the evaluation, its cost and complexity, and the extent to which equity issues are addressed.

We use **program theory models** (see Chapter 10) to articulate the assumptions on which the project **design** was based and to ensure the evaluation focuses on the issues of concern to stakeholders. Program theory also helps us understand how project implementation, **outcomes**, and impacts are affected by the political, economic, institutional, environmental, and cultural

BOX 2.1

Boundaries and Boundary Choices

"Boundaries determine what is 'in' and what is 'out' of any endeavor. In evaluation, boundary choices have a special purpose. They identify the criteria by which we evaluate the intervention since criteria delineate between what has merit, worth and significance, and what does not. . . . Not only do [boundaries] help determine the evaluation criteria but they also force you to consider the ethical and political implications of the intervention. . . . [The approach adopted in this book] formally acknowledges that evaluation design has practical, political and ethical implications, and that these need to be deliberated on as part of the design process."

Source: Williams & van't Hof (2014).

context within which each project is implemented, and contextual analysis should form part of the scoping study. RWEs use both qualitative (**QUAL**) and quantitative (**QUANT**) methodologies, and there should be no a priori preference for either, and there are many advantages in using mixed-method designs that draw on the strengths of both QUAL and QUANT methodologies. The chapter concludes by showing how the scoping phase is used to identify cost, time, data, and political constraints that a particular evaluation will face and how this analysis is used to identify and assess the possible RWE designs that could be used.

1. STAKEHOLDER EXPECTATIONS OF IMPACT EVALUATIONS

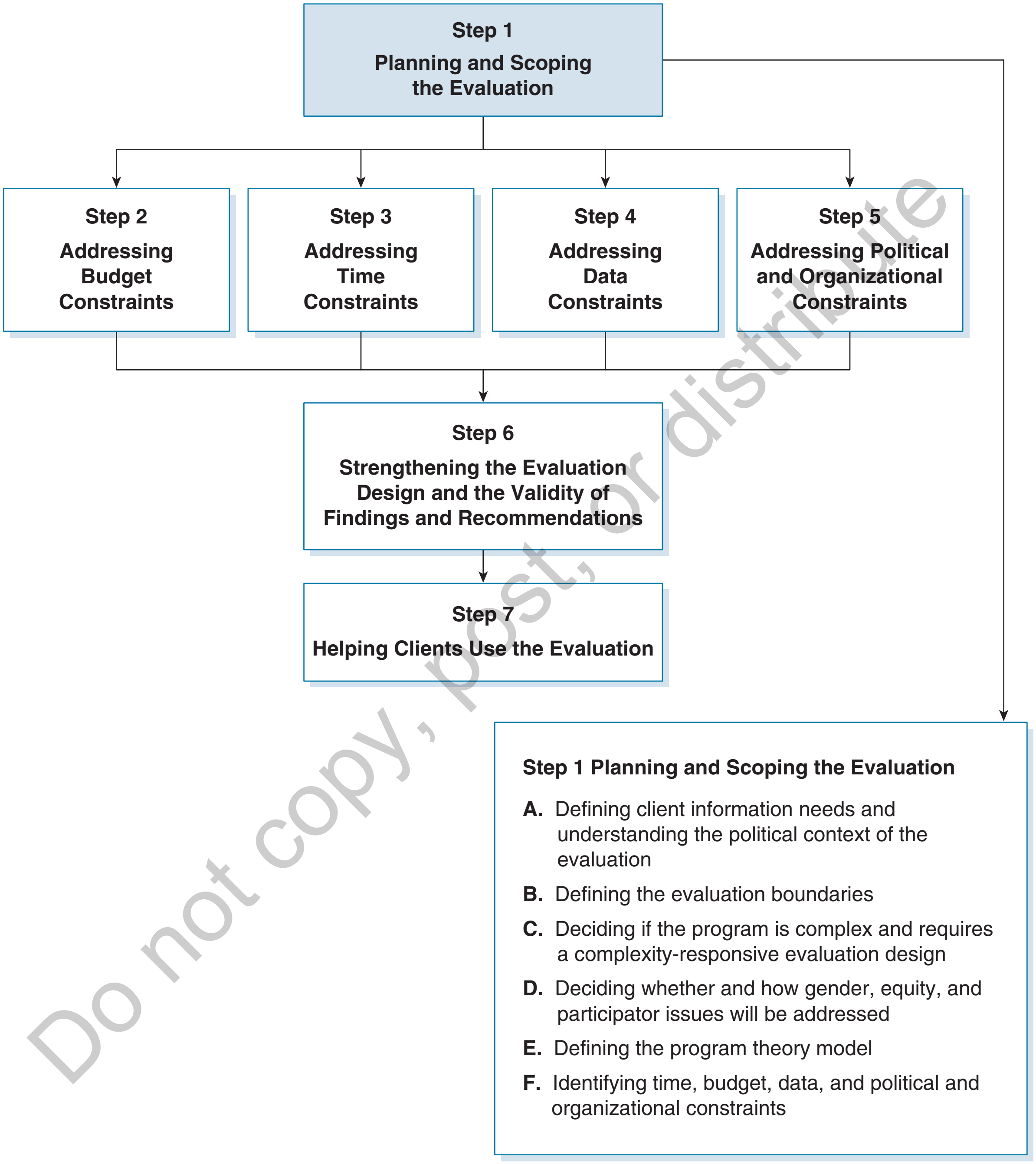
Before discussing the different possible impact evaluation designs, it is first necessary to determine whether in fact an impact evaluation is required. Evaluations are conducted for many different reasons, including impact assessment, accountability to funding agencies, transparency, improving program implementation, ensuring programs are reaching and benefiting poor and vulnerable groups, and ensuring resources are being used in an efficient and cost-effective manner.[1] Many of the above issues do not involve the assessment of impacts.

There is a wide variety of understandings of what is involved in conducting an impact evaluation and what can be expected from the results. These include those who believe that every impact evaluation must be a sophisticated, "scientific," randomized, or quasi-experimental design.[2] On the other end of the continuum are those who believe that QUAL methods are needed to understand programs and their impacts and how they are experienced by stakeholders. And there are

[1] The Organization for Economic Development/Development Advisory Committee (OECD/DAC) evaluation guidelines identify five sets of evaluation criteria that are widely used by development agencies as the framework for conducting evaluations: relevance, effectiveness, efficiency, impact, and sustainability. In fact, many agencies often do not include impact as one of the criteria when commissioning evaluations.

