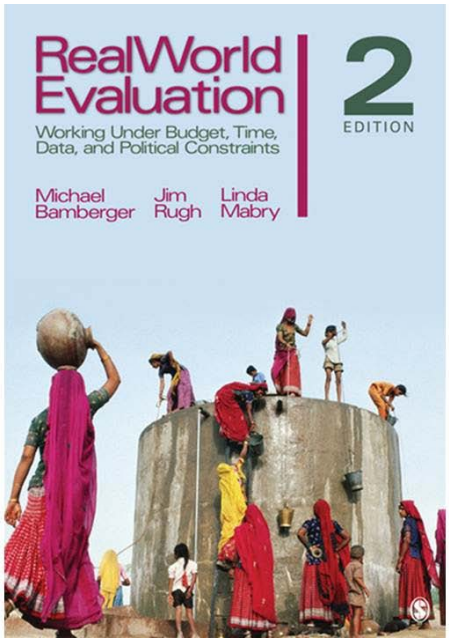


A Condensed Summary Overview

Michael Bamberger
Independent Consultant, Former World Bank Senior Sociologist

Jim Rugh
Independent Consultant, Former Evaluation Director for CARE International

Linda Mabry
Washington State University, Vancouver

SAGE
Los Angeles | London | New Delhi
Singapore | Washington DC

To order the book: www.sagepub.com/rwe

Contents

Preface

This book addresses the challenges of conducting program evaluations in real-world contexts where evaluators and the agencies commissioning evaluations face budget and time constraints and where critical data are not available or are of poor quality. Evaluators must also adapt the evaluation to a range of political pressures and influences and must work within organizational systems that often involve many different agencies and actors and where administrative procedures may not be well suited to the conduct of a rigorous evaluation. Evaluators are also often subject to pressures from many sides concerning the "right" evaluation methods to use, what should be studied (and not studied), who should be involved, and how and to whom the findings should be disseminated. While trying to juggle and reconcile all of the constraints (factors), it is also essential for the evaluator to follow standards of professional practice and to adhere to evaluation codes of conduct.

We were inspired to write a book about *RealWorld Evaluation* (RWE) because it became clear from our evaluation workshops and consultancies that most of the above topics were not systematically addressed in most evaluation textbooks. So while it was very easy to find textbooks and workshops explaining how to conduct a rigorous impact evaluation when budget and time are not major constraints, as well as when critical data are either available or can be collected, it was very difficult for evaluators and for agencies commissioning or using evaluations to provide practical guidance on how to conduct evaluations of an acceptable level of rigor and validity when working under budget, time, and data constraints, as well as when seeking to reconcile different political perspectives while working within organizational structures not well suited to conducting a rigorous evaluation.

Since the publication of the first edition of *RealWorld Evaluation* in 2006, there have been a number of important developments in the field of program evaluation, and the second edition has been considerably expanded to address these new issues as well as the feedback we continue to receive from the RWE workshops. These include continuing debates on appropriate evaluation designs, the challenges of evaluating complex development programs, how to manage evaluations and promote the institutionalization of evaluation systems, the emergence of mixed methods as a distinct approach to evaluation design, new developments in program theory, and the need to find ways to reduce sample size while ensuring statistically acceptable standards. Encompassing all of these is the recognition of the importance of following professional evaluation standards and codes of conduct and an understanding of the complex ethical issues affecting evaluation practice.

We will say more about this below and in the final chapter (copied below).

What's New in the Second Edition?

- A greater focus on responsible professional practice, codes of conduct, and the importance of ethical standards for all evaluations.
- Some new perspectives on the debate over the "best" evaluation designs. While experimental designs can address the important issues of selection bias, such statistical designs are potentially vulnerable to a number of important threats to validity. These include process and contextual analysis, collecting information on sensitive topics and from difficult-to-reach groups, difficulties in adapting to changes in the evaluation design, and implementation strategies. Experience also suggests that strong statistical designs can be applied only in a very small proportion of evaluations.

- There are many instances in which well-designed nonexperimental designs will be the best option for assessing outcomes of many programs, particularly for evaluating complex programs and even "simple" programs that involve complex processes of behavioral change.
- The importance of understanding the setting within which the evaluation is designed, implemented, and used.
- Program theory as a central building block of most evaluation designs. The expanded discussion incorporates theory of change, contextual and process analysis, multilevel logic models, using competing theories, and trajectory analysis.
- The range of evaluation design options has been considerably expanded, and case studies are included to illustrate how each of the 19 designs has been applied in the field.
- Greater emphasis is given to the benefits of mixed-method evaluation designs.
- A new chapter has been added on the evaluation of complicated and complex development interventions. Conventional pretest-posttest comparison group designs can rarely be applied to the increasing proportion of development assistance channeled through complex interventions, and a range of promising new approaches—still very much "work in progress"—is presented.
- Two new chapters on organizing and managing evaluations and strengthening evaluation capacity. This includes a discussion of strategies for promoting the institutionalization of evaluation systems at the sector and national levels.
- The discussion of quality assurance and threats to validity has been expanded, and checklists and worksheets are included on how to assess the validity of quantitative, qualitative, and mixed-method designs.

The RWE approach, originally called "shoestring evaluation," initially focused on problems of conducting evaluations in developing countries under budget, time, and data constraints. However, it soon became clear that despite the widespread use of evaluation in industrialized countries such as the United States, Canada, Europe, Australia, New Zealand, and Japan, many evaluators in these countries faced similar constraints. It also became clear that for many evaluators, political and organizational constraints were seen as bigger challenges than budget and time. Consequently, RWE now addresses a broader range of challenges and constraints faced by evaluators in both developing and developed countries.

Outline of the *RealWorld Evaluation* book (Second Edition)

	Preface	Introduction to the second edition of this RWE book
Part I: The Seven Steps of the RealWorld Evaluation Approach: *An overview for managers and readers who are not evaluation specialists and a refresher for "old hands"*		
Chapter 1	An overview of the full RWE approach	This chapter provides a preview of the seven steps and the basics of how the RWE tools and techniques can help evaluators and their clients cope with a variety of constraints typically faced when conducting evaluations in the RealWorld.

Chapters 2–8	Presentation of the seven steps of the RWE approach	Relatively nontechnical discussion that provides an introduction to the different stages of an evaluation. Cross-referenced to more detailed discussion of these topics in Part II.
Part II: A Review of Evaluation Methods and Approaches and Their Application in RealWorld Evaluation: *For those who would like to dig deeper*		
Chapter 9	Ensuring competent and ethical practice in the conduct of the evaluation	The basic "good practice" principles and standards that guide the evaluation profession.
Chapter 10	Theory-based evaluation	Theory-based evaluation is an essential building block in the RWE approach.
Chapter 11	Evaluation designs	Reviews the stages in the selection of an evaluation design and presents a wide range of experimental, quasi-experimental, and nonexperimental designs. Statistical rigor is only one of several dimensions of a methodologically sound design, and strong statistical designs are often weak on other important dimensions, while there are many situations in which nonexperimental designs are the best option. Appendix F presents case studies illustrating how each of the 19 designs has been used in the field.
Chapter 12	Quantitative approaches and methods	Reviews and contrasts the key elements of QUANT and QUAL approaches to evaluation. Shows how mixed-method design can combine the strengths of both approaches. Mixed methods is an integrated strategy involving unique approaches at each stage of the evaluation.
Chapter 13	Qualitative approaches and methods	
Chapter 14	Mixed-methods approaches	
Chapter 15	Sampling for RealWorld evaluations	Reviews approaches to sample design for QUANT, QUAL, and mixed-method evaluations. Explains the role of statistical power and effect size in estimating sample size. Introduces Lot Quality Acceptance Sampling (LQAS) as a practical operational tool for working with small samples.
Chapter 16	Evaluating complex, multicomponent development interventions	Discusses the move toward more complex, multicomponent, national-level development support and the demand that this creates for new evaluation designs as conventional evaluation designs can rarely be applied at these levels. A range of promising new approaches is discussed.
Part III: Organizing and Managing Evaluations and Strengthening Evaluation Capacity: *For readers involved with the funding and management of evaluations*		

Chapter 17	Organizing and managing evaluations	Defines and discusses the main stages in the preparation, recruitment of evaluators, implementation, dissemination, and use of an evaluation.
Chapter 18	Strengthening evaluation capacity at the agency and national levels	This covers quality assurance and threats to validity checklists, working with program management to design "evaluation-ready" projects, evaluation capacity development, and institutionalizing evaluation systems.
Chapter 19	Conclusions and challenges for the road ahead	Our final comments on some of the main approaches and issues discussed in the book.
Appendices		
Appendix A	Workbook for Assessing Threats to the Validity of the Findings and Recommendations of Quantitative (Experimental and Quasi-Experimental) Impact Evaluation Designs	
Appendix B	Workbook for Assessing Threats to the Validity of the Findings and Recommendations of Qualitative Impact Evaluation Designs	
Appendix C	Integrated Workbook for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Evaluation Design (Standard Version)	
Appendix D	Example of a Completed Threats to Validity Workbook	
Appendix E	Integrated Checklist for Assessing Threats to the Validity of Findings and Recommendations of a Mixed-Method Impact Evaluation (Advanced Version)	
Appendix F	A More Detailed Look at the RealWorld Evaluation Design Frameworks	
Appendix G	Using Concept Mapping as a Tool for Program Evaluation	
Appendix H	Applying the Complicated and Complex Program Evaluation Methodologies to Country Strategies	
	Glossary of Terms and Acronyms	
	Useful websites pertaining to RealWorld Evaluation	

