



# OVERVIEW

## RealWorld Evaluation and the Contexts in Which It Is Used

The chapter begins with an overview of the RealWorld Evaluation (RWE) approach, the contexts in which RWEs are conducted, and the many different constraints, pressures, and influences under which evaluations are formulated, conducted, disseminated, and used. The **RWE** approach was originally developed to address four of the most common constraints evaluators face: budget, time, and data constraints, and political influences. Subsequently, participants in workshops highlighted the importance of issues concerning organizational structures, and management and administrative arrangements. Two common RWE scenarios are reviewed. The first is when the evaluator is brought in at the start of the **program**<sup>1</sup> but with constraints on the types of information that can be collected or the designs that can be used. The second concerns retrospective evaluations where the evaluator is not called in until the program is nearing completion. For most retrospective evaluations, no baseline data have been collected and no comparison (control) group has been identified. The widespread use of retrospective evaluations is due to the fact that this is an approach used by the interdependent evaluation offices that form part of most multilateral development banks and many UN and bilateral development agencies.

### 1. WELCOME TO REALWORLD EVALUATION

---

Most evaluators are familiar with situations in which programs are nearing completion before clients begin to think seriously about evaluating whether the programs are achieving their objectives and producing the intended **impacts**. Usually, the belated interest in evaluation is motivated by the need for solid **evidence**<sup>2</sup> on which to base decisions about whether the program should be continued or perhaps expanded. When the evaluations do finally get underway, many have to be conducted under budget and time constraints, often with limited access to baseline data and comparison groups. Consequently, it is difficult to apply the most rigorous evaluation designs.

---

<sup>1</sup>Unless there is a need to be more specific, we use the term “program” to refer to any kind of intervention at the project, program, sector, national, or international level.

<sup>2</sup>Bold technical terms are defined in the Glossary at the end of this book.

Although more resources are often allocated to evaluation in developed countries, many evaluators in the United States, Canada, Europe, Japan, and Australasia report that they operate under similar constraints to those faced by their colleagues in developing countries.<sup>3</sup> As if these problems were not enough, many evaluations are often conducted in *political* environments in which funding agencies, clients, and key stakeholders have strongly held views on what the “right” evaluation methods should be, what types and amounts of information should be collected, and which groups should and should not be asked to comment on (or even see) the **findings**. New evaluators soon discover that seemingly straightforward “technical” issues—such as whether to use randomized selection of program and control groups; the choice of qualitative, quantitative, or mixed-method designs; and whom to interview and what questions to ask—can provoke strong reactions from clients and stakeholders.

Despite the difficult circumstances under which many evaluations have to be conducted, there is a growing demand from funding agencies, governments, and civil society for systematic impact evaluations, including whether the program met its objectives and should be continued or expanded to other communities or locations. Consequently, there is a strong demand from many sides for evaluators to answer questions such as those proposed by Stern et al. (2012):

- To what extent can a specific net impact be attributed to the intervention?
- Did the intervention make a difference?
- How has the intervention made a difference?
- Will the intervention work elsewhere?

Many evaluators also address the more detailed questions proposed by Realist Evaluation (Pawson, 2013; Pawson & Tilley, 1997) concerning: who benefits? where? when? how? and why? There is also an increasing awareness that, in order to be considered credible, evaluation conclusions need to be supported by sound evidence and not just opinions—although there are often major disagreements as to what constitutes credible evidence.<sup>4</sup>

The RWE approach presented in this book was developed in response to the demand for guidance on how to conduct sound evaluations when faced by these kinds of constraints, accommodating organizational structures and administrative procedures, while at the same time ensuring maximum possible methodological rigor within the particular evaluation context.

RWE is based on the following seven-step approach, summarized more specifically in Figure 1.1 and described in detail in Chapters 2 through 8:

- Step 1: *Planning and scoping the evaluation*. Before selecting the evaluation design, it is important to fully understand the purpose of the evaluation, the information needs and expectations of the clients and stakeholders, and the constraints and pressures under which

<sup>3</sup>One of our colleagues who has worked with major U.S. foundations that support community-level initiatives stated that there is a huge unmet need in the United States for material on how to conduct evaluations when working with very limited financial and professional resources. He stated that his “and other foundations make lots of small grants. There is often not enough money in the grants to hire an external consultant. And the recipients of these small grants don’t have the capacity to do internal evaluation. The evaluation work done by these nonprofits is usually pretty bad. I don’t really know of any materials targeted to this group.”

<sup>4</sup>For two publications on the question of credible evidence, see Donaldson, S., Christie, C., & Mark, M. (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage; and Rieper, O., Leeuw, F., & Ling, T. (2010). *The evidence book: Concepts, generation and use of evidence*. New Brunswick, NJ: Transaction.

they are working. What is the client's bottom line? What do different stakeholders really want from the evaluation, and how will the results be used? Difficult choices have to be made to accommodate budget and time constraints or to recognize the limitations of the available data. This step also includes getting agreement on the articulation of the **program theory model/ logic model** (see Chapter 10), which in addition to clarifying the underlying model on which the program is based, also helps identify the critical hypothesis and linkages in the program implementation model on which the limited evaluation resources should focus. It is also essential to identify and spell out key assumptions on which the program design is based. Often program management and program designers have deeply held beliefs about how and why the program will work, but often these are not spelled out in program documents.