[2] See, for example, MIT's Poverty Action Lab (www.povertyactionlab.com) and the International Initiative for Impact Evaluation (3ie; www.3ieimpact.org).

FIGURE 2.1 ● Step 1: Scoping the Evaluation


Step 1
Planning and Scoping the Evaluation
Step 2
Addressing Budget Constraints
Step 3
Addressing Time Constraints
Step 4
Addressing Data Constraints
Step 5
Addressing Political and Organizational Constraints
Step 6
Strengthening the Evaluation Design and the Validity of Findings and Recommendations
Step 7
Helping Clients Use the Evaluation

Step 1 Planning and Scoping the Evaluation

A. Defining client information needs and understanding the political context of the evaluation

B. Defining the evaluation boundaries

C. Deciding if the program is complex and requires a complexity-responsive evaluation design

D. Deciding whether and how gender, equity, and participator issues will be addressed

E. Defining the program theory model

F. Identifying time, budget, data, and political and organizational constraints those who prefer multisite or multiprogram studies to examine broader impact. In later chapters we will identify at least six impact evaluation approaches that can be considered when selecting the best design for any conventional evaluation plus two additional designs for evaluating complex programs or where the evaluation incorporates big data (see Chapters 7 and 11). The choice of evaluation design is influenced by the size and complexity of the program being evaluated, the context within which the intervention is implemented, and the specific purpose of the evaluation—as well as by the methodological preference of stakeholders. As we will explore in more depth in subsequent chapters, RWE budget, time, data, and political constraints can also affect the choice of methods.

2. UNDERSTANDING INFORMATION NEEDS

The agencies responsible for commissioning and conducting the evaluation must consider its purpose and therefore which designs would be appropriate and feasible. Table 1.1 (Chapter 1) shows that RWE can be commissioned at the beginning of a program, during implementation, or at the end; it also describes the different purposes for which evaluations are used at each of these points in the program cycle. The process of defining the evaluation purpose begins with a stakeholder analysis to understand the expectations of key stakeholders and often to negotiate with them what should and can be done, given constraints of money, time, data availability, and political considerations.

A clear understanding of the priorities and information needs of clients and other key stakeholders is an essential first step in the design of a good evaluation and an effective way for the RealWorld evaluator to eliminate unnecessary data collection and analysis, hence reducing cost and time.

While it is usually a simple matter to define the evaluation clients (those commissioning the evaluation), a more difficult issue is to define the range of stakeholders whose concerns should be taken into account in the evaluation design, implementation, and dissemination. Time and budget constraints often limit the range of stakeholders who can be consulted and involved. The evaluator should try to assess whether these constraints exclude some important groups—particularly, vulnerable groups who may be difficult to reach. It is useful to distinguish between primary stakeholders, who are consulted regularly throughout the evaluation, and secondary stakeholders, whose role in the evaluation is less clearly defined and who may not be consulted on a regular basis. Evaluators and clients may sometimes disagree on who is a stakeholder and who should be consulted. This can become a sensitive issue if the evaluator believes that certain groups who are affected by the project should be consulted and the client wishes to limit consultations to the primary stakeholders.

Typically, an evaluator (or evaluation team) is commissioned to conduct an evaluation according to the terms set forth by the client. Some agencies call these the **terms of reference** (**ToR**), while other agencies refer to the Scope or Statement of Work (SoW). But when should a conscientious evaluator propose that other stakeholders' perspectives be included and that the evaluation be made relevant to them? As Robert Chambers (1997) asks, "Whose reality counts?" Related to the previous point, *Realist Evaluation* (Pawson, 2006; Pawson & Tilley, 1997) suggests asking the following questions:

1. Who benefits from the program?
2. How do they benefit?

3. When do they benefit?
4. Why?
5. Who does not benefit and why?

An important part of the scoping phase is to clarify the **evaluand**, that is, what is being evaluated. As White and Bamberger (2008) have pointed out, evaluators need to pay more attention to the "factual" (the evaluand), as many evaluation designs are based on wrong or at least untested assumptions about how the program actually works. So among issues that need to be addressed during the evaluability assessment are the following:

1. How well developed and tested is the program design, and is this adequately captured in the program theory framework on which is the evaluation is based?
2. If the program is still in a pilot development phase, the evaluation will need to assess how well the program's organization and delivery systems work. It is probably not worth conducting a rigorous assessment of impacts if the basic systems are not yet tested or working.
3. If the program has been operating for some time and is well tested, it is possible to consider a more rigorous assessment of outcomes and impacts.

Appendix 2.1 presents a checklist identifying 14 dimensions that must be taken into consideration when designing the evaluation. The first 11 describe the characteristics of the evaluand (program being evaluated), while the final three refer to methodological dimensions referring both to client and stakeholder preferences and what is feasible within budget, time, and data constraints. For example, the appropriate evaluation design would be quite different when evaluating a complex, national-level intervention with a large evaluation budget and for which the evaluation is being planned before the intervention than it would be for the evaluation of a small project with a small evaluation budget and for which the evaluation does not begin until the project is nearing completion. The purpose of the evaluation will also influence the appropriate evaluation design, as will the type of client who commissions the evaluation and the skills of the consultant(s) hired to conduct the evaluation. It makes no sense to discuss the "best" evaluation design until all of these dimensions are fully understood.

Once the evaluation design options have been narrowed down, there will always be several different ways that the evaluation could be designed (see Appendix 2.1 and Chapter 11). The design must reflect the methodological preferences of the client and key stakeholders as well as the constraints imposed by the evaluation scenario.

An evaluability assessment may reveal that the scope of the original ToR must be modified if some of the proposed questions cannot be addressed at this time and within the evaluation's budget, time, and data constraints. For example, it may be too early in the life of the program to measure impacts, the lack of comparative data may limit the use of more rigorous statistical designs, the absence of relevant secondary data or its poor quality and reliability may limit the possibility of reconstructing reliable baseline data, and political or ethical constraints may limit the design (e.g., selecting control individuals or groups for randomized control trials), the people who can be interviewed, or the questions that can be asked.