The Seven Steps of the RWE Approach

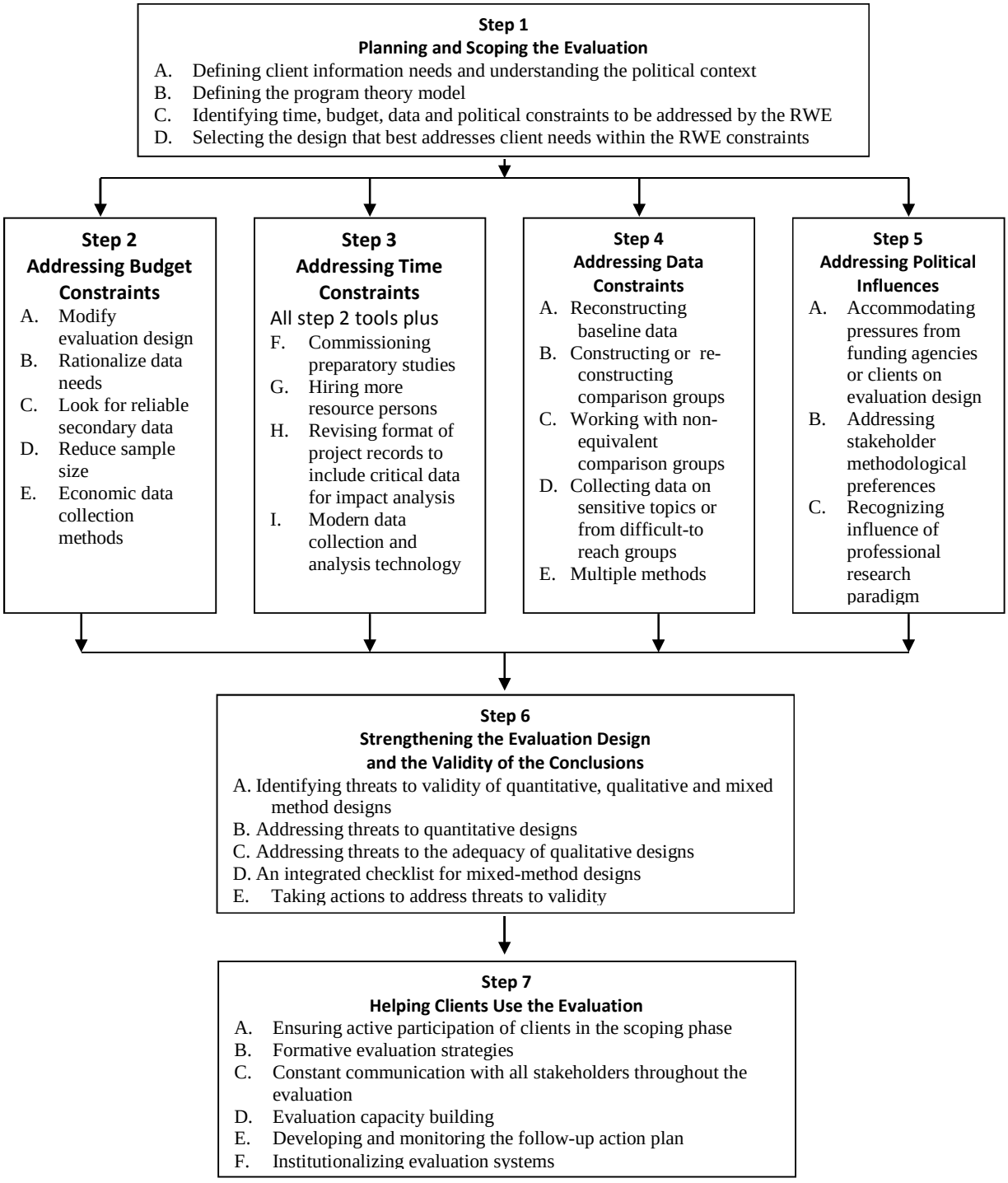


Table 1.1 RealWorld Evaluation Scenarios: Conducting Impact Evaluations with Time, Budget, Data, and Political Constraints

Time	*Budget*	*Data*	*Political*	*Typical Evaluation Scenarios*
X				The evaluator is called in late in the project 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 project is well advanced. Consequently, no baseline survey has been conducted either on the project 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 project 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	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 them.
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; and no comparison group has been identified.

NOTE: To simplify the table, the possible combinations of political constraints with the other three factors have not been included.

Purposes for RWE (Summary of Chapter 1)

- 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 main scenarios. The first is when the evaluator is called in at the start of the project 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 project has been underway some time or may even be 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 usually no comparison group has been identified.
- We have found that the RealWorld Evaluation approach is applicable to varying degrees in all countries.

Scoping the Evaluation (Summary of Chapter 2)

- 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 is 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.

Table 2.2 Seven Basic Impact Evaluation Design Frameworks

Key: *P = Project participants* *C = Control/comparison Group (Note 1)* *P_1, P_2, C_1, C_2 = First and second and any subsequent observations* *X = Project intervention*	Start of project [baseline / pretest] (Note 2)	Project Intervention (Note 3)	Mid-term evaluation	End of project evaluation [endline]	Post-project evaluation (some time after intervention ended) [ex-post]
Time period for evaluation event:	**T_1**		**T_2**	**T_3**	**T_4**
1. Longitudinal design with pretest (baseline), mid-term, posttest (endline) and ex-post observations of both project and comparison groups. [Note 4]	P_1 C_1	X	P_2 C_2	P_3 C_3	P_4 C_4
2. ***Pretest + posttest project and comparison group design*** **i.e. before-and-after plus with-and-without comparisons.**	P_1 C_1	X		P_2 C_2	
3. Truncated ***pretest + posttest of project and comparison groups where the initial study is not conducted until the project has been underway for some time*** (most commonly at the mid-term evaluation)		X	P_1 C_1	P_2 C_2	
4. ***Pretest + posttest comparison of project group combined with posttest (only) of comparison group***	P_1	X		P_2 C_1	
5. ***Posttest (only) comparison of project and comparison groups***		X		P_1 C_1	
6. ***Pretest + posttest of project group (****no counterfactual comparison group)*	P_1	X		P_2	
7. ***Posttest (only) analysis of project group (****no baseline or statistical comparison group)*		X		P_1	

Notes: (1) Technically a *control group* is only used in an experimental design (as randomization supposedly ensures there is no systematic difference in the distribution of subject characteristics between the two groups (i.e. selection *controls* for differences) and a comparison group is used in quasi-experimental designs where different selection procedures are used for the non-treatment group (sometimes called a "non-equivalent control group"). However, we will follow the common practice of using *comparison group* as shorthand for all kinds of matched groups, except when we wish to specifically indicate that randomization ***was*** used in which case we will use the term "control group".

(2) In this simplified table the point at which data is first collected on the project group (P_1) is also the time at which the evaluation begins. In Table 10.4 we distinguish between evaluations that start at the beginning of the project (and where baseline data is collected through primary data collection), and evaluations that start late in the project but where baseline data is obtained from secondary sources or through the baseline reconstruction techniques discussed in Chapter 5.

(3) The project intervention is usually a process that occurs over time, i.e. past the mid-term to the end of the life of the project.

(4) Though these designs call for direct collection of baseline data through surveys or other data collection instruments, there may be alternative ways for obtaining baseline data, e.g. from secondary census or survey data, or it may be reconstructed (see Chapter 5).

Table 2.3 Essential evaluation design components to complement any 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 is discussed in Chapter 9. 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 9
3. Multiple data collection methods	Many evaluations use a single method of data collection, for QUANT designs typically data is collected using a structured questionnaire. This is not adequate for collecting information on sensitive topics or on multi-dimensional indicators.	See chapters 12 and 13
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 socio-economic characteristics of target communities. This can result in wrong estimations of project impact, often leading to under-estimation of impacts (due to increased variance of the estimations).	See chapter 9
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 is enhanced when two or more independent estimates can be compared.	See Chapters 12 and 13

Coping with Budget Constraints (Summary of Chapter 3)

- Five strategies can be considered for reducing costs of evaluation planning, data collection, and analysis. (It should be noted that each of these may reduce the validity of results obtained.)
- The first is to simplify the evaluation design, usually by eliminating the collection of data on the project or comparison group before the project begins (pretest) or on the comparison group after the project is implemented (posttest) (see Chapter 11). In the simplest design, when data are collected on only the posttest project group, the data collection budget can be reduced by as much as 80%.
- The second is to agree with clients on the elimination of nonessential information from the data collection instruments.