Evaluators have come to recognize the importance of working with stakeholders to define the *boundaries* of the program and consequently of the evaluation. Is the goal of the program to produce benefits for a clearly defined group of program beneficiaries, or is it intended to also affect a broader population (other families in the program communities, neighboring communities, or wider population groups)? Is it only intended to achieve a limited number of clearly defined outcomes, or is it hoped that the program will contribute to broader outcomes, perhaps over a longer period of time? How these boundaries are defined will have an important effect on program design and potential impacts and also on how the evaluation is designed. The clarification of boundaries will have an important effect on the evaluation. The situation will often arise where the evaluators believe that the program could have broader outcomes (both positive and possibly negative) than those defined in the program results framework. It is essential for evaluators to reach a clear understanding with program management on whether it is possible to assess outcomes that are broader than those in the results framework and the program design. Often management only want evaluators to assess the defined program objectives, so it is essential for the scope to be clarified at the start of the evaluation. As we will see in Chapter 17, the issue of boundaries is often quite sensitive in gender-focused evaluations.

The scoping phase also involves identifying and assessing evaluation design options that are feasible within the cost, time, and data constraints that a particular evaluation will face, followed by assessing the strengths and weaknesses (i.e., threats to validity and adequacy) of each option. The different design options are then discussed with clients, emphasizing the trade-offs involved in each option, and an agreement is then reached on which design would be most feasible and acceptable to the client. (We get into more detailed coverage of evaluation designs in Chapter 11.) While the debate on whether there is a "best" evaluation design continues, the steady increase in the use of mixed and multiple methods designs has resulted in a recognition of the benefits of combining quantitative and qualitative designs (discussed in Chapter 14). It is also important to recognize that there are at least six widely used evaluation designs, each of which has strengths and weaknesses for addressing different kinds of evaluation questions. We argue that there is no single "best" evaluation design, and that the choice of design will be determined by the questions being addressed, the real-world constraints, the understanding of contextual factors, and the methodological preferences of the different stakeholders.

A final issue concerns whether the program should be considered "complex" and if so, whether a complexity-responsive evaluation design is required. The evaluation of complex programs will usually require the use of more expensive and methodologically rigorous evaluation designs (see Chapter 16).

Finally, the scoping phase must also understand the broader political, economic, socio-cultural, historical, and environmental context within which the program will be implemented. Strategies for contextual analysis are discussed in Chapter 16.

- *Step 2: Strategies for addressing budget and other resource constraints.* How many evaluators have been told by the client, “We really need a rigorous and professional evaluation as it is important to assess impacts, but . . . unfortunately our budget has been cut.” Step 2 describes options for reducing costs. These include simplifying the evaluation design, reducing the amount of data to be collected, making greater use of secondary data, revising the sample design and sample size, and streamlining data collection and analysis. In addition to budget, it is also important to assess organizational constraints, technical and other human resource constraints, and increasingly constraints on access to new information technology capacity.

The rapid emergence of new information technology<sup>5</sup> over the past few years means that one option for reducing costs (and time) may be to incorporate smartphones and big data into the collection and analysis of evaluation data (see Chapter 18).

- *Step 3: Strategies for addressing time constraints.* In addition to many of the approaches used in Step 2, strategies include planning ahead to avoid delays and bottlenecks, particularly during the short periods when outside consultants are in the field; building impact-related indicators into routine program monitoring data collection; and using videoconferencing to reduce travel and to permit more frequent interactions between the evaluation team and agency staff. ICT can also reduce the amount of time required for data collection and analysis.

- *Step 4: Strategies for addressing data constraints.* These include addressing problems concerning the lack of important data or data quality when the evaluation is not commissioned until late in the program cycle. Thus, one needs to consider a number of approaches for *reconstructing* baseline data. These include using secondary data sources, recall, key informants, focus groups, construct mapping, and participatory group techniques such as PRA (participatory rural appraisal). Chapter 4 covers techniques for collecting information on sensitive topics and on difficult-to-reach groups. As these groups are often the poorest and most vulnerable, there will often be pressures to ignore sensitive questions and inaccessible groups.

### BOX 1.1

#### The Implications of New Information Technology for RealWorld Evaluation

New information technology (NIT) comprises new technologies for the collection, analysis, and dissemination of data. It combines big data, Information and Communication Technology (ICT), and the Internet of Things (IOT). Big data sources include, but are not limited to, satellite images and remote sensors, digital

financial transactions such as ATMs, telephone call records and purchases of airtime, call-in radio programs, social media such as Twitter and Facebook, administrative records, and survey databases from government agencies, donor agencies, and nongovernmental organizations (NGOs). ICT data comes from

*(Continued)*

<sup>5</sup>New information technology (NIT) covers handheld devices such as smartphones and tablets and wearable and remote sensors (all of these are usually referred to as Information and Communication Technology [ICT]) and big data, such as satellite images; data streams from Twitter, Facebook, and other social media; phone records and electronic financial transfers; and audio and video recordings. In addition to the huge volume and speed with which the data is collected, and the need for advanced computing facilities for its analysis, most big data shares the characteristic that it was generated remotely for a purpose other than the evaluation or research for which it can be used. New information technology is discussed in Chapter 18.