It is also important to understand the context within which the evaluation is to be conducted. Many evaluation designs are based on only a limited understanding of the evaluation context—often assuming that the project will be implemented as planned, without understanding the many political, cultural, organizational, economic, and perhaps environmental factors that could affect how the project is actually implemented and who benefits. Some of the factors that could be considered include:

1. The ethnic composition of the target population and any conflicts and divisions that could make it difficult for the project to reach and benefit the whole population
2. The need to understand gender relations
3. Population movements and demographic trends
4. How the local and national political context may affect the project
5. The multiple effects of the local and national economic situation

Meeting as early as possible with clients and key stakeholders helps ensure that the reasons for commissioning the evaluation are understood. It is particularly important to understand policy and operational decisions to which the evaluation will contribute and to agree on the level of precision required in making these decisions. Typical questions that decision makers must address include the following:

1. Is there evidence that the project achieved (or will achieve) its objectives? Which objectives were (or will be) achieved and which were not (or will not be) achieved? Why?
2. Did the project aim for the right objectives? Were the underlying causes of the problem(s) the project is designed to ameliorate accurately diagnosed and adequately addressed?
3. Are outcomes sustainable and benefits likely to continue?
4. What internal and/or external contextual factors determine the degree of success or failure?
5. Did the program satisfy the Organization for Economic Cooperation and Development/Development Advisory Committee (OECD/DAC) criteria of relevance, effectiveness, efficiency, impact, and **sustainability**?

Many of these questions do not require a high level of statistical precision, but they do require reliable answers to additional questions:

1. Are there measurable changes in the characteristics of the **target population** with respect to the impacts the project was intended to produce?
2. What impact has the project had on different subsets of the target population, including the poorest and most vulnerable groups? Are there different impacts on men and women? Are there ethnic, religious, or similar groups who do not benefit or who are affected negatively?
3. Is it likely the same impacts could be achieved if the project were implemented in a different setting or on a larger scale?
4. It may also be useful to address the realist evaluation questions referred to earlier: Who benefits from the program? How do they benefit? When do they benefit? Why? Who does not benefit and why?

The RealWorld evaluator needs to distinguish between critical issues that must be explored in depth and less critical issues that can be studied less intensively. It is also essential to understand when the client needs rigorous statistical analysis to legitimize the evaluation findings to members of Congress, funding agencies, or those critical of the program and when more general analysis and findings would be acceptable. Answers to such questions can have a major impact on the evaluation design, budget, and time required.

3. DEVELOPING THE PROGRAM THEORY MODEL

Theory-based evaluation (TBE), also known as program theory evaluation, is "an explicit theory or model of how the program causes the intended or observed outcomes" (Rogers, Petrosino, Huebner, & Hacsi, 2000, p. 5). All programs are based on an explicit or implicit hypothesis or theory about how intended program outputs will lead to desired outcomes and impacts and the factors constraining or facilitating their achievement. TBEs are particularly useful for RWE as a framework to identify critical areas and issues on which limited evaluation resources or time should focus. A TBE can also help explain whether failure to achieve objectives is due to faulty expectations or ineffective project implementation (Lipsey, 1993; Weiss, 1997), or to contextual factors that are largely beyond the control of program managers.

There are at least four different ways in which a TBE model can be developed. The first is where the program theory is developed by program staff during project design before the evaluation begins. The evaluators will normally use this model as a starting point, although it may be updated periodically. The second is where the TBE is developed by the evaluator with only minimal input from most stakeholders. In most cases the evaluator tries to involve stakeholders in the process, but this is either not possible because of time constraints or because of lack of interest. The third is where the evaluator constructs the TBE in consultation with a few key stakeholders such as the funding agency and program management. The fourth is where there is a participatory consultation process that involves a wide range of stakeholders, including different sectors of the target population. Which of these approaches is used has important implications for how much buy-in there is to the evaluation process and how the findings are used and disseminated.

TBE models are relatively easy to describe for projects that have a relatively simple structure with a limited number of inputs intended to achieve well-defined and measurable outputs and outcomes. They also have a relatively linear structure (see discussion in Chapters 9 and 15) and often a defined start and end date. However, many projects have less clearly defined objectives and often no defined end date. Program theory can still be applied in these latter cases, but a more creative approach will often be required to identify objectives and the underlying assumptions on which the project is based.

A key factor that is frequently ignored in program theory models is the concept of *emergence.* Many program theory models implicitly assume that projects are operating in a relatively stable environment, with no major changes in the project design, the services provided, and how the project is organized. However, in practice many projects, particularly those being implemented over a period of several years, are likely to experience significant changes in all of the above. After several years the range of services being offered has changed, as well as how they are delivered. Realist evaluation (see Chapter 10) is one of the approaches that specifically addresses issues of emergence and how the response of different sectors of the target population can affect how projects evolve.

As interventions become larger and more complex, it becomes harder to apply program theory models. When applied to programs with a number of different components, it becomes necessary to use a multilevel program theory model (Bamberger, Vaessen, & Raimondo, 2016; Funnell & Rogers, 2011; Leeuw, 2016a; Rogers, 2008). The application of TBE models for complex intervention is discussed in Chapter 16.

Figure 2.2 illustrates how a program theory model can be applied to a typical project. The model describes the seven stages of the project cycle:[3]

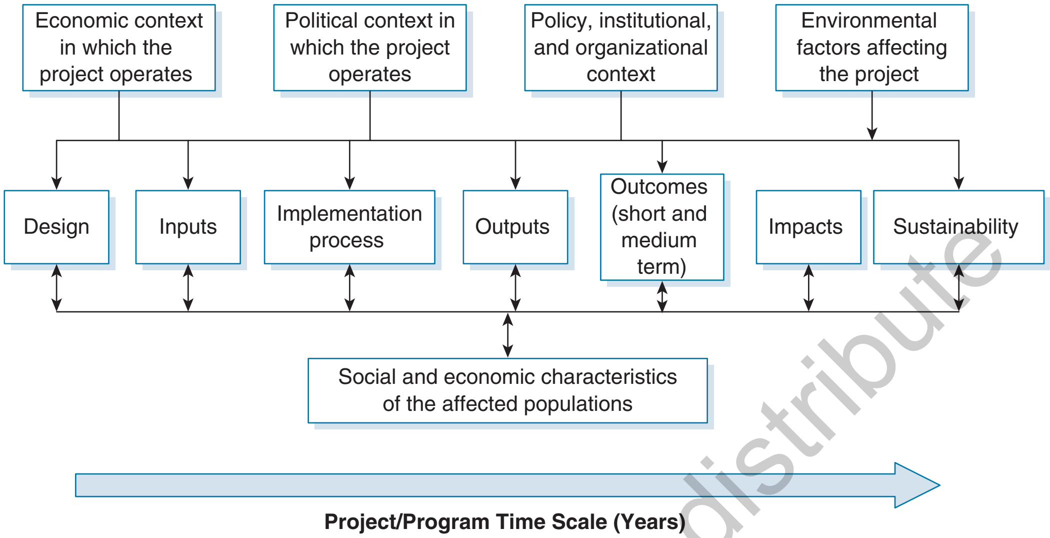
1. ***Design***. How the project was designed (e.g., was it top-down, were there participatory consultations, was a standard blueprint used, or was it adapted to the local context)?
2. ***Inputs***. The financial, human, material, technological, and information resources used in the project
3. ***Implementation process***. The actions taken or work performed through which inputs, such as funds, technical assistance, and other types of resources, are mobilized to produce specific outputs; to what extent and how intended beneficiaries were involved
4. ***Outputs***. Products and services resulting directly from program activities
5. ***Outcomes***. The intended or achieved short- and medium-term effects of an intervention's outputs. Outcomes represent changes in development conditions that occur between the completion of outputs and the achievement of impact. Higher-level outcomes, often referred to as impacts, usually require acknowledging the collective efforts (plausible contributions) of partners and other actors.
6. ***Impacts***. Long-term economic, sociocultural, institutional, environmental, technological, or other effects on identifiable populations or groups produced by a project, directly or indirectly, intended or unintended
7. ***Sustainability***. Continuation of benefits after a project has been completed