- The third is to maximize the use of existing documentation (secondary data). See Chapter 5 for more details.
- The fourth is to reduce the sample size. Although this can produce significant savings, if the sample becomes too small, there is the danger of failing to detect statistically significant project effects even when they do exist. See Chapter 15 for more details.
- The fifth is to reduce the costs of data collection through methods such as the use of self-administered questionnaires, direct observation (instead of surveys), automatic counters, inputting data through handheld devices, reducing the number of periods of observation, prioritizing informants, and hiring and training students, nurses, and other more economical data collectors. It should be noted, however, that although these methods may reduce the cost of data collection, they will not necessarily reduce, or may even increase, the costs of data analysis.
- Most of the above strategies for reducing costs involve trade-offs because they pose threats to the validity of the evaluation findings and recommendations. The chapter concludes with a brief introduction to the assessment of threats to validity discussed in more detail in Chapter 7.

Coping with Time Constraints (Summary of Chapter 4)

- When identifying strategies to reduce time, it is important to determine whether there are also budget constraints or whether it is possible to increase expenditures to save time.
- Most of the cost-saving strategies discussed in the previous chapter can also be used to save time.
- Often, the main time pressure is on outside (often foreign) consultants who are available for only a short period of time. Their time can be used more efficiently by commissioning agency staff or a local consultant to prepare background studies or to do preparatory work for developing the methodology. Videoconferencing can also save outside consultant time (and often money). This also means that consultants can contribute to the evaluation design at the critical points when their input is most useful.
- Time can also be saved by increasing the number or raising the professional level of data collectors, field supervisors, and data analysts.
- In cases where the evaluator is involved at the start of the project, it may be possible to incorporate impact indicators into some of the administrative data forms used by the implementing agency to collect information that can later be used in the measurement of impacts.
- Modern handheld computers and similar technology can sometimes be used to reduce the time required for data collection and analysis.

As discussed in the previous chapter with respect to cost-saving strategies, most of the time-saving strategies involve trade-offs because they pose threats to the validity of the evaluation findings and recommendations. The chapter concludes with a brief introduction to the assessment of threats to validity, which is discussed in more detail in Chapter 7.

Coping with Data Constraints (Summary of Chapter 5)

- When evaluations do not begin until after the project has been underway for some time, the evaluator will often find that no baseline data have been collected and that no comparison group has been identified or studied.
- A number of strategies can be used to try to reconstruct the baseline conditions that existed at the time the project began. These include the use of documentary (secondary) data sources, interviews with key informants, using participatory methods such as PRA to help the community to recreate historical data and timelines, and the use of recall.
- While documentary (secondary) data are a valuable source of information, they were normally collected for a purpose other than evaluation of a particular project, and it is necessary to identify any biases or other factors that might limit the utility of some secondary sources.
- Additional challenges exist when reconstructing comparison groups because it is necessary to identify a group or community that is comparable to the project population as well as collecting information from this group.
- Many evaluations require the collection of sensitive data or collecting information from difficult-to-reach groups. This is a particular challenge for RWE because this information is often expensive and time consuming to collect, so there are often pressures to ignore these questions or groups. Some of the techniques for reaching difficult-to-locate groups include participant observation, use of key informants, tracer studies, snowball samples, and sociometric techniques.
- Like the two previous chapters, the present chapter ends with a discussion of some threats to validity arising from the innovative approaches that have to be used to reconstruct or obtain difficult and sensitive information.

Coping with Political Pressures (Summary of Chapter 6)

In evaluation, values and politics are inescapable, and no evaluation can ever be value free or completely objective. Decisions as to what to study, which methods to prefer, and whose criteria to use in determining program success all involve human judgment. Issues include maintaining access to data if political snags arise or if some groups are difficult to reach, respecting the values of various stakeholders, prioritizing conflicting interests appropriately, balancing stakeholders' and the public's "right to know" with client interests. Evaluators may encounter pressures to produce overly positive findings or to manipulate dissemination of reports.

Strengthening the Evaluation Design and Addressing Threats to Validity (Summary of Chapter 7)

The worksheets for assessing validity for QUANT, QUAL, and mixed-method designs all have three parts, each of which is targeted to a different audience (see Figure 7.3 below):[1]

[1] Detailed Threats to Validity checklists can be found in Appendixes A-E of the book.

Part 1. The cover sheet provides a one-page summary for senior management and for partner agencies. This explains the purpose of the evaluation and the reason for conducting the threats-to-validity assessment. It also summarizes the main conclusions of the validity assessment and the recommended follow-up actions.

Part 2. The summary assessment for each component is intended for middle-level management. It presents a half-page text summary of the validity assessment of each of the five components and a summary numerical rating (1 = very strong to 5 = serious problems). This provides sufficient detail for middle-level management to understand the main strengths and weaknesses of the evaluation and how these affect the validity of the findings and recommendations. In cases where only a general assessment of the evaluation quality is required, only Parts 1 and 2 of the worksheet may be used. However, when a more rigorous and comprehensive validity assessment is required, Part 3 can also be used.

Part 3. Between four and six checklists are included that permit a comprehensive technical analysis by an evaluation specialist of each of the components. Each indicator can be rated on a scale of 1 to 5 (Column A), where 1 indicates that the methodology is sound and 5 indicates there are significant methodological problems. A summary score can then be calculated for each checklist, indicating the overall strength or weaknesses of the methodology for each component. Column B can then be used to check how many of the methodological problems (i.e., those with a rating or 4 or 5) have important operational or policy issues for the purposes of this evaluation. The distinction between Column A and Column B is important because, depending on the purpose for which the evaluation was commissioned, and the types of decisions that will be based on its findings, there will often be a number of methodological weaknesses that may not have important implications for the intended users

The worksheets can be used at various points in an evaluation:

- During the evaluation design phase to identify potential threats to validity or adequacy. When important problems or threats are identified it may be possible to modify the design to address them. In other cases, if some of the potential threats could seriously compromise the purpose of the evaluation, further consultations may be required with the client or funding agency to consider either increasing the budget or the duration of the evaluation (where this would mitigate some of the problems) or agreeing to modify the objectives of the evaluation to reflect these limitations. In some extreme cases the evaluability assessment may conclude that the proposed evaluation is not feasible and all parties may agree to cancel or postpone the evaluation.
- During the implementation of the evaluation (for example a mid-term review). If the checklist had also been administered at the start of the evaluation it is possible to assess if progress has been made in addressing the problems. Where serious problems are identified it may be possible to adjust the evaluation design (for example, to broaden the sample coverage or to refine or expand some of the questions or survey instruments).
- Towards the end of the evaluation – perhaps when the draft final report is being prepared. This may still allow time to correct some (but obviously not all) of the problems identified.
- When the evaluation has been completed. While it is now too late to make any corrections, a summary of the checklist findings can be attached to the final report to provide a perspective

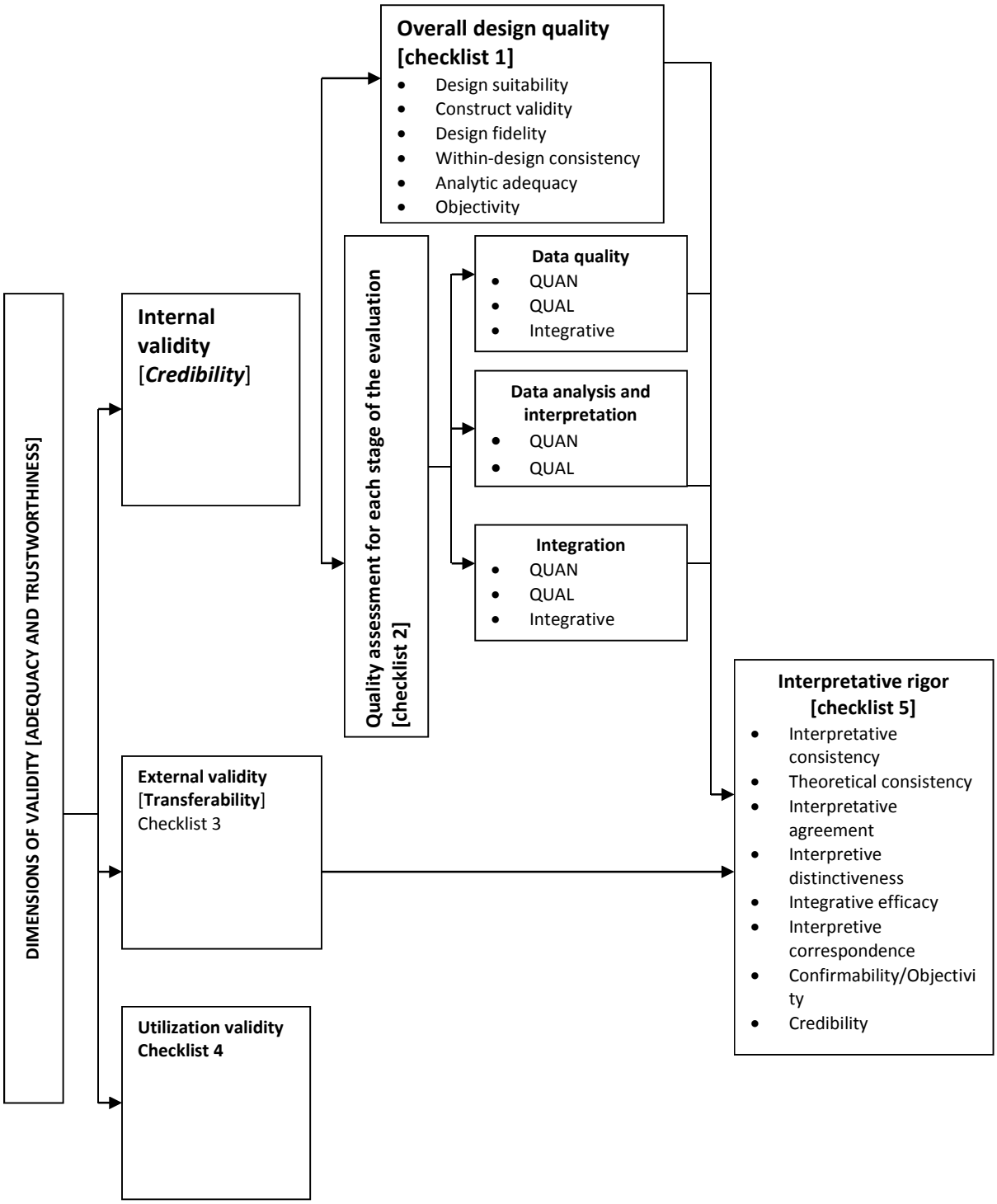
for readers on how to interpret the evaluation findings and recommendations and to understand what caveats are required.