(Continued)

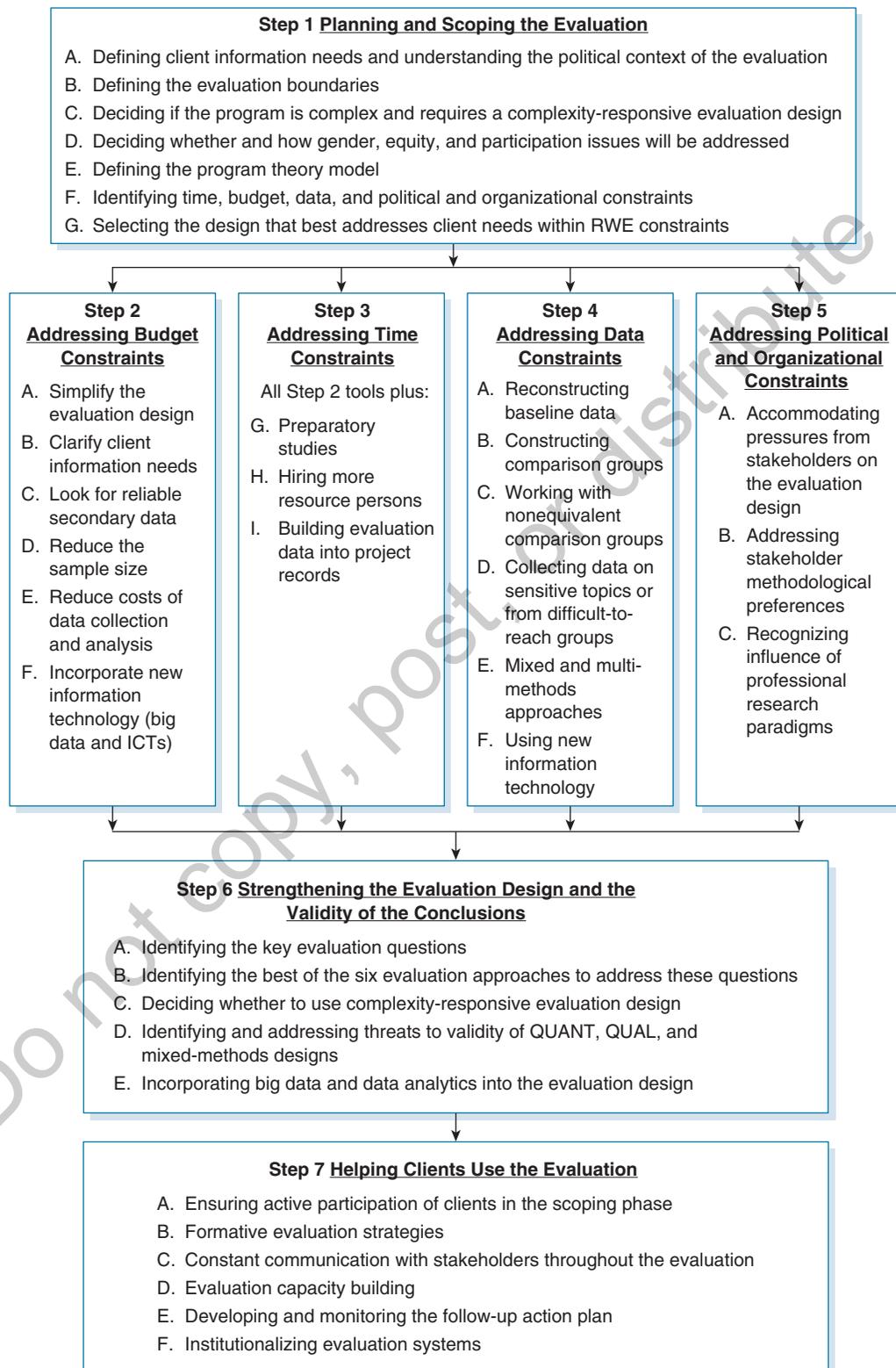
mobile phones and other handheld devices, and IOTs include devices for recording health, travel, and sensors attached to devices such as refrigerators or domestic and public electrical supply systems. However, these huge data sources are of limited practical utility without the use of smart data analytics that transform big data into user-friendly applications such as interactive maps and charts for data visualization, the identification of patterns and associations, and prediction. As we will emphasize throughout the book, NIT has major implications for development evaluation as it makes possible (i) the economical collection of much greater amounts of data, (ii) the collection of many new kinds of information that were not previously possible, (iii) the use of new forms of data analytics and predictive modeling, and (iv) the visualization of large and complex data sets in an easy to access and understand format. Furthermore, the data often can be collected, analyzed, and disseminated in almost real time.

However, as we will continually emphasize, NIT also brings challenges and potential risks for both evaluators and policymakers. For example, most big data was collected for a different purpose, and it is often analyzed through proprietary algorithms; therefore, it is often difficult to surmise the quality or meaning of the data or to detect potential biases (e.g., ethical or political) in how it is used and the social consequences of these uses. There are also organizational and institutional challenges, as coordination between data analysts and evaluators is often quite weak, with the result that much of the analysis and interpretation is done by data scientists who are often not familiar with conventional evaluation approaches, and who use analytical methods with which most evaluators are still not very familiar.

Big data offers tremendous opportunities for the future directions of evaluation, but it also presents many challenges.

Since the publication of the second edition, there have been rapid advances in the fields of new information technology, making it possible to collect and analyze data much faster and more economically (see Box 1.1 and Chapter 18). There are also issues concerning whether the new methods of collecting and analyzing data are promoting more participatory and *inclusive* evaluation approaches, giving greater voice to poor and vulnerable groups, or whether they may lead to more *extractive* approaches, whereby information is collected from and about vulnerable groups, often without their knowledge, and where decisions are made about development programs and policies without any consultation with the affected groups.