Although the first four components of this model—design, inputs, implementation process, and outputs—may be controllable by those managing the project, by contrast, the outcomes, impacts, and sustainability depend to a considerable degree on external factors over which the project agency usually has little or no control. Some forms of logic models (e.g., logframes) refer to these as *assumptions*. Since the success of a project achieving higher-level results depends on those external assumptions, it is important that they be verified. If essential external conditions change, it will be necessary for the project design to adapt to those changes. And, of course, there is the assumption that impacts will be *sustained* over the intended life of the project.

Some of the different ways in which the concept of impact is used in evaluation are described in Box 2.1. There are also a number of agencies that do not use the concept of impact, believing that it is methodologically or philosophically too difficult to define, measure, or interpret. However, we go along with the majority of development practitioners and try to evaluate impacts—while fully recognizing all the methodological and philosophical limitations on how well impacts can be measured/assessed/inferred, particularly in RWE contexts.

[3] Several of the definitions given here are adapted from Organization for Economic Cooperation and Development/Development Assistance Committee (OECD/DAC) (2002). This source is widely used by the evaluation departments of international development agencies.

FIGURE 2.2 ⬢ A Simple Program Theory Model


Economic context in which the project operates
Political context in which the project operates
Policy, institutional, and organizational context
Environmental factors affecting the project
Design
Inputs
Implementation process
Outputs
Outcomes (short and medium term)
Impacts
Sustainability
Social and economic characteristics of the affected populations
Project/Program Time Scale (Years)

An **evaluability assessment** may also be conducted during the scoping phase. This is an assessment of the feasibility of assessing project outputs, outcomes, and impacts with the available resources and data. While design, inputs, implementation processes, and outputs can be directly observed, measured, and documented by the project's monitoring system, indicators of outcomes and impacts usually require additional data collection (e.g., sample surveys or in-depth QUAL data collection), often using one of the designs discussed in Chapter 11. Whether the project makes plausible contributions to such outcomes and impacts must be tested or inferred. Consequently, one of the purposes of the evaluability assessment is to determine whether resources will permit collecting the types of additional data required to assess outcomes and impacts.

Outcomes are the short- and medium-term effects and impacts are the long-term effects of a project. In other words, these are the changes that can be wholly or partly attributed to the interventions of the project, perhaps by a counterfactual that estimates what would have been the economic, sociocultural, institutional, or other conditions of the intended beneficiaries in the absence of the project's interventions. The difference between the observed conditions of the beneficiaries and the counterfactual is the estimated impact of the project. The methodology for assessing project impacts through a variety of evaluation designs is discussed in Chapter 11.

The logic model depicted in Figure 2.2 also identifies five sets of **contextual variables** that may affect implementation and outcomes. These include the economic, political, organizational, operational, and environmental settings of the project and the socioeconomic and cultural characteristics of the affected populations (Hentschel, 1999; Patton, 2008). The following are examples of how each of these contextual variables can affect the project and how their analysis can strengthen interpretation of evaluation findings:

- *Economic factors:* In a dynamic economy in which jobs are being created and demand for products and services is growing, people are often more willing to invest time or resources in developing marketable skills or in launching small businesses. It is often hypothesized, for example, that parents are more willing to pay for their daughters to stay in school (and to forgo the daughters' assistance with domestic and farming activities) if labor market conditions create the expectation that extra education will help them get better jobs.

- *Political factors:* Support from local government agencies (who happen to be from the same political party as the national or state government sponsoring a project) can significantly improve project performance by mobilizing community support or providing free resources such as transport, workers, or buildings. Inversely, politically induced opposition to a project can seriously affect its success or even its ability to operate. Sometimes projects can become affected by political campaigns. In Zambia in the late 1970s, a donor agency was trying to convince the Ministry of Housing to charge full economic rent for low-cost housing. One of the candidates in the municipal election campaign promised families that if he was elected, all rents would be subsidized, which contributed to the reluctance of families to pay their rent to the project.

- *Organizational and institutional factors:* Many projects require support from government agencies and other organizations such as **NGOs** (nongovernmental organizations) or religious organizations. The effectiveness of this cooperation can vary considerably from one community, district, or city to another. In some cases, this is due to personalities, in other cases to local politics, but in many cases, it is mainly due to differences in staff, finances, or other resources. Sometimes something as basic as the Ministry of Health in one town having a jeep, whereas in the next location it does not, can have a major impact on the level and effectiveness of support.

- *Environmental factors:* Agricultural and rural development projects are directly affected by variations in the local environment. The new grain varieties being introduced may prosper well on flat land but not on hillsides, or they may be very sensitive to variations in seasonal rainfall. Urban development projects may be affected by erosion or flooding. All these factors may produce dramatic differences in crop yield or in the success of water and sanitation projects.

- *Socioeconomic and cultural characteristics of the target communities:* Many countries in Africa and other developing regions have literally hundreds of different tribal groups, each with its own farming practices, rules concerning use of natural resources, marriage practices, and attitudes concerning the mobility and economic participation of women. In one village in Uganda, bicycles proved an effective way to transport water and reduce women's time burden (because water was carried in square metal jerry cans that could easily be transported on the bicycle's luggage rack), but in a neighboring village, bicycles failed to produce this benefit because water was transported in round clay pots that could not easily be transported on a luggage rack.