- For organizations that regularly commission or conduct evaluations, a very useful exercise is to conduct a meta-analysis to compare the ratings for different evaluations to determine whether there is a consistent pattern of methodological weaknesses in all evaluations (or all evaluations in a particular country, region or sector). We discussed earlier how the checklist can be used at different points in the evaluation – for example, at the evaluation design stage, during implementation and when the draft final report is being prepared. When the scale is applied at these different points, it is possible to detect whether any of the threats are corrected or mitigated over time, or whether on the other hand some of them get worse. Differential sample attrition (between the project and control groups) is a familiar example where a problem may get worse over time as differences between the characteristics of subjects remaining in the project and the control samples may increase.

Chapter Summary:

- When referring to data the term *validity* is roughly similar to *accuracy.* It is a criterion used to assess whether data adequately and accurately represent actual conditions.
- The validity and adequacy of an evaluation are affected by (a) the appropriateness of the evaluation focus, approach, and methods; (b) the availability of data; (c) how well the data support valid findings; and (d) the adequacy of the evaluation team in terms of methodology, the specific field of the program, and the available resources.
- The validity of QUANT evaluations is usually assessed in terms of statistical conclusion validity, internal validity, construct validity, and external validity.
- QUANT evaluation designs can often be strengthened by (a) ensuring that random sample selection has been properly applied, (b) using triangulation to obtain independent estimates of key indicators, (c) correct selection of statistical procedures, and (d) using peer review and meta-evaluation.
- Once threats to QUANT validity have been detected, measures can be taken to correct or reduce their effects.
- The capacity of a QUAL design to support valid conclusions can be considered in terms of descriptive validity, interpretive validity, and evaluative validity.
- The Integrated Checklist for assessing evaluation validity and adequacy assesses QUAL and mixed-method evaluations in terms of confirmability, reliability, and dependability; credibility and authenticity; transferability and fittingness; and use, application, and action orientation.
- QUAL evaluation designs are strengthened through triangulation, validation, meta-evaluation, and peer review.
- Once threats to the validity of QUAL, QUANT, or mixed-method evaluations have been identified, measures can be taken to correct or reduce their effects.
- Measures can be taken to improve the validity of evaluations during the design stage, during implementation, or when preparing and reviewing the report on evaluation findings and conclusions.

Figure 7.3 Dimensions of Threats to Validity for the assessment of Mixed Method Evaluation Designs [QUAN terms are in bold and parallel QUAL terms are in bold italics]



Helping Clients and Other Stakeholders Utilize the Evaluation (Summary of Chapter 8)

Box 8.3 Guidelines for Helping Clients Use the Evaluation

1. Scoping the Evaluation

a. Understand the clients information needs
b. Understand the dynamics and timetable of the decision-making process.
c. Define the program theory on which the program is based in close collaboration with key stakeholders
d. Identify budget, time and data constraints and prioritize their importance.
e. Understanding the political context
f. Preparing a set of RWE options to address the constraints and strategizing with the client to assess which option is most acceptable

2. **Formative evaluation strategies**

a. Try to incorporate *process evaluation* and other methods that provide periodic feedback to clients on ways to improve project implementation.

3. Constant communication with clients throughout the evaluation

a. Keep clients informed about the progress of the evaluation and the preliminary findings and hypotheses.
b. Ensure there are "no surprises" for clients in the main evaluation reports

4. Strategies for overcoming political and bureaucratic challenges

a. Redesigning management systems to focus on results
b. Creating incentives for higher program performance
c. Developing agreement on key national, state or community indicators
d. Developing performance partnerships

5. "Carrots, sticks and sermons"

a. Providing rewards for individuals, departments and agencies that are committed to the implementation and use of evaluations
b. Providing sanctions for those who do not implement or use evaluations
c. Indications of support from high level and prestigious figures.

6. Evaluation capacity building

a. Actively involve clients and users in the scoping phase.
b. Ensure the program theory model is developed in a participatory way
c. Ensure users understand the trade-offs in the choice between RWE designs
d. Invite users to participate in the evaluation training programs for practitioners
e. Encourage users to participate in the periodic progress briefings on the evaluation
f. Involve users as resource people in briefings for other organizations that are planning evaluations.

7. Communicating the findings of the evaluation

a. Understand what users what to know
b. Understand how different users like to receive information
c. Understand the kinds of evidence users want (statistics, case studies, photos etc.)
d. Ensure presentations are pitched at the right technical level

e. Consider separate customized presentations targeted for different audiences
f. Ensure reports are available in the user's language

8. Developing a follow-up action plan

a. Ensure there is user buy-in to the evaluation so they are prepared to consider using relevant findings and recommendations.
b. Identify options but where possible let users decide the actions to be taken.
c. The role of the evaluation in the preparation of the action plan should be as a low-key technical resource and facilitator. Sometimes better not to attend all action planning meetings to allow more freedom to the users.
d. A key role for the evaluator is to ensure an action plan is prepared. The content should be left to the users to define as far as possible.

Developing a follow-up action plan

Many evaluations present detailed recommendations with very little practical utility because the recommendations are never implemented – even though all groups might have indicated their agreement with the proposals. What is needed is an agreed action plan with specific, time-bound actions, clear definition of responsibility, and procedures for monitoring compliance. Many government and international agencies have standard procedures to monitor the implementation of evaluation recommendations but such systems are used much less frequently for RWE-type evaluations. For example, many agencies keep a log of all recommendations included in their evaluations, management response to these recommendations and the agreed actions, and periodic follow-ups to report on the status of the agreed actions.

For RWE, as for many other evaluations, the definition of a follow-up action plan is an effective way to promote utilization of the evaluation findings. Some of the steps include:

- As we have stressed above, a key strategy involves ensuring client and stakeholder "buy-in" to the evaluation process so that there is willingness to review, and where there is agreement, implement the evaluation findings.
- The evaluation report must identify the key issues on which decisions must be taken and follow-up actions agreed. However the external evaluator needs to be cautious about presenting specific recommendations so as to not discourage users from taking ownership of the action plan. In preparing the report the evaluator, in consultation with the clients, must decide whether it is better to:
    - Present a list of issues but not propose specific actions;
    - Present a number of follow-up options but not recommend which one is best;
    - Present specific recommendations on follow-up actions. This may be appropriate when discussing technical issues (for example which financial management package is compatible with the computer systems used by the agency).
- The action plan must be developed by the interested organizations with the evaluator as a technical resource and possibly facilitator. It is sometimes better for the evaluator not to participate in the action planning meetings so as to give more feeling of ownership and freedom of action to the agencies themselves.
- Often the evaluator can help develop measurable indicators and timetables to monitor progress. One of the evaluator's key contributions is to ensure that the action plan is actually developed before she or he leaves.

Standards and Ethics (Summary of Chapter 9)

As their potential for harm to persons became manifest, researchers and evaluators became subject to restrictions from national, international, and professional organizations. Professional codes of conduct suffer the enforceability available with statutory guidelines but offer more specific guidance to practitioners in each of the various fields of social science. These guidelines typically target competent and ethical practice.

In evaluation, the *Guiding Principles for Evaluators* of the American Evaluation Association (2004) and the more detailed and multi-organization *Program Evaluation Standards* (Yarbrough et al. 2011) call for ethical treatment of all stakeholders and stakeholder groups, sensitivity regarding cultural contexts, and actions that promote fairness and social justice. Challenges regarding time and political constraints ensue for evaluators in the real world.