- Step 5: *Understanding and coping with political and organizational factors influencing how the evaluation is designed, implemented, disseminated, or used.* It is important to identify the key actors and their political perspectives and how these affect their orientation to the evaluation. We identify political issues arising at the outset of an evaluation, during implementation, reporting, and use of the evaluation, and we propose strategies for addressing all these issues. There are also important professional and ethical issues concerning who should be given information on the evaluation and when. Often, the client would like to limit who sees and comments on the evaluation draft, whereas the evaluator may feel that the report should be given to the mass media and to the different stakeholder groups potentially affected by the program. We will return to these ethical issues throughout the book. The design, implementation, and use of evaluations are also affected by institutional and organizational constraints, often referred to as *the political economy of evaluation*. As programs become larger and more complex, these factors can play an increasingly important role in how effectively an evaluation can be conducted. Issues of coordination or competition among multiple agencies and departments can constrain the kinds of information that can be collected and how evaluation findings are disseminated and used. The ongoing discussions on how to evaluate the Sustainable Development Goals (SDGs) illustrates the challenges of conducting multicomponent, multisectoral evaluations in over 100 countries with multiple stakeholders at the international, regional, national, and local levels.

**FIGURE 1.1** • The RealWorld Evaluation Approach

- *Step 6: Strengthening the evaluation design and the validity of the conclusions.* We argue there is no single “best” evaluation design, and the choice of design should be determined by the questions of concern to stakeholders, the real-world context within which the evaluation will be implemented, and the findings used. We identify six sets of widely used evaluation designs, plus three additional designs that are applicable to complexity-responsive evaluations (see Chapter 16). Each of these designs has areas of application and its respective strengths and weaknesses. Given the widespread popularity of experimental designs (the most common of which is the randomized control trial), we discuss the great importance of the experimental approach to evaluation, as well as the many practical and methodological limitations of RCTs (randomized control trials) in real-world contexts. We also discuss the need to understand *threats to validity*, how these affect the findings and recommendations of the evaluation, and how the threats can be addressed once they have been identified. Appendices 7.1 through 7.4 present worksheets for assessing the validity of QUANT, QUAL, and mixed-method designs, and for communicating the evaluation findings and recommended follow-up to managers and policymakers.

- *Step 7: Helping clients use the evaluation.* It is important to ensure clients and other key stakeholders are actively involved from the start and that they “buy into” the evaluation; maintain contact with clients throughout the evaluation and ensure that by the time the major reports are published they do not contain any surprises for the client; and adapt the presentation of findings to the preferred communication style of different stakeholders. On a broader level, this also involves helping institutionalize evaluation systems at the sector and national level (see Chapter 19). Smartphones and data analytics now offer attractive ways to easily present findings (e.g., data visualization) and to reach wider audiences.

## 2. THE REALWORLD EVALUATION CONTEXT

The RWE approach was developed to assist evaluators in both developing and developed countries to conduct evaluations with budget, time, data, political, and organizational constraints. In one common scenario, the client delays contracting an evaluator until late in the program when the funding agency (government, international development agency, foundation, etc.) is about to decide whether to continue to support a program or possibly launch a larger second phase. Such tardiness occurs even when evaluation was built into the original program agreement (see Box 1.2). With the decision point approaching, the funding agency may suddenly realize that it does not have solid information on which to base a decision about future funding of the program, or the program-implementing agency may realize it does not have the evidence needed to support its claim that the program is achieving its objectives. An evaluator called in at this point may be told it is essential to conduct the evaluation by a certain date and to produce “rigorous” findings regarding program impact although, unfortunately, no comparable baseline data are available.

In other scenarios, the evaluator may be called in early to help develop the monitoring and evaluation (M&E) plan but may find that for budget, political, or methodological reasons it will not be possible to collect data on a comparison group to determine program impact by comparing participants with nonparticipants (a **counterfactual**). In some cases, it may not even be possible to collect baseline data on the program population for purposes of analyzing progress or impacts over time. Data constraints may also result from difficulties of collecting information on sensitive topics such as HIV/AIDS, domestic violence, post-conflict reconstruction, or illegal economic activities (e.g., commercial sex workers, narcotics, or political corruption).

Determining the most appropriate evaluation design under these kinds of circumstances can be a complicated juggling act involving trade-offs between available resources and acceptable standards of evaluation practice. Often, the client's concerns are more about budgets and deadlines, and basic principles of sound evaluation design may receive a lower priority. Box 1.2 illustrates this difficult trade-off between budgets and deadlines on the one hand and desired standards of methodological rigor on the other. Failure to reach satisfactory resolution of these trade-offs may also contribute to a much-lamented problem: low use of evaluation results (see Chelimsky, 1994; Operations Evaluation Department, 2004 and 2005; Patton, 1997). RWE is a response to the all-too-real difficulties in the practical world of evaluation.

### BOX 1.2

#### A Familiar Evaluation Story

When a social development fund was launched in an African country a few years ago, it was suggested that a baseline study be conducted as the first phase of a longitudinal impact evaluation study. The program manager asked, "What is the point of spending money and time on a baseline study when we do not know if the program model will work in our country?" He also indicated that staff members were under pressure to launch the program and could not spend time on

something that would not be useful until the program was completed. Three years later, when the possibility of a second program was being discussed, consultants were called in to conduct an impact evaluation study. It was agreed that it was unfortunate that no baseline data were available to permit a rigorous measurement of the changes produced by the program. The consultants had to try to reconstruct baseline data using methods described in Chapter 5.