An analysis of these contextual factors can often help explain why two identical projects may have very different outcomes in different communities. In one community, the economy may be thriving, whereas in another it is in decline—so parents are more willing to pay for their daughters' continued education in the first than in the second; in one community, most of the farmland is flat and well drained, whereas in the next community, most of the land is hilly and the new variety of grain does not prosper. For these reasons, evaluators are strongly encouraged to incorporate contextual analysis into the evaluation design.

A key element of program theory models is the identification and monitoring of critical assumptions about inputs, implementation processes, and the expected linkages with outcomes. There are two types of assumptions: internal and external. Internal assumptions, or hypotheses, describe the logical cause-and-effect links between interventions and outcomes. External assumptions refer to factors beyond the direct control of a project—for example, whether the project should address policy issues rather than take it for granted that it can have no influence over them. Even those external factors that truly cannot be changed by the project need to be monitored if the success of the project depends on the correctness of those assumptions or on needed adjustments in response to changing external conditions.

BOX 2.2

Defining Outcomes and Impacts

Outcomes are defined by OECD/DAC (2002) as "the likely or achieved short and medium term effects of an intervention's outputs." In this book, we also focus on impact evaluations. However, there is a wide variety of definitions of and assumptions related to the meaning of *impact* and to the related term *outcomes*. Impact evaluation goes beyond an examination of outputs or outcomes produced by a project's interventions to determine higher-level and longer-term effects.

The definition of impact adopted by the OECD/DAC (2002) is "Positive and negative, primary and secondary long-term effects produced by a development intervention on identifiable population groups, directly or indirectly, intended or unintended. These effects can be economic, sociocultural, institutional, environmental, technological or of other types." This definition emphasizes that an impact evaluation is conducted late in the project cycle to assess the long-term effects, but the definition does not require that a particular methodology be used.

On the other hand, many quantitative researchers define impact evaluation in terms of the methodology, stating that impacts can only be assessed through the definition of a statistical counterfactual (generated through experimental or quasi-experimental designs), but most of these definitions do not state the point in the project cycle at which an impact evaluation can be conducted.

There are other definitions or nuances, including the following:

- Some writers consider that outcomes are the observed changes in the variables the project seeks to affect, whereas impacts are the proportion of the changes that can be attributed to the project. Outcomes (changes in conditions) can be observed, whereas impacts (the influence a project had on those changes) can only be inferred through the use of an analytical process such as an experimental or a quasi-experimental design.
- Some define impact as "higher-level" outcomes (CARE International's definition is "equitable and durable improvements in human wellbeing and social justice"; CARE International, 2003) whether or not changes at these levels can be directly attributable to a project. A project can be held accountable for direct attribution to outputs and more immediate outcomes/short-term effects, and thus to their plausible contributions to higher-level sustainable impact, along with other influences that must be identified and acknowledged.
- The dictionary definition of impact refers to influence, the effect or impression of one thing on another—for example, what difference did the project make?
- But influence on what? A project could have an "impact" on staff paychecks, on direct (but superficial) benefits of services provided to participants, on the capacities of local organizations, on the conditions of the target population, on the empowerment of individuals and community groups, on national policy, on the achievement of the Millennium Development Goals[4] . . . the list could go on. There needs to be agreement among stakeholders (including intended beneficiaries, donors, and partners) on what "success" (and therefore impact) would look like. It depends on their values and expectations—and on what is reasonable to expect from a time- and resource-bound project.
- There are those who refer to impact in terms of scope or scale—for example, how many people's lives were influenced (impacted) in some way?
- Others refer to it in terms of degree or depth—that is, whether a project had a minor influence or made a significant difference in the quality of life of beneficiaries in important ways.
- There are also the intended/unintended dimensions of impact, desired/negative impact, and direct/indirect impact (e.g., multiplier effect of people adopting a practice beyond those who participated directly in a project).

[4] See www.un.org/millenniumgoals.

3.1 Theory-Based Evaluation (TBE) as a Management Tool

TBE is widely used as a management tool to define and monitor progress toward the achievement of program objectives (results, goals, outcomes, and impacts), to test the validity of the assumptions on which program design is based, and to draw lessons for the design and implementation of future projects. In order to do this, the program theory is represented graphically by a logic model that is then translated into a set of tables that operationalize inputs, outputs, outcomes, and impacts with related indicators that can be measured and against which progress can be tracked. Today, the most common framework for doing this is results-based management (RBM), which is a refinement of logical framework analysis. RBM is discussed in Chapter 10.

An important management tool is the incorporation of a results chain (see Chapter 10) that spells out in more detail the steps through which the outputs, outcomes, and impacts are to be achieved. When these are not fully achieved, the results chain enables management to identify the link in the chain at which the problems arose. The results chain also helps identify the critical hypotheses and assumptions that need to be tested by the evaluation.

Leeuw (2016b) identifies five sets of management and policy problems to which TBE can contribute:

- Problem 1: What can TBE contribute to define and operationalize key performance indicators for program, policy, and other interventions?
- Problem 2: What can TBE contribute to defining the counterfactual when it is not possible to use statistical (experimental and quasi-experimental) evaluation designs?
- Problem 3: What can TBE contribute at the program design stage in assessing how effective the proposed programs, policies, or other interventions can be?
- Problem 4: What can TBE contribute to find out—during implementation—how plausible the effectiveness of a program, policy, or intervention probably will be?
- Problem 5: What can TBE contribute when the findings of an impact evaluation are not clearly explained?

While many evaluators assume that TBE will mainly be used at the end of a project to assess outcomes and impacts and the overall effectiveness of the project design, Leeuw shows how it can be used at all stages of a program.

4. IDENTIFYING THE CONSTRAINTS TO BE ADDRESSED BY RWE AND DETERMINING THE APPROPRIATE EVALUATION DESIGN

The final part of Step 1 of the RWE approach includes preliminary identification of those budget, time, data, and political and organizational constraints that can be anticipated. This can lead to a determination of which of the options for Steps 2, 3, 4, and 5 will need to be used. Once the evaluators have identified what they consider to be the best options for addressing these constraints, the proposed strategy will then be discussed with the client and, ideally, key stakeholders. Often this may involve a period, sometimes quite long, of negotiation and revision of the proposed strategy. In some cases, the evaluation team may try to convince the client that the requested reductions in budget or time are not possible without prejudicing the purposes for which the evaluation is being conducted. In these cases, it is extremely important for the client to understand the types of information, findings, and recommendations that can and cannot be provided within these constraints and the levels of precision, validity, and adequacy that the evaluation can be expected to achieve. Importantly, the client should also understand how these compromises are likely to affect the credibility of the findings with different stakeholders.