Applications of Program Theory in RealWorld Evaluation (Summary of Chapter 10)

- A program theory is an explicit theory or model of how a program is intended to produce the intended outputs, outcomes, and impacts, and the factors affecting or determining its success.
- A program theory is particularly helpful in planning an evaluation under RWE constraints because it helps identify the critical issues or hypotheses on which scarce evaluation resources should focus.
- Evaluators often have to work with clients and stakeholders to construct the implicit program theory because it has often not been formally articulated in project documents.
- Program theories often combine an impact model and an implementation model.
- The impact (change) model describes the linkages between project inputs, the implementation strategy, and the intended outputs, outcomes, and impacts. It also examines how performance is affected by *mediators*—factors affecting performance that can be modified by the project (e.g., willingness of different groups to support the project or to change their behavior)—and by *contextual factors* (such as the economic, political, organizational, natural environment, and characteristics of the affected populations) that affect performance but over which project managers have little control.
- The implementation (action) model describes how staff and resources are to be used to deliver the program services to the target population.
- Logical framework analysis (logframe) or Results-Based Management (RBM) are forms of program theory or modeling used by many agencies to monitor program performance against a set of measurable indicators. Such logic models can also be used to identify and test critical assumptions about conditions that will affect program success (e.g., actions that the government must take, the stability of the local and national economy, the willingness of communities or individuals to change traditional forms of behavior).
- Some, but not all, program theory practitioners believe that program theory can be used to help explain causality in situations where it is not possible to use randomized control trials or quasi-experimental designs.

Evaluation Designs (Summary of Chapter 11)

There are those who consider experimental research designs (or randomized control trials, RCTs) the methodologically strongest QUANT (statistical) designs for impact assessment.

However, the appropriate role for randomized designs for the evaluation of complex social development programs is hotly debated in the United States as well as in the field of international development. Even without these debates, the application of randomized designs—particularly for RWE scenarios—is limited by the technical, budgetary, ethical, and political difficulties in using a randomly selected control group that will be excluded from receiving the services being offered to the project group.

- One of the limitations of the typical RCT or experimental design is that it only measures quantitative indicators before and after, and with and without the intervention being tested. Typically, it does not provide a perspective over time, including an ex-post evaluation of the sustainability of any impact. A longitudinal design that combines pre- and posttest observations with data collected during project implementation and data collected after the project has been completed offers a more powerful design. However, longitudinal designs are used infrequently due to the additional cost and time requirements.
- Quasi-experimental designs (QEDs) seek to approximate the experimental design as closely as possible while adapting to realities of the RealWorld social context. One of the most robust QEDs involves pretest-posttest comparisons of the project group and a comparison group selected to approximate as closely as possible relevant characteristics of the project group.
- Even under the most favorable conditions, experimental designs or even QEDs, when used in isolation, have a number of limitations for impact assessment. A purely QUANT pretest-posttest design does not examine the project implementation process or the contextual factors (economic, political, organizational, environmental, and the characteristics of the affected populations) affecting project performance and results. It is also difficult to quantify many important input, process, and outcome indicators, and the rigidity of the design makes it difficult to adapt to changes in the project design and/or the evolution of the internal and external contexts over time.
- Seven variations of the basic QEDs are described that can be applied in RWE contexts. The different designs adapt to cost and time constraints and to the fact that many evaluations do not start until the project has been underway for some time and, consequently, no relevant baseline data were collected. Some of these designs save cost or time by eliminating one or more of the four major data collection points (pretest-posttest of project and comparison groups).
- The less robust designs involve trade-offs as the elimination of data collection points increases vulnerability to different threats to validity of the conclusions. A number of strategies can be used to strengthen all these designs, including (a) the use of mixed-method designs, (b) using a program theory model, (c) incorporating process analysis, (d) making judicious use of secondary data and recall, and (e) using multivariate analysis to more closely match the project and comparison groups.

Quantitative Evaluation Methods (Summary of Chapter 12)

- QUANT and QUAL methodologies represent two distinct social science traditions. It is important for evaluation practitioners, clients, and users to have a basic understanding of the two approaches; most social scientists have a professional or personal preference for one or the other of the two traditions, and this will often affect their approach to evaluation research.
- It is also useful for clients to understand these differences because QUANT- and QUAL-oriented evaluators may adopt quite different approaches to the same evaluation.

- Some of the advantages of QUAL approaches are that findings can be generalized to broader populations, subjects are selected to ensure their statistical representativity, QUALitative estimates can be obtained of project impacts, and the QUALitative contribution of intervening variables can be assessed. The combination of clearly documented procedures and standard instruments means that the research can be replicated in other settings.
- Some of the limitations of QUAL approaches are (a) many types of information are difficult to express numerically, (b) some groups are particularly difficult to reach using structured data collection instruments, (c) contextual factors are often ignored, (d) research methods are expensive and difficult to adapt to changing circumstances, and (e) statistical associations may be misleading if important explanatory variables are not included in the analysis.
- QUAL evaluators tend to use experimental (randomized) or quasi-experimental designs that rely on the application of structured data collection instruments that must be administered in exactly the same way at different points in the project to ensure comparability.
- Some of the principal data collection instruments include structured questionnaires, structured observation, physical measurements (height, weight), and knowledge and aptitude tests.
- Secondary data are an important, but often underutilized, source of data for QUAL evaluations.
- The management of the process of data collection is often as critical to the success of evaluation as the research design. Some of the common data management problems for RWE are (a) the best available sampling frame may not cover all the target population; (b) the actual sample selected may not correspond exactly to the sampling plan due to lack of enumerator supervision; (c) there may be high nonresponse rates; (d) enumerators may lack experience, may not speak the local language, or the team may not have the right ethnic, sex, or age composition; or (e) interviews may have to be conducted in the presence of other people, which may have affected the responses.
- QUAL data analysis may include descriptive data analysis, comparisons and relationships between groups, hypothesis testing, and analytical models.

Qualitative Evaluation Methods (Summary of Chapter 13)

Qualitative data collection, analysis, and reporting make heavy demands on real-world evaluations, but failure to attend to these demands compromises the validity of findings. QUAL designs are emergent, offering flexibility and increasingly well-targeted foci. Analysis and reporting are complex—attentive to many perspectives, diverse experiences, contrasting theoretical frameworks and criteria—and tend to offer site-specific (rather than broad) findings. The personal nature of much QUAL data raises challenging ethical issues. Yet, QUAL methods offer insights otherwise unavailable and lead to deep understanding. These insights and understandings facilitate constructive application and use of findings.

Mixed-Method Evaluation (Summary of Chapter 14)

- Mixed-method designs involve the planned use of two or more QUANT and QUAL methods of data collection and analysis.

- Building on the strengths of both QUANT and QUAL approaches, mixed methods can combine more comprehensive coverage with in-depth analysis of individual cases and a holistic understanding of the context within which each project is implemented.
- There are at least five reasons for using mixed-method designs: (a) strengthening validity through triangulation, (b) using the results of one method to help develop the sample or instrumentation of the other, (c) extending the comprehensiveness of findings, (d) generating new insights, and (e) incorporating a wider diversity of values.
- Mixed methods can be used at any stage of an evaluation.
- Mixed methods can either be used where one approach (either QUANT or QUAL) is *dominant* and the other approach is used as a complement, or both approaches can have equal weight. Mixed methods are used differently and bring different benefits depending on which approach is dominant.
- Mixed methods can be used either *sequentially,* when one approach is used after the other, or *concurrently,* when both approaches are used at the same time. The sequential approach is more widely used because it is simpler to manage.
- Although mixed-method approaches can be used at just one stage of the evaluation, a fully integrated mixed-method design involves more than simply combining data collection methods. A fully integrated approach involves (a) attention to the composition of the research team and allowing sufficient time to build relations between members from different professions; (b) integrating different conceptual frameworks and planning approaches during the design phase; (c) integrated data collection methods; (d) systematic use of triangulation during data collection; (e) integrating different approaches during data analysis, including the possibility of returning to the field to verify or elaborate on initial findings; and (f) combining different methods for the presentation of findings.

Sampling Strategies and Sample Size Estimation for RealWorld Evaluation (Summary of Chapter 15)

- Sampling issues are important for both QUAL and QUANT evaluations, but the approaches to sampling tend to be quite different in each case.
- QUAL evaluations tend to use *purposive sampling* to carefully select a small number of cases that represent all the main categories of interest to the study. Although random sampling would not be appropriate with these kinds of small samples, each sample is selected to ensure that the maximum amount of information is obtained.
- In contrast, QUANT evaluations normally use random sampling procedures to ensure that the selected sample is statistically representative of the total population so that generalizations can be made from the sample to this population with a measurable level of statistical precision.
- For QUANT evaluations, questionnaire administration and other forms of data collection usually represent one of the largest cost items in an evaluation; therefore, when RWE constraints require cost reductions, reducing the size of the sample is always a tempting option.
- However, if the sample is too small (as is often the case when there are budget and time constraints), it will not be possible to identify statistically significant relations between the project interventions and the production of the desired outcomes and impacts—even when they do exist.
- Consequently, deciding what is the appropriate sample size to achieve the desired levels of precision of the evaluation findings is one of the critical evaluation design decisions.

- Two key factors in the estimation of sample size are the estimated *effect size* (how large a change the project is expected to produce if it is successful) and statistical *power analysis* (the required level of significance that the project effect will be detected if it really exists). The smaller the expected effect size, the larger the sample needed to detect the effect. The higher the required level of confidence (power), the larger the required sample size..
- Estimating the effect size and adjusting the power of the test are two of the key factors in estimating sample size.
- When there are cost pressures to reduce sample size, this can be achieved either by accepting a lower power (a higher risk that a real project effect will not be detected) or by finding ways to increase the effect size (e.g., studying only those groups where the project is expected to have a larger effect).