## 3. THE FOUR TYPES OF CONSTRAINTS ADDRESSED BY THE REALWORLD APPROACH

Table 1.1 illustrates the different ways in which RWE constraints interact in the contexts in which evaluations are conducted. In some cases, the evaluator faces a single constraint. For example, the budget may be limited but there is plenty of time. Or the evaluation may begin at the start of the program with no time constraint, but the evaluator is told that for political or ethical reasons it will not be possible to collect data on a comparison group. However, many unlucky evaluators find themselves simultaneously contending with several—or all—of these constraints!

### 3.1 Budget and Other Resource Constraints

Sometimes funds for the evaluation were not included in the original program budget, and the evaluation must be conducted with a much smaller budget than would normally be allocated. As a result, it may not be possible to collect the desired data or to reconstruct baseline or comparison group data. Lack of funds may create or exacerbate time constraints because evaluators may not be able to spend as much time in the field as they consider necessary. Box 1.3 makes the point that it is important to understand whether the main constraint is budget or time (or both), because the best strategy will often be different in each case.

**TABLE 1.1** RealWorld Evaluation Scenarios: Conducting Impact Evaluations With Time, Budget, Data, and Political and Organizational Constraints

The constraints under which the evaluation must be conducted				
Time	Budget	Data	Political and Organizational	Typical Evaluation Scenarios
X				The evaluator is called in late in the program and told that the evaluation must be completed by a certain date so that it can be used in a decision-making process or contribute to a report. The budget may be adequate, but it may be difficult to collect or analyze survey data within the time frame.
	X			The evaluation is allocated only a small budget, but there is not necessarily excessive time pressure. However, it will be difficult to collect sample survey data because of the limited budget.
		X		The evaluator is not called in until the program is well advanced. Consequently, no baseline survey has been conducted either on the program population or on a comparison group. The evaluation does have an adequate scope, either to analyze existing household survey data or to collect additional data. In some cases, the intended program impacts may also concern changes in sensitive areas, such as domestic violence, community conflict, women's empowerment, community leadership styles, or corruption, on which it is difficult to collect reliable data even when time and budget are not constraints.
			X	The funding agency or a government regulatory body has requirements concerning acceptable evaluation methods. For example, in the United States, the No Child Left Behind Act of 2001 includes funding preference for certain types of research designs. In other cases, a client or funding agency may specifically request qualitative data, tests of statistical significance regarding measured program effects, or both.
			X	There is overwhelming indication that the evaluation is being commissioned for political purposes. For example, an evaluation of the effects of conservation policy might be commissioned to stall its expansion.
			X	There is reason to suspect that the evaluation will be used for political purposes other than or contrary to those articulated in preliminary discussions. For example, an evaluator might suspect that an evaluation of charter schools might be used (and even misused) by a client with known advocacy for privatization of education.
			X	Problems of coordination or even rivalries among agencies with different objectives, implementation strategies, and monitoring systems often affect the design, implementation, and utilization of the evaluation.
X	X			The evaluator has to operate under time pressure and with a limited budget. Secondary survey data may be available, but there is little time or few resources to analyze it.
X		X		The evaluator has little time and no access to baseline data or a comparison group. Funds are available to collect additional data, but the survey design is constrained by the tight deadlines.
	X	X		The evaluator is called in late and has no access to baseline data or comparison groups. The budget is limited, but time is not a constraint.
X	X	X		The evaluator is called in late, is given a limited budget, and has no access to baseline survey data; no comparison group has been identified.

*Note:* To simplify the table, the possible combinations of political and organizational constraints with the other three factors have not been included in the table.

**BOX 1.3****Budget and Time Constraints Have Different Implications for the Evaluation Design**

While budget and time constraints often have similar consequences for the evaluation design, in other cases they can require very different approaches. For example, if an evaluation must be completed by a certain date, the process of data collection can often be speeded up by bringing in consultants, hiring more experienced researchers, or increasing the number

of interviewers. All these measures may require significant budget increases. If, on the other hand, budget is the main constraint, the decision might be made to contract with a local university that would use cheaper though less experienced graduate students who might require more time for data collection because they cannot work full time.

**3.2 Time Constraints**

The most common time constraint is when the evaluator is not called in until the program is already well advanced and the evaluation has to be conducted within a much shorter period of time than the evaluator considers necessary. Time constraints often make it impossible to conduct a pretest–posttest evaluation design with a baseline study that can be repeated after the program has been implemented. The time available for planning stakeholder consultations, site visits and fieldwork, and data analysis may also have to be drastically reduced to meet the report deadline. These time pressures are particularly problematic for an evaluator who is not familiar with the area or even the country and who does not have time for familiarization and for building confidence with the communities and the agencies involved with the study. The combination of time and budget constraints frequently means that international evaluators (and out-of-town U.S. evaluators) can be in the country or the state for only a short period of time—often requiring them to use shortcuts that they recognize as methodologically questionable.

**3.3 Data Constraints**

When the evaluation does not start until late in the program cycle, there is usually little or no comparable baseline information available on the conditions of the target group before the start of the program. Even if program records are available, they are often not organized in the form needed for comparative before-and-after analysis, or they measure activities and outputs but not outcomes. Program records and other documentary data often suffer from reporting biases or poor record-keeping standards. Even when secondary data are available for a period close to the program starting date, the data may not fully match the program populations on important indicators. For example, employment data may cover only larger companies, whereas many program families work in smaller firms in the informal sector, or school records may cover public schools but not religious and other private schools.