Chapters 3 through 6 review the options for addressing budget, time, data, and political constraints, respectively.

5. DEVELOPING DESIGNS SUITABLE FOR REALWORLD EVALUATION CONDITIONS

We now address a very important decision that needs to be made by those commissioning and conducting an evaluation: What evaluation design would be most appropriate for responding to the priority questions determined during the assessment of client needs, and which design options are even possible, given the constraints and the stage the project has reached? As we saw in Table 1.1, the earlier in the life of the project this decision is made, the more design options are available. Though the subject of evaluation designs will be covered in more detail in Chapter 11, we provide a brief introduction here.

It is helpful to break the choice of the appropriate design for a particular evaluation into three sequential decisions. These decisions are based on the identification of the key questions that the evaluation must address (see Chapter 1) and the real-world constraints under which the evaluation must be conducted (see earlier in this chapter).

- Decision 1: Which of the main kinds of evaluation design (more than one design may be combined) is/are most appropriate (see Box 2.2).
- Decision 2: Should the evaluation use a quantitative (QUANT), qualitative (QUAL), or mixed-method design (see Chapter 4)?
- Decision 3: If it is decided in Decision 1 to use an experimental or quasi-experimental design, which design (or designs) should be used? The main options are listed in Box 2.3 (see below).

BOX 2.3

The Main Kinds of Evaluation Design

The following are the most widely used evaluation designs; in most cases the evaluator will select one, or sometimes more than one, of these designs.

1. Experimental or quasi-experimental designs
2. Theory-based evaluations
3. Case-based designs, including Qualitative Comparative Analysis designs (QCA)
4. Qualitative designs
5. Systematic reviews
6. Statistical designs
7. Complexity responsive designs
8. Big data analytic designs

Chapter 11 identifies and describes a framework for classifying the evaluation design structures (see Table 11.3—repeated in Appendix 2.2 for easy reference). They are classified in terms of when the evaluation began, whether baseline (or only posttest) data were collected, whether a **comparison group** design was used, and if so, how the comparison group was selected. These structures or **scenarios** should only be considered as the skeleton of the evaluation design, and a wide range of different methodologies (quantitative, qualitative, and mixed-method) can be used within each structure. Given the ongoing debates concerning whether there is a "best" evaluation design, it is important to stress that we do not believe there is a single best design and that the choice of the appropriate design must be made on the basis of the purpose of the evaluation and a review of types of evaluation presented earlier.

It is useful to distinguish between designs that are statistically strong (experimental and quasi-experimental) and methodologically weaker (from a statistical perspective) designs. However, Box 2.2 is an important reminder that many of the "strong designs" referred to in the research literature are, in fact, only strong in their ability to address statistical threats to selection bias, but they are potentially weak with respect to other important methodological areas. These questions are discussed in more detail in Chapter 11 and in Chapters 12, 13, and 14, in which the relative strengths and weaknesses of quantitative, qualitative, and mixed-method designs are compared.

The evaluation design frameworks presented in Appendix 2.2 can be classified into three categories depending on whether there is a control/comparison group and, if so, how it is selected:

- *Experimental designs* (randomized control trials) in which subjects are randomly assigned to the project and treatment groups. This is the strongest statistical design in terms of control for selection bias but, as noted in Appendix 2.3, these designs have a number of potential methodological weaknesses—some of which can be addressed through incorporating a mixed method design.
- *Quasi-experimental designs* in which a comparison group is used, but it is selected separately from the project group so that there are potential problems of selection bias. When large samples are used or good secondary data are available, it is possible to use statistical matching procedures such as propensity score matching. When this is not possible, judgmental matching procedures, which are statistically weaker, must be used.
- *Nonexperimental designs* that do not use a statistical comparison group. While these designs are often used when an evaluation must be conducted under time and budget constraints and where as a consequence the methodology is often weak, many other evaluations use methodologically strong qualitative and mixed-method designs (see Chapters 13 and 14).

So this framework should be considered as a starting point for identifying the most appropriate evaluation design and not as a definitive list of evaluation designs and certainly not as a ranking of "strong" and "weak" designs.

5.1 How the Availability of Data Affects the Choice of Evaluation Design

As indicated earlier, evaluations can be based mainly on primary data collection, they can use a combination of primary and secondary data, or they can be based mainly on secondary data from surveys (or other sources such a big data) that have already been conducted or generated.

TABLE 2.1 ⬢ Evaluation Design Components to Strengthen All of the Basic Evaluation Designs

Essential Evaluation Design Component	Why Required	How to Implement
1. Basing the evaluation on a program theory model	The purpose of an evaluation is not just to estimate *how much* change has occurred but also to explain *why* and *how* the changes were produced. Clients also wish to know to what extent the changes were due to the intervention and whether similar changes would be likely to occur if the program is replicated in other contexts. In order to achieve the above objectives, it is necessary to explain the underlying theory and the key assumptions on which the program is based and to identify how these can be tested in the evaluation.	The design and use of program theory are discussed in Chapter 10. That chapter also illustrates how the theory can be articulated graphically through a logic model.
2. Process analysis	Project outcomes are affected by how well a project is implemented and by what happens during implementation. Without process analysis, it is not possible to assess whether failure to achieve outcomes is due to design failure or to implementation failure.	See Chapter 10.
3. Multiple data-collection methods	Many evaluations use a single method of data collection. For QUANT designs, typically, data are collected using a structured questionnaire. This is not adequate for collecting information on sensitive topics or on multidimensional indicators.	See Chapters 13 and 14.
4. Contextual analysis	Projects implemented in an identical way in different locations will often have different outcomes due to different local economic, political, or organizational contexts or different socioeconomic characteristics of target communities. This can result in wrong estimations of project impact, often leading to underestimation of impacts (due to increased variance of the estimations).	See Chapter 10.
5. Identification and use of available secondary data	Many evaluations do not identify and use all of the available secondary data. Secondary data can often reduce the costs of data collection and provide independent estimates of key variables.	See Chapter 5.
6. Triangulation	The validity of data and the quality and depth of interpretation of findings are enhanced when two or more independent estimates can be compared.	See Chapters 13 and 14.