Evaluating Complicated, Complex, Multicomponent Programs (Summary of Chapter 16)

- Development agencies are providing an increasing proportion of their development assistance through complex interventions that may involve multiple components or cover more than one sector, and that frequently operate at the national or international level, often in cooperation with one or more other donor agencies.
- Following the Paris Declaration and a number of follow-up international agreements, the assistance is also being provided more often in the form of general budget support to governments, making it difficult to identify how the funds have been used.
- All of these developments make it more difficult to evaluate the impacts of these interventions. A fundamental methodological challenge is the difficulty of defining a counterfactual that can be used to compare the situation with and without the project.
- Given these methodological challenges, many agencies believe that it is not possible to estimate the impacts that their particular resources and initiative have had; consequently, the focus is often on assessing the *plausible contribution* rather than *direct attribution*.
- Despite these challenges, there are wide ranges of methodologies available that could potentially be used to define alternatives to the conventional statistical counterfactual and that could provide estimates of the impact of the interventions, even though these estimates will often not be as statistically rigorous as is possible when evaluating the direct effects caused by simple project interventions.
- Potential methodologies for impacts of complex programs include quantitative, qualitative, mixed-method, and theory-based approaches. Rating scales, usually adapted from the OECD-DAC evaluation guidelines, are also used by many agencies to rate impacts as well as efficiency, effectiveness, relevance, sustainability, and sometimes other dimensions such as gender equity or systematic procedures for learning and disseminating lessons from the evaluation.
- There are also a number of general strategies that can be used to strengthen all complex evaluation designs: (a) disaggregating programs into simpler and more easily evaluated components; (b) portfolio analysis through which all activities that support a broad development objective are rated on a set of scales assessing the quality of design and implementation and potential outcomes; (c) using the techniques for reconstructing baseline data discussed in Chapter 5 to construct pretest-posttest counterfactuals; (d) creative use of available secondary data; and (e) taking advantage of ongoing or planned studies that could generate program or comparison group data.

Organizing and Managing the Evaluation Function (Summary of Chapter 17)

All evaluations are developed, managed, and used within a particular administrative system and within a particular political context, both of which affect how the evaluation is designed, implemented, disseminated, and used. An approach to evaluation that works well in one administrative system might work less well in a different system. There is also a wide range of ways in which the political system can influence how evaluations are designed, implemented, disseminated, and used.

The planning and management of an evaluation typically involves six main steps:

Step 1: Preparing the evaluation. This requires (1) the preparation of the evaluation framework or a Statement of Work (SoW) and the definition of what is to be evaluated, the activities needed to do the evaluation, responsibilities for each stage, timing, proposed methods, and resources, as well as where they will come from; (2) involving the stakeholders; and (3) defining the management structure. Often an additional preparation activity will be to commission a diagnostic study to help understand the context and nature of the program being studied.

Step 2: Recruiting the evaluators. A first decision is to decide whether the evaluation should be conducted internally or whether external evaluators should be commissioned. External evaluators can provide two main services: evaluation implementation and evaluation technical support. The latter can include a training program to strengthen evaluation capacity, assistance in scoping the evaluation, assessing evaluation proposals, evaluation quality assurance, and providing a resident evaluation adviser to provide long-term support.

Each organization has its own contracting procedures. There are a number of factors to be considered during the recruitment process: deciding whether to use a fixed-cost contract or level of effort (number of consultant days), whether to contract a firm or an individual consultant, and broadening the range of research expertise. With respect to the latter, many agencies traditionally draw on a limited range of evaluation expertise (e.g., mainly quantitative or mainly qualitative research), and the decision may be made to design the selection process to encourage a broader range of evaluators to participate.

Typical stages in consultant selection and recruitment include inviting EOI, issuing the RFP, and preparing the ToR (which will be included in the RFP but may be revised based on consultants' comments on the RFP or feedback from the inception report). Many agencies require an Inception Report, which can be used to revise and finalize the ToR.

Step 3: Designing the evaluation. (Evaluation design is discussed in more detail in Chapter 11.) This will typically involve the following: (1) formulating the main evaluation questions, (2) assessing the evaluation scenario, (3) selecting the appropriate evaluation design, and (4) commissioning an evaluability assessment.

Step 4: Implementing the evaluation. This typically involves the following steps: (1) defining the role of the evaluation department, (2) preparing and reviewing an inception report, (3) managing the implementation of the evaluation, (4) working with stakeholders, and (5) building in quality assurance procedures. There are additional issues to address when managing multiagency and complex evaluations.

Step 5: Reporting and dissemination. This involves providing feedback on the draft report and dissemination of the final report (see Chapter 8).

Step 6: Ensuring implementation of the recommendations. It is important to have clearly defined procedures to ensure that agreed actions on the recommendations are implemented. This involves coordinating the management response and follow-up and facilitating dialogue with partners.

Strengthening Evaluation Capacity at the Agency and National Level (Summary of Chapter 18)

1. Most evaluations are intended to provide information on program performance and recommendations on actions that management or policy makers should take to improve performance of ongoing programs or the selection and design of future programs. It is important to build in quality assurance procedures to ensure the findings are accurate and unbiased and that recommendations are based on all available evidence.
2. Some of the most common quality assurance procedures are (1) preliminary feasibility studies to ensure that the evaluation objectives could be achieved, (2) evaluability analysis to assess whether the proposed evaluation design could answer all of the questions in the Terms of Reference, and (3) a threats to validity analysis. The chapter presented three complementary threats to validity analysis designed to evaluate quantitative, qualitative, and mixed-method evaluations.
3. Threats to validity have traditionally focused on internal threats to validity (reasons why conclusions about the contribution of project and program interventions to explaining observed changes in the beneficiary population may not be valid) and external threats to validity (reasons why recommendations about the replicability of the program in other settings may not be valid). The approaches proposed in this chapter also assess utilization validity (arguing that if stakeholders do not use the evaluation findings, there must be questions about the validity of the design).
4. The threats to validity assessment for mixed-method evaluations is more complicated as this must assess both how effectively quantitative and qualitative methods are used as well as how well these approaches are integrated into a mixed-method approach at each stage of the evaluation.
5. Appendixes A, B, and C present worksheets that can be used to assess validity for quantitative, qualitative, and mixed-method approaches, respectively. Each worksheet includes three parts, each targeted for a different audience: a one-page summary for senior management summarizing the findings of the assessment and recommended actions that might be required before the evaluation report is accepted; a half-page summary of each component of the assessment for mid-level management, providing more detail on the reasons for the ratings given to each component; and checklists providing a detailed set of technical indicators for assessing the methodological quality of each component. This latter will be used when a detailed technical assessment is required and an evaluation specialist, either from within the client agency or contracted externally, will normally be required to conduct this technical assessment.
6. The second topic concerns ways in which closer coordination between evaluators and program management can improve the quality of evaluation design. Frequently, the evaluator has no involvement in the design of the program, and as a result, opportunities to use participant selection procedures such as randomized selection or regression discontinuity may be missed. Similarly, project data sources such as baseline studies, information about beneficiaries and those not selected, or monitoring

data on project implementation, which could be very valuable for the evaluation, are collected or archived in ways that cannot easily be used for the evaluation. Anther major frustration for many evaluations is that changes are made in how beneficiaries are selected or in the packages of services provided by the program, but these changes are not systematically documented. Consequently, it is very difficult for the evaluator to know the criteria on which different participants were selected or what services different groups of participants actually received. Finally, it is frequently the case that, due to administrative problems or resource constraints, not all beneficiaries receive all of the planned services. Frequently, the failure to deliver services as planned is not documented, so the evaluation may be designed on the assumption that everyone receives the complete package of services, while this may not be the case. Where closer cooperation between evaluators and program management can be achieved, which is admittedly often difficult to ensure, it is possible to design programs that are "evaluation ready," and the quality of the evaluation findings and recommendations can be improved.

7. The third topic concerns strategies for evaluation capacity development. While much evaluation training is targeted for the evaluation practitioners who conduct the evaluations, an effective evaluation program requires that five main groups between them have the capacity to understand how to define the need for evaluation and the kind of evaluation that is required at a given point in time, how to select the appropriate evaluation design, how to implement or manage the evaluation, and how to disseminate and use the evaluation findings and recommendations. Evaluation capacity can be developed in a number of ways, including courses and workshops, on-the-job training, guidance manuals and publications, and study tours.
8. The final topic covered is the institutionalization of evaluation systems. International and national donors have supported the conduct of high-quality evaluations in many countries, but there are still very few countries in which these individual evaluations have resulted in the development of a nationally managed evaluation system. One of the reasons for this is that many donors support evaluations that respond to their information needs and do not necessarily address the priority questions of concern to national governments. Similarly, the choice of evaluation methods is often strongly influenced by the methodological preferences of the donor.
9. Progress toward institutionalization involves most of the following: country ownership of the evaluation system, including the selection of the policies or programs to be evaluated; an increasing proportion of evaluation funding comes from national sources; there is a system for the selection, commissioning, design, implementation, dissemination, and use of evaluation with clearly defined audiences and purposes; and systems are put in place to strengthen the capacity of national evaluators. It has also been argued that the system must include "carrots," "sticks," and "sermons" to provide incentives, sanctions, and moral support to agencies to build evaluation into the programs they manage.
10. In countries that have made progress toward the institutionalization of evaluations, one of three paths tends to have been followed. The first starts with ad hoc selection of evaluations, taking advantage of funding opportunities or the interest of a strong figure in government. As these evaluations are found to be useful, central agencies gradually become more involved in the selection, management, and dissemination of the evaluations. Colombia is cited as an example of this approach. The second path is where a concerted effort is made to implement high-quality evaluations in a particular sector

such as education, health, or cash payments to the poor. Where these evaluations are found to be useful and where they gain international recognition, the decision may be made to expand the rigorous approaches to other sectors. Mexico is an example of this approach. The third path, and the least common, is where a national government makes a commitment to develop a national evaluation system that is introduced step by step over a period of years. Chile is an example of this approach.