Many clients are only interested in collecting data on the groups or communities with which they are working. They may also be concerned that collecting information on nonbeneficiaries might create expectations of financial compensation or other benefits for which the program has no budget, further discouraging the collection of comparison group data. Even if funds are available, it is also often difficult to identify a comparison group, because many program areas have unique characteristics. Where intended program impacts concern sensitive topics such as women's empowerment, contraceptive usage, or domestic violence, especially in paternalistic societies, information may be difficult to collect even when funds are available (see Box 1.4). Similar data problems can arise when working with difficult-to-reach groups such as drug addicts, criminals, ethnic minorities, migrants, or illegal residents.

Since the publication of the second edition, there have been rapid advances in the fields of new information technology, which open up a wide range of new, and radically different, tools for the collection and analysis of data—offering both a vast new range of data that could hardly have been imagined even a few years ago, and new approaches to the synthesis and analysis of this data. These make it possible to collect and analyze data much faster and more economically. However, despite the great potential, evaluators have been much slower to adopt big data than have other areas of development programs such as early warning signals and emergency relief. The opportunities and challenges for incorporating new information technology are discussed in Chapter 18.

#### BOX 1.4

##### Problems in Capturing Information From or About Women

- Many household surveys only interview the “household head,” who is often considered to be a male. He often does not have all the information on female household members or gives low priority to their concerns. Many men, for example, say their wives are happy to spend several hours per day walking to collect water or fuel because they “sing and chat with their friends as they walk.”
- Women are often interviewed in the presence of other household members where they may not feel free to express their views.
- Donor agencies often insist that women be invited to attend community meetings to discuss proposed programs. However, the women often do not feel free to speak in public, or they always say they agree with their husbands.
- In many parts of the world, sexual harassment is one of the main reasons women do not use public transport. However, it is culturally impossible for women to mention this to an outside interviewer, so this major problem is often not captured in surveys.

### 3.4 Political and Organizational Influences and Constraints

We use the term *political influences and constraints* in a broad sense to refer not only to pressures from government agencies and politicians but also to include the requirements of funding or regulatory agencies, pressures from stakeholders, and differences of opinion within an evaluation team regarding evaluation approaches or methods.

Evaluations are frequently conducted in contexts in which political and ethical issues affect evaluation design and use. All programs affect some portion of the public, and most programs consume public funds, always limited and often scarce. Decisions based on evaluation results may intensify competition for funding, expand or terminate programs needed by some and paid for by others, or advance the agenda of a politically oriented group. Box 1.5 shows how political pressures often affect the evaluation design—in this case, forbidding the use of a comparison group.

While evaluators are always quick to spot the political or ideological biases of their clients and stakeholders, they are often less aware (or open) about their own ideological orientations. Many of the ongoing debates between quantitative and qualitative evaluators are fueled by the search for the “correct” or “best” research **paradigm**.

**BOX 1.5****Political Influence on the Evaluation of a Hydroelectric Program in Asia**

Consultants were asked to design an evaluation to assess the impacts of a hydroelectric power program in an Asian country that would involve the forced resettlement of a large number of villages in the area where the dam was to be constructed. Families who had title to their land would receive compensation. The consultants proposed that the evaluation should

include a comparison group of families who did not have land title. They were informed by the power authority that it would not be possible to do this because this would create expectations that these families would also receive compensation for being relocated, and funds for this were not included in the program budget.

## 4. ADDITIONAL ORGANIZATIONAL AND ADMINISTRATIVE CHALLENGES

In addition to budget, time, data, and political constraints, all evaluations must conform to the organizational arrangements under which they are commissioned, and to the administrative procedures of the agencies involved in commissioning, financing, managing, and using the evaluations. Often there are several different agencies involved in the evaluation, and they often have different goals for the evaluation. These may involve the kinds of information to be obtained, the preferred methodology, the stakeholders that should be involved and who is asked to comment on or approve the evaluation reports, the extent and form in which target populations are or are not involved, and how and to which audiences the evaluation will be disseminated and used. When several international agencies are involved, the logistics of arranging joint planning and supervision missions can be a major challenge, sometimes delaying implementation for significant periods of time.

Balancing the preferences and operating styles of different agencies can be a major challenge for the evaluation team, particularly in cases where there may be differences of opinion among stakeholders or lack of definition of their respective roles. Even when only a single agency is involved, its administrative and operating procedures may provide further constraints and challenges. For example, when local counterparts have to be contracted, the procurement procedures of the funding agency or the host government may produce long delays or require the use of contractual procedures that do not work well for a particular evaluation. In other cases, the requirement to prepare an inception report and to delay the start of fieldwork until different departments have commented on the report can cause significant delays in the start of the evaluation. In some cases, the date for the completion of the evaluation report is not changed, despite time lost waiting for feedback, so that the effective time for consultants to work on the evaluation may be significantly reduced. Another common problem is that a fixed amount of time is allowed for fieldwork in every country, even when it is well known that considerable numbers of days are likely to be lost in certain countries arranging travel to difficult-to-reach parts of the country or waiting for government clearance for travel. Often, when the evaluation consultants bring up these logistical problems, the evaluation manager will respond, “I entirely agree with you, but unfortunately, this is our administrative policy, so you will just have to do the best you can.” Box 1.6 illustrates common organizational and administrative challenges that many evaluations face.