TABLE 2.2 ● Determining Possible Evaluation Designs[a]

Question[b]	If the Answer Is Yes	If the Answer Is No
1. Was the evaluation preplanned? That is, was the evaluation design included in the project's monitoring and evaluation plan from the beginning?	Use that preexisting plan as the guide for the project evaluation. The evaluation should include an assessment of the appropriateness of the monitoring and evaluation plan and should acknowledge and use it as much as possible.	This is going to have to be an ad hoc, one-off evaluation (e.g., Design 5 or 7). This limits the rigor of the evaluation design, but there are things that can be done, even so.
2. Was there a baseline (pretest)?	That will make a before–after comparison (Design 1, 2, 4, or 6) possible—if the baseline was done in a way that can be compared with the posttest (end-of-project evaluation).	Too bad. You'll either have to make do with retrospective analysis, a with–without (comparison group at endline only, Design 5) or cope with a "one snapshot" limitation.
3. Was there a comparison group for the baseline?	Recommend Design 1 or 2 if there can be the same or a comparable control group for the posttest (see next question).	Too bad. You could still use Design 3 or 4, hoping that the posttest comparison group was similar to the participants at the beginning of the project.
4. Even if there was no comparison group in the baseline, can there be a comparison group for the posttest (end-of-project evaluation)?	Design 3, 4, or 5 could be used. Do all possible to verify that the comparison group was similar to the participants at the beginning in all ways except for the intervention.	Consider looking for secondary data that may give general trends in the population to compare with the group that participated in the project.
5. Was reliable monitoring information collected on outcome and/or impact indicators during project implementation?	Very helpful! Quasi-experimental longitudinal Design 1 may be possible, including examining trends over time.	Well, pretest + posttest with comparison group (Design 2) isn't bad. You might still look for secondary data indicating trends.
6. Will it be possible to conduct an ex-post evaluation sometime (e.g., several years) after the end of the project?	An extended longitudinal Design 1 will provide more certain evidence of sustainability (or lack thereof).	Without an ex-post evaluation, predictions about sustainability will have to be made based on the quality of the project's process and intermediary outcomes.

[a] These are the kinds of questions that should be asked by an evaluation team when called in to evaluate an ongoing project. Obviously, if these questions are considered at the same time as a project is designed, the evaluation plan can be stronger. Otherwise, the evaluation team will have to cope as well as it can with the given situation. See Table 2.1 for the designs referenced here.

[b] Readers not familiar with any of the terms used in this table are referred to Chapter 11, where all the evaluation designs are discussed.

Over the past few years the new category of *big data* is becoming of increasing importance for development programs and is gradually being introduced into development evaluation. Big data has many quite distinct characteristics compared to conventional sources of evaluation data and often requires the use of new analytical tools (see Chapter 18). While big data is often discussed as a completely unique kind of data (many books are published exclusively on the collection and

[5] Recall techniques (see Chapter 5) involve asking individuals or groups to give their recollections of their personal situation or the situation of their community at an earlier point in time. For impact evaluations, the earlier time will usually be the time at which the project was starting.

analysis of big data), it is important for evaluators to recognize that there is a *data continuum* going from big data at one extreme, through *large data* (typically sample survey data and administrative data), through *small data* (from case studies and qualitative research). Many evaluations will draw on and combine data from different points on the data continuum. In practice, triangulation and mixed methods are powerful tools to combine the strengths and address the weaknesses of each type of data.

When good-quality secondary data are available, the range of design options is increased: pretest/posttest comparison designs can be used even when the evaluation does not start until late in the program cycle. It becomes more feasible to use a comparison group, and the procedures for selecting a comparison group can be strengthened.

Whichever design is selected, the RWE approach strongly recommends that the basic statistical design be complemented by a number of mixed-method design components to address some of the weaknesses of the statistical designs. Table 2.2 describes six essential components of all RWE evaluation designs.

While these frameworks and designs have often been discussed within the context of QUANT evaluation designs, the scenarios (e.g., whether or not the evaluation begins with a baseline assessment, whether there are to be before-and-after comparisons, whether or not there is some form of counterfactual analysis) apply equally to QUAL and mixed-method evaluations. Most QUANT evaluations are based on either an experimental or a quasi-experimental design in that they seek to directly measure changes in a set of QUANT variables and to assess whether the changes are associated with the project interventions. Even if using QUAL methods to determine people's perspectives of changes that have taken place, there will be discussion of what things were like before the intervention started, what changes might be attributed to the project, and how these might compare with what has happened in other communities. These considerations make it possible to identify a set of evaluation designs that cover most RWE scenarios and that include QUANT, QUAL, and mixed-method approaches. Table 2.2 provides a decision tree matrix to help decide which design is possible under different scenarios.

To repeat the caveat mentioned previously, it should be noted that whereas the various quasi-experimental evaluation designs given in Appendix 2.2 are more commonly associated with QUANT methods, whether there can be a before–after and/or a with–without comparison applies to both QUANT and QUAL methodologies. The major differences between these designs have to do with the stage of the project at which the evaluation team collects data (e.g., baseline, midterm, final, ex-post). A separate distinction has to do with whether the data-collection *methods* are QUANT, QUAL, or mixed and whether they also rely on secondary sources or the **recall**[5] perspectives of key informants and participants.

In the situation depicted by Design 7, in Appendix 2.2 (posttest analysis without baseline or comparison group), for example, the evaluators would not only want to measure (QUANT approach) or describe (QUAL approach) the present status of the condition the project aimed to change (indicator or other form of evidence); they would also need to find some evidence of how that condition changed over the life of the project and a comparison of how that change may have been different for those participating in the project compared with others under similar conditions who did not. This calls for finding secondary data or collecting the perspectives of knowledgeable people and the use of recall. Whether the evaluator does that by measurement (collecting numbers) or descriptions (words) has to do with methodology. How much of that data is obtained from primary sources (e.g., surveys, observation, key informants) or from secondary sources has to do with evaluation design.

5.2 Developing the Terms of Reference (Statement of Work) for the Evaluation

Those commissioning evaluations may find the following set of questions helpful when preparing the terms of reference (ToR) or scope of work (SoW) for the evaluation. (This topic is covered with more detail in Chapter 19.) The evaluators might also find this checklist helpful, particularly for identifying points not covered in the ToR and that must be clarified with the client before the evaluation is designed.