11. At the time of writing, most progress has been made toward the institutionalization of evaluation in Latin America. However, progress is being made in other regions, and it is likely that over the next five years, it will be possible to cite a number of countries in Africa, South and East Asia, and perhaps the Middle East and Eastern Europe that have made significance progress toward the institutionalization of their evaluation systems.

Conclusions and Challenges for the Road Ahead (Summary of Chapter 19)

1. Conclusions

1.1. The RWE Perspective on the Methods Debate

There are a wide variety of purposes, evaluands, contexts, designs, and methods for conducting evaluations. Some of them were introduced in Chapters 2 and 11. Nevertheless, our main focus in this book has been on evaluations whose main purpose is to ascertain the impact of programs. There are also a variety of ways to design and conduct impact evaluations. But since this is still such a hotly debated subject among evaluators and those who call for "rigorous impact evaluation," we feel the need to clarify our perspectives on this important issue.

Experimental and quasi-experimental (QED) impact evaluation designs[2] are an essential part of the evaluation toolkit. In situations where randomized controlled trials (RCTs) can be used, they are able to statistically control for selection bias, a major cause of misinterpretation of evaluation findings. When only QEDs can be applied, they are able to partially control for selection bias and also, when properly applied, can identify the issues that have not been controlled for and the implications for the interpretation of findings and recommendations. Even in the many situations where experimental designs cannot be used, the logic of the experimental design can still provide a reference framework against which to judge the validity of findings from nonexperimental or weaker QED evaluation designs.

While the continuing debate on impact evaluation designs has had many positive effects, one of the unfortunate consequences has been that the often heated debates have caused many evaluators to react against what they consider the exaggerated claims of some advocates of RCTs. Consequently, the important reference point that experimental designs can provide for understanding potential threats to validity of nonexperimental designs has been ignored by many evaluators from other camps. On the other side of the debate, many of the websites promoting RCTs and other strong statistical designs could be accused of a selective presentation of the evidence as most

[2] We use "experimental" designs as a generic term covering randomized designs and strong quasi-experimental designs

of these sites only present examples where experimental designs have been used successfully, and it is extremely difficult to find any information on these sites about the frequency with which efforts to use these designs were not successful. So unsuspecting visitors to these websites could get the impression that experimental designs almost always work well and that they should be used wherever possible. However, experienced evaluators know there are many situations in which attempts to use experimental designs have run into problems (e.g., because of changes in the project design and treatments, changes in the participant selection criteria, problems in maintaining the comparison group, difficulties in implementing the project as planned, and external events that dramatically changed the project context). So there is still a need for a franker assessment of what has been the experience in the use of these designs and more objective discussions of when they are and are not likely to work.

It is also important to make a clearer distinction between *statistically strong* designs (designs that incorporate a robust counterfactual and control for selection bias) and *methodologically strong* designs. While experimental and strong QEDs provide statistically strong designs, they do not automatically ensure that the overall design is methodologically sound, and the logic on which these designs are based exposes them to a number of potential methodological vulnerabilities:

- *Inflexibility.* Most experimental designs require the replication of the same or a similar data collection instrument in the pre- and posttest applications to measure change.
- *Hard to adapt sample to changing circumstances.* They also require that the same, or a similar, sample be used in the pre- and posttest comparisons.
- *Hard to adapt to changing circumstances.* These requirements make it difficult to adapt the design to situations where the project design (services provided) or participant selection criteria may change. They also make it difficult for changes in the control group when, for example, parts of the control group are incorporated into the project or may vanish due to migration, urban renewal, or other factors.
- *Problems with collecting sensitive information.* Experimental designs usually require the use of a structured questionnaire that is often administered in a formal setting.
- *Mono-method bias.* Many experimental designs rely on one principal method of data collection and do not systematically incorporate triangulation. This is always a potential weakness, particularly for studying complex, multidimensional constructs such as poverty, empowerment, and vulnerability.
- *Difficult to identify and interview difficult to reach groups.* Many evaluations require interviewing difficult to reach groups such as the homeless, sex workers, people who are HIV positive, the landless, and illegal squatters/immigrants. Reaching these groups often requires a more qualitative approach that can be difficult to incorporate into the experimental design and sample frame.
- *Lack of attention to the project implementation process.* Many experimental designs rely on a pretest-posttest methodology that does not adequately assess the process of project implementation (the so-called black box problem). As most projects experience some deviations from the implementation plan, and there are often serious implementation problems, it is difficult to judge whether failure to achieve expected outcomes is due to *design failure* or to *implementation failure.*
- *Lack of attention to context.* Many experimental designs do not systematically analyze contextual factors such as the local and national economic,

political, organizational, sociocultural, and natural environmental factors that can affect outcomes in different project locations.

- *Focus on one intervention.* Typically, experimental designs test one intervention at a time. The unstated ideal is that of a "silver bullet" or "panacea"—one relatively simple intervention that will, by itself, lead to desired impact. In most RealWorld situations, combinations of multiple interventions (by more than one agency) or preconditions are required to achieve higher-level impact. In other words, they call for multilayered, more complex logic models, where plausible contributions from multiple sources are acknowledged and taken into account when determining what led to any observed changes in impact-level indicators.
- *Limitation of direct cause-effect attribution.* Again, the typical RCT tests directly attributable results of an intervention, redefining *impact* to be what others may consider short-term or intermediate outcomes rather than *impact* defined as higher-level, long-term, sustainable improvements in human well-being (like the indicators of the Millennium Development Goals (MDGs).[3]

Here's another way to express the caveat regarding RCTs: For an individual evaluator to attempt to conduct an impact evaluation of a program using only one predetermined tool is to suffer from myopia, which is unfortunate. On the other hand, to prescribe to donors and senior managers of major agencies that there is a single preferred design and method for conducting all impact evaluations can and has had unfortunate consequences for all of those who are involved in the design, implementation, and evaluation of national and international development programs. There is much more to impact, to rigor, and to "the scientific method" than RCTs. Serious impact evaluations require a more holistic approach.

In any case, experimental designs, whatever their merits, can only be applied in a very small proportion of impact evaluations in the real world.

A crucial issue that often gets overlooked in the methods debate is the fact that experimental designs can only be applied in a very small proportion of program impact evaluations where it is feasible (and ethical) to randomly select "intervention" and "control" subjects. While no hard statistics are available, it is often estimated that RCTs can probably only be applied in less than 5% of impact evaluations, and many would estimate the figure is much lower. Even strong quasi-experimental designs have been applied in perhaps only 10% to 25% of impact evaluations. So for many evaluators, the debate on the merits of experimental designs is largely academic as they may never have a chance to apply these in their whole professional career as evaluators.

Furthermore, most development agencies are moving from support of individual projects to support of broad development programs, often with multiple components, many different funding and implementing agencies, and often a lack of defined objectives or even target population. For most of these complicated and complex programs, conventional (relatively simplistic) experimental and quasi-experimental designs are generally not applicable.

So for all of these reasons, we are left with the question, "What kinds of impact evaluation designs are appropriate for the vast majority of development interventions where conventional experimental designs do not apply?" In contrast to

[3] As a reminder, the oft-quoted definition of impact by the Organization for Economic Cooperation/Development Advisory Committee (2002:24) is "the positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended. These effects can be economic, sociocultural, institutional, environmental, technological or of other types."

the very large literature on rigorous, quantitative experimental research designs, the evaluation literature has had very little to offer to the majority of funding and implementing agencies and evaluators. *This is a "missing piece" where we hope the RealWorld Evaluation (RWE) approach will make a contribution.*

1.2. How Does RealWorld Evaluation Fit Into the Picture?

The RWE approach was designed to address situations where evaluations have to be conducted under budget, time, and data constraints (see Chapters 3–5) and where it is necessary to reconcile different political interests (see Chapter 6) and to accommodate organizational structures and administrative procedures that may add further complications to the design, implementation, dissemination, and use of the evaluation (see Chapter 17). Strategies are proposed for addressing all of these challenges while at the same time seeking the most rigorous methodology that is possible within each context. An important component of the strategy is the use of a "threats to validity" worksheet to assess the methodological validity of each component of the evaluation, to identify potential threats to validity that affect the achievement of the evaluation's objectives, and to propose actions to address the challenges (see Chapter 7 and Appendixes A–E). The worksheet is designed in three parts so that the findings of the technical assessment of validity can be communicated in a brief and nontechnical language to various stakeholders, including program participants, evaluation managers, donor agency senior management, partner agencies, and (where appropriate) policy makers.