**BOX 1.6****How Organizational and Administrative Considerations Can Affect Budget, Time, Data, and Political Constraints**

Organizational and administrative arrangements can affect how evaluations are funded; the ease or difficulty of planning, implementing, and disseminating evaluation findings; access to data; and how political pressures are controlled or exacerbated:

- *Budget constraints.* Many organizations have cumbersome, multistage budget approval processes and mechanisms for releasing funds. Some funds get “lost” at different stages of the authorization and release process. These delays can mean that some of the evaluation funds are released too late in the financial year to be used.
- *Time constraints.* Many programs involve multiple funding and implementing agencies, and coordination among these agencies and the approval of the evaluation plans can cause considerable delays. In one large population program in South Asia, several international funding agencies wished to be involved in the evaluation, and difficulties agreeing on a date for the planning and supervision missions caused delays of up to 6 months in the start of some phases of the evaluation.
- *Access to data.* There are several organizational constraints on access to data. First, funding agencies often each require their own M&E indicators and they are reluctant to accept standardized indicators. This makes it difficult to establish standard, comparable indicators across agencies working on the same project component. Second, some agencies may be unwilling to collect or make available certain kinds of data. This may be due to the inconvenience and cost of collecting or providing access to data. In other cases, such as gender-responsive evaluations, some agencies may not believe that gender is a relevant issue on which time and money should be spent. Issues such as gender, equity, or participation may also be considered controversial or sensitive.
- *Political constraints.* When different ministries or funding agencies are involved in a program, there may be competition to be the lead agency. Being appointed as the lead agency brings benefits such as invitations to international conferences, extra resources, and political leverage. Agencies that do not win this lead position may lose interest in a project or may even try to sabotage. One of the common ways to exert political pressure is through often complex organizational structures and coordination mechanisms. Even in the best of circumstances, coordination mechanisms are often slow and cumbersome, so there are many subtle ways in which the process of cooperation with the evaluation can be slowed even further.

## 5. THE REALWORLD APPROACH TO EVALUATION CHALLENGES

Although RWE does not develop many new data-collection or analysis methods, the approach makes several contributions to the conduct of evaluations under real-world budget, time, data, and political constraints. First, RWE draws from a wide range of evaluation approaches to the four types of constraints described earlier. Second, the systematic use of mixed methods is emphasized for several reasons: (a) It permits the evaluator to draw on the widest possible range of evaluation methods and tools, (b) it increases the validity of conclusions by providing two or more independent estimates of key indicators (triangulation), (c) it permits a deeper and richer analysis and interpretation of the context in which a program operates, and (d) it offers ways to reduce the costs or time of data collection (see Chapters 3, 4, and 14).

Third, RWE's seven-step approach offers corrective measures that can be introduced in different phases of the evaluation process, even after a draft evaluation report has been produced, helping to enhance the quality of the evaluation. Quality promotes credibility and utility of findings, which, in turn, helps ensure evaluation contributes to the public good.

Fourth, many quantitative evaluations rely on a pretest–posttest statistical counterfactual design to estimate the changes and impacts produced by a program. This approach, when used in isolation, has two serious limitations: (a) It does not take into account the different socioeconomic and political contexts affecting each program, and (b) it implicitly assumes that each program is implemented as planned and in exactly the same way in each location. One of the contributions of RWE is to look inside the “black box” of the program implementation process to examine what actually happens during implementation and how much variation there is between different program sites (see Box 1.7). RWE also focuses on quality of implementation. This is critical because in many real-world contexts, some program components are not implemented at all or the quality is so low that it is hardly surprising that the intended impacts were not achieved. In other cases, the intended impacts were achieved, but what went on within the program was quite different from what had been planned.

#### BOX 1.7

##### Getting Inside the “Black Box”

Many impact evaluations assume that programs are implemented exactly as planned and in exactly the same way in each location. In fact, there are often major differences in how each program is implemented depending on local cultural, economic, administrative, and political factors. In some cases, the pretest–posttest evaluation is faithfully conducted without realizing that some of the program components were never implemented at all. Women did not apply for loans because it was too

far to travel to the bank in town, teachers did not come to school during the planting season, textbooks never reached many of the schools, and parents in some areas did not send their daughters to school.

Unless the evaluation looks inside the “black box” of the program's implementation process, many of the findings of an impact evaluation can be very misleading and of little practical utility.

Fifth, RWE stresses the need to keep up to date with the development of new information technologies and to assess the potential benefits from incorporating some of the new tools into RWE designs. Evaluators have tended to lag behind other development practitioners in the adoption of many of these potentially important approaches, so evaluators are urged to broaden their horizons and to be open to the adoption of these approaches.

Finally, when designing an evaluation and during the consultations with diverse stakeholders, the evaluator should clarify the values on which the evaluation is based. How central are values such as equity, gender equality, empowerment, giving voice to vulnerable groups, and concern for the future of the planet to the evaluation? We argue there is no such thing as a value-free evaluation (even when the evaluator's and the stakeholders' values are not explicitly stated). Values affect the questions that are asked, what is measured, who is consulted, and who are the intended beneficiaries. Confusion is often caused when values are not explicitly stated or are only made explicit late in the evaluation cycle.

## 6. WHO USES REALWORLD EVALUATION, FOR WHAT PURPOSES, AND WHEN?

---

There are two main users of RWE. First, evaluation practitioners use RWE to

- identify ways to conduct adequately rigorous evaluations given limitations of time and financial resources;
- overcome data constraints, particularly the lack of baseline and comparison data; and
- identify and address factors affecting the validity and adequacy of the findings of the evaluation.