1. Who asked for the evaluation? Who are the key stakeholders? Do they have preconceived ideas regarding the purpose for the evaluation and expected findings (political considerations)?
2. Who should be involved in planning the evaluation?
3. Who should be involved in implementing the evaluation?
4. What are the key questions to be answered?
5. Will this be a **developmental** or **formative** or **summative evaluation**? Is its purpose primarily for learning and improving, accountability, or a combination of both?
6. Will there be a next phase, or will other projects be designed based on the findings of this evaluation?
7. What decisions will be made in response to the findings of this evaluation? By whom?
8. What is the appropriate level of rigor needed to collect and analyze the information needed to inform those decisions?
9. What is the scope/scale of the evaluation/evaluand (program or intervention being evaluated)?
10. How much time will be needed/available?
11. What financial resources are needed/available?
12. What evaluation design would be required/is possible under the circumstances?
13. Should the evaluation rely mainly on QUANT methods, QUAL methods, or a combination of the two?
14. Should participatory methods be used? If so, who should be included? What roles should they play?
15. Can/should there be a survey of individuals, households, or other entities?
16. Who should be interviewed?
17. What sample design and size are required/feasible?
18. What form of analysis will best answer the key questions (see the fourth question above)?
19. Who are the audiences for the report(s)?
20. How will the findings be communicated to each audience?

Summary

- Clients and other stakeholders can have widely varying expectations of what an impact evaluation is and what it can produce. These can range from detailed statistical measurements to case studies on how a program has affected the lives of individual communities, families, or schools.
- An evaluation should be based on a sound understanding of why the evaluation is being commissioned, how the findings will be used, and the political context within which it will be conducted. Understanding the client's *bottom line*—what information and analysis are essential and what would simply be "nice to have"—is critical when decisions have to be made on what can and cannot be cut in the light of budget and time constraints.
- All programs are based on an explicit or implicit model of how the program is expected to operate, how the intended program outputs and impacts are to be achieved, and the factors facilitating or constraining achievement. Defining the program theory helps focus the evaluation and identify the key hypotheses and linkages that the evaluation must test.
- The scoping step should end with an agreement between the client and the evaluator on the RWE design that best responds to the purposes for which the evaluation is being commissioned while at the same time adapting to the budget, time, data, and political constraints under which it must be conducted.

Further Reading

Altschuld, J., & Kumar, D. (2010). *Needs assessment: An overview*. Thousand Oaks, CA: Sage.

This is the introduction to a five-volume needs assessment kit. It explains the importance and applications of needs assessment in different organizational contexts and provides a three-step generic needs assessment model that can be applied in many different organizational contexts.

Bamberger, M. (2016). *Integrating big data into the monitoring and evaluation of development programs.* New York: UN Global Pulse with support from the Rockefeller Foundation. Available at http://www.unglobalpulse.org/big-data-monitoring-and-evaluation-report

A review of the main types of big data and how these are, or potentially could be, used in development evaluation.

Carvalho, S., & White, H. (2004). Theory based evaluation: The case of social funds. *American Journal of Evaluation, 25*(2), 141–160.

An example of the application of program theory to the evaluation of social investment funds (a widely used model for providing health, education, water supply, and other local infrastructure in developing countries). The article illustrates how program theory can be reconstructed during the evaluation when it was not defined in the project documents. The article is also interesting because it presents the concept of an "anti-theory" based on the views of critics as to the potential negative outcomes of the project interventions.

Chambers, R. (1997). *Whose reality counts? Putting the first last.* London, UK: ITDG.

Making the case for ensuring that all stakeholders and affected groups, particularly the poorest and

(Continued)

(Continued)

most vulnerable, are involved in program design, implementation, and evaluation.

Donaldson, S. (2007). *Program theory-driven evaluation science: Strategies and applications.* Thousand Oaks, CA: Sage.

Recent overview of program theory. Includes eight case studies illustrating diverse applications of program theory.

MercyCorps. (2005). *Design, monitoring and evaluation guidebook*. Portland, OR: Author.

Includes chapters on project design and criteria for useful evaluations.

Morra-Imas, L. G., & Rist, R. C. (2009). *The road to results: Designing and conducting effective development evaluations.* Washington, DC: World Bank.

Very practical guidance for the evaluation of international development programs. Includes understanding the context, developing the program theory of change, considering the evaluation approach and design, developing evaluation questions, and much more.

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

Probably the most widely cited text on how to ensure that evaluations respond to the needs of stakeholders and that the findings will be used.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London: Sage.

They argue convincingly for the need for a better understanding of the intervention, what happens during implementation, who benefits, and why.

Rogers, P. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation, 14*(1), 29–48.

Introduction to the concept of complexity and the challenges of applying logic models to complex programs. Also includes a review of recent developments in logic modeling that help address the challenges of modeling complex programs.

Rogers, P., Hacsi, T., Petrosino, A., & Huebner, T. (Eds.). (2000). *Program theory in evaluation: Challenges and opportunities.* New Directions for Evaluation No. 87. San Francisco, CA: Jossey-Bass.

A useful overview of program theory and still relevant even though published in 2000. All the chapters include extensive reference sources.

Rossi, P., Lipsey, M. W., & Freeman, H. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

This classic textbook has a number of chapters covering the topics discussed in this chapter, including tailoring evaluations, identifying issues and formulating questions, assessing the needs for a program, and expressing and assessing program theory.

Thomas, A., & Mohan, G. (2007). *Research skills for policy and development: How to find out fast*. London: Sage/The Open University.

This publication explains how research is designed and used to contribute to policy and action. It focuses on the kinds of information that policymakers need and shows the wide variety of sources and methods that can be used to generate this information. Chapter 1, "Information Needs and Policy Change," illustrates the kinds of information required for different kinds of policy decisions and makes the important point that as the policy context changes, so do the information needs. Researchers and evaluators must be sufficiently attuned to the policy environment to be able to adapt the focus of their evaluation.

Weiss, C. H. (2001). Theory-based evaluation: Theories of change for poverty based programs. In O. Feinstein & R. Picciotto (Eds.), *Evaluation and poverty reduction* (pp. 103–114). New Brunswick, NJ: Transaction.

A discussion of how program theory models can be applied to the evaluation of poverty reduction programs.

White, H., & Bamberger, M. (2008). *Impact evaluation in official development agencies.* Sussex, UK: IDS Bulletin.

Considerations for the determination of evaluation designs, based on extensive experience with international agencies.