Second, government agencies, international development agencies, and foundations who commission evaluations and/or use evaluation findings will find the RWE approach useful to

- identify ways to reduce the costs and time of evaluations—or at least be aware of what an adequate budget and time frame would be required to conduct the kind of evaluation they may have in mind;
- be more fully aware of the various constraints under which evaluations are conducted, and what can be done to address those constraints; and
- understand the implications of different RWE strategies on the ability of the evaluation to respond to the purposes for which it was commissioned.

Table 1.2 shows that RWE can be conducted at three different points in a program: during the planning stage, when the program is already being implemented, or at the end. When the evaluation begins at the start of the program, RWE is used (a) to understand client information needs and the political context within which the evaluation will be conducted, (b) to help identify different options for minimizing costs or time required for evaluation while still providing valid information to meet stakeholders' needs, (c) for deciding what evaluation design would be appropriate, (d) for deciding what data need to be collected by the monitoring system during the implementation of the program, and (e) for deciding how to make the best use of available data.

When the evaluation does not begin until program implementation is already underway, RWE is used to identify and assess the different evaluation design options that can be used within the budget and time constraints and to consider ways to reconstruct baseline data. Attention will be given to assessing the strengths and weaknesses of monitoring and administrative data available from the program and the availability and quality of secondary data from other sources. The feasibility of constructing a comparison group may also be considered. When the evaluation does not begin until toward the end of the program (or when the program has already ended), RWE is used in a similar way to the previous situation except that the design options are more limited because it is no longer possible to directly observe the program implementation process. One of the innovative RWE approaches is to suggest measures that can be taken to strengthen the validity of the findings even up to the point when the draft final evaluation report is being reviewed.

TABLE 1.2 ● Who Uses RWE, for What Purposes, and When?

When Does the Evaluation Start?	Evaluation Practitioners Who Design or Implement the Evaluation	Managers and Funding Agencies
At the beginning of a program (baseline)	<ul style="list-style-type: none"> <li>• Identify a life-of-program evaluation design that will meet the needs of key stakeholders, given anticipated budget, time, and data constraints</li> <li>• Advise management how to reduce costs and time while achieving evaluation objectives</li> <li>• Negotiate with managers to relax some of the constraints (e.g., provide adequate budget and time) to reduce some of the threats to validity and adequacy</li> <li>• Advise management on plans for a baseline study consistent with evaluation objectives</li> </ul>	<ul style="list-style-type: none"> <li>• Be realistic in estimating the budget and time required for the proposed evaluation design, including the baseline study</li> <li>• Assess the relevance, required level of rigor, and quality of the proposed life-of-program evaluation design</li> </ul>
During program implementation	<ul style="list-style-type: none"> <li>• Identify ways to produce the best evaluation under budget, time, and data constraints</li> <li>• Identify ways for relevant monitoring data to be collected and documented that inform implementers and are relevant for evaluation purposes</li> <li>• If there was no baseline, reconstruct baseline data</li> <li>• Ensure maximum quality under existing constraints</li> </ul>	<ul style="list-style-type: none"> <li>• Identify ways to strengthen the ongoing monitoring and evaluation (these measures may be directly implemented by program management or funding agencies or recommended to the agency conducting the evaluation)</li> <li>• Keep data collection minimized and prioritized on information that informs decision making and learning</li> </ul>
At the end of the program	<ul style="list-style-type: none"> <li>• Identify ways to meet evaluation objectives within limitations of budget, time, political considerations, and data availability</li> <li>• Use the RWE checklist to identify and deal with threats to validity and reliability</li> <li>• Reconstruct baseline data</li> <li>• Ensure maximum quality under existing constraints</li> </ul>	<ul style="list-style-type: none"> <li>• Be clear on the purpose of evaluation and the relevant degree of rigor required</li> <li>• Identify ways to correct weaknesses in the evaluation within the budget and time constraints and/or be willing to allocate more funds and time to achieve required credibility</li> </ul>

## Summary

- Many evaluations are affected by budget, time, and data constraints or by political influences that limit the design options and data-collection methodologies available to the evaluator. We call these the *RWE constraints*.
- RealWorld evaluators most frequently face one of two scenarios. The first is when the evaluator is called in at the start of the program but the choice of evaluation design is constrained by budget or time pressures, by technical and administrative difficulties in collecting certain kinds of data, or by pressures from clients and stakeholders. The second and probably the most

common scenario is when the evaluator is not called in until the program has been underway for some time or is nearing completion. Often the evaluator is again subject to budget and time constraints and political pressures, but even when budget and time are adequate, it is usually the case that no systematic baseline data have been collected and no comparison group has been identified.

- We have found that the RealWorld Evaluation approach is applicable to varying degrees in all countries.

## Further Reading

American Evaluation Association. (2004). *Guiding principles for evaluators*. Fairhaven, MA: Author. Retrieved from [www.eval.org/Publications/GuidingPrinciples.asp](http://www.eval.org/Publications/GuidingPrinciples.asp)

The evaluation guidelines approved by the American Evaluation Association.

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

Covers many of the topics addressed by RWE.

Operations Evaluation Department. (2004). *Influential evaluations: Evaluations that improved performance and impacts of development programs*. Washington, DC: World Bank.

Case studies of evaluations that had a demonstrable influence on clients and stakeholders and a discussion of the factors determining whether evaluations will be used.

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

One of the most cited texts on how to design evaluations that will be utilized.

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

Discussion of the limitations of conventional impact evaluation designs, particularly for the evaluation of complex and evolving programs.

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

Chapter 2 introduces the evaluator-stakeholder relationship, and Chapter 12 discusses the social context of evaluation and the ethical issues discussed in this chapter.