1.3. Selecting the Appropriate Evaluation Design

RWE also identifies a wide range of evaluation design options that could be considered in different situations. In Chapters 2 and 11, we proposed an eight-step strategy for selecting the appropriate evaluation design. In addition to a technical analysis of feasible design options, the process involves consultation and sometimes negotiation with the clients. It is essential that they understand and are in agreement with the proposed design. Where trade-offs must be made to accommodate to budget, time, data, administrative and political constraints, it is essential that clients fully understand the methodological and political consequences of decisions to cut budgets, reduce time, or respond to priorities or concerns of key stakeholders.

We start with the premise that there is no single "best" evaluation design that will work well for all kinds of evaluation. It is essential to begin with a scoping study (Step 1 of the RWE approach—see Chapter 2) to fully understand the following:

- What is the purpose of the evaluation, how do clients intend to use the evaluation, and what key kinds of information do they require?
- In addition to the clients paying for the evaluation, who are the other key stakeholders, what are their perspectives on the program and the purpose of the evaluation, how do they relate to each other, what roles should they play in the evaluation, and what kinds of information would be useful to them?
- What are the contextual factors likely to influence the choice of evaluation questions, how the evaluation will be implemented, the kinds of constraints that will be faced, and how the findings will be used (or not used)?

What are the key characteristics of the "evaluand" (the program or activity being evaluated)? This will include dimensions such as purpose, size, budget, implementation strategy, previous experience, qualifications of the staff, and

relations with the target population (including whether and how they were involved in the choice and design of the project).

It is then necessary to understand the circumstances and context within which the evaluation will be conducted, including the following:

- Why is the evaluation being conducted?
- What is it supposed to achieve?
- Who promoted it, who supports it, and who does not? What are the opinions of the clients? Are they in favor or opposed? Do they understand the proposed methodology and do they agree with it?
- What are the preferred methodologies of the funding agency, the national government agencies, and the implementers? Do they have strong preferences for or against quantitative and qualitative methods? Do they support or oppose participatory approaches?

To assist with this process, RWE identifies 7 basic evaluation design framework scenarios (see Table 11.2) and a total of 19 more nuanced evaluation design options (see Table 11.3) that take into consideration the following:

- When the evaluation primary data collection is conducted (at the start of the project, during implementation, at the time the project closes, or sometime afterwards)
- What is used as a counterfactual—that is, whether a comparison group is used (at some or all points of the evaluation) and how it is matched with the project group (random assignment, statistically matched quasi-experimental comparison group, or judgmental matching), or whether secondary data or key informants or recall or other methods are used to determine an adequate counterfactual.
- Whether relevant baseline data were collected and used and how they were obtained (primary data collection, use of secondary data or through "reconstruction" of the baseline when the evaluation does not begin until late in the project cycle)

The choice of the appropriate design is critical and must involve both the findings of the scoping study (see Chapter 2) and methodological considerations and an evaluability assessment to determine if the preferred design is feasible. It is important to recognize that there is almost always more than one possible design option, and the final choice must combine technical considerations as well as the purpose of the evaluation, client preferences and a full understanding of the evaluation context. Although many clients will expect the evaluation "expert" to tell them which is the "best" design or the appropriate sample size, it is important to understand that there are political dimensions to the choice and that the client and other key stakeholders must be fully involved in the decision.

1.4. Mixed Methods: The Approach of Choice for Most RealWorld Evaluations

Almost all program evaluations require a combination of depth, to understand the lived experiences of individuals and groups affected by the program; and breadth, to generalize from in-depth qualitative methods to the broader population (see Chapter 14). This requires a combination of quantitative (QUANT) and qualitative (QUAN) methods. While many, perhaps most evaluators claim to have used mixed methods, in many cases they are only including one additional method of data collection to complement an approach which is mainly QUANT or QUAL. In fact mixed methods is an integrated approach to evaluation that has a unique approach to all stages of the evaluation from hypothesis development, through sample design, data collection, data analysis and triangulation, interpretation and dissemination of findings.

Among the many contributions that mixed methods make to RealWorld evaluation are: the following:

- Hypothesis development strategies that combine deductive (QUANT) hypotheses that can be tested statistically, on one hand, and inductive, emergent (QUAL) hypotheses, on the other hand, that have the flexibility to adapt as more information is obtained on the program and its context
- Sample designs that permit the use of QUAL purposive sampling to select a small sample of subjects that provides the highest value for the purposes of the evaluation, with QUANT random sampling procedures that permit generalizations from the QUAL sample to the broader program population
- Triangulation that helps strengthen the reliability and validity of data by using two or more independent methods to check for consistency or to provide a broader context for understanding the complexity and multidimensionality of what initially appeared to be "simple" concepts.

1.5. Greater Attention Must Be Given to the Management of Evaluations

When we first began organizing workshops on "shoestring" evaluation and, later, "RealWorld" evaluations, we had assumed that budget, time, and data would be the main constraints. But it quickly became clear that for many evaluators, the main challenges involve accommodating different organizational priorities and perspectives, as well as having to work within administrative systems that are often inflexible and not designed to accommodate the real world within which evaluations are designed, implemented, and used. So while all evaluators would always like more time and money, as well as easy access to all of the required information, often their major headaches concern things such as

- Pressures from their own organization or partner agencies (e.g., other donors or host country government agencies) to not "rock to boat" by being "too negative" and to avoid raising sensitive questions
- Pressures to not interview certain groups, including control groups and critics of the program, and not to address certain issues
- Inflexible procurement arrangements for hiring consultants
- Unrealistically short amounts of time that consultants can spend in the field. Allowable days in each country often ignore the significant amounts of time that everyone knows are required to obtain travel clearance from government (once in the country) or the logistical problems and delays in arranging in-country travel.
- Long delays in obtaining feedback on evaluation designs, inception, and progress reports Difficulties in obtaining information from partner agencies or arranging interviews with key people in these agencies
- Internal coordination problems (not to mention rivalries) within partner agencies

For reasons such as these, two additional chapters were added to this second edition of the RWE book: "Organizing and Managing Evaluations" (Chapter 17) and "Strengthening Evaluation Capacity" (Chapter 18). Some of the key conclusions from these chapters and from the many references to these issues in earlier chapters include the following:

- The importance of developing an evaluation framework (some agencies call this the "Terms of Reference"; others call it a "Scope of Work" or "Statement of Work") that spells out the basic purpose, objectives, major questions to be answered, proposed

methodology, division of responsibilities, resource requirements and sources, timelines, and deliverables. This should be developed and negotiated in consultation with all partners. This step is crucial to avoid confusions or misunderstanding at a later point. It is also important to ensure that all budget requirements have been realistically estimated and that budgetary approval has been obtained.

- Ensure that the evaluation proposal reviews alternative designs and explains why a particular design has been proposed. This is important because many clients (and evaluators) have preferred designs that they propose for all evaluations without studying the specific requirements of each evaluation.
- Commission an evaluability analysis to ensure that the proposed evaluation design is feasible within the RealWorld constraints and that it can respond to all of the information requirements. Sometimes evaluations are commissioned too early in the life of the project before it is possible to have achieved and measured outcomes and impacts, so this should receive particular attention.
- An important aspect of the evaluability analysis is the trajectory analysis in which the way impacts are expected to evolve (both in terms of time and the shape of the impact trajectory) is assessed (see also Chapter 10). Evaluations that are conducted too early or too late (in cases where impacts are not expected to have a very long duration) can fail to capture impacts that may really have occurred.
- A follow-up plan to ensure agreed-to recommendations are acted on is critical. Due to the pressure of other activities, often the agreements are forgotten, so a management log that tracks implementation of agreements can be helpful.
- The following section emphasizes that an important management function is to build in a quality assurance system so that the quality of evaluation findings and recommendations are routinely assessed.

1.6. Quality Assurance

Many evaluation departments and agencies that commission evaluations do not have systematic procedures for assessing the quality of the evaluation and for taking actions to address any weaknesses that are identified. Sometimes evaluations are assessed through a peer review process where reviewers are not given clear guidance on the assessment criteria and where consequently each reviewer uses his or her own assessment criteria. Often, as a result, important aspects of the evaluation design may not get assessed. Even more frequently, there is no plan, budget, or time to follow up on any recommendations concerning issues with the design or findings. Consequently, it is recommended that the Terms of Reference (ToR) for the evaluation should include budget and time to allow for any follow-up actions that might be required. And consultants must agree to ensure the appropriate members of their team would be available to respond to these requests.

Experience shows that the success of quality assurance depends to a significant degree on the specificity of the ToR concerning the methodology and responsibilities of the consultants. Frequently, the consultants will respond to critiques or requests for further analysis by stating, often with justification, that these requirements were not clearly stated in the ToR.

Chapter 7 proposes a set of threats to validity worksheets and checklists that can be used to assess the validity of findings and recommendation of QUANT, QUAL, and mixed-method evaluations. Appendixes A to C present the worksheets, and Appendix D gives an example of a completed worksheet. Part III of each worksheet permits a detailed technical

assessment of validity on each major dimension, Part II provides a short summary for evaluation managers, and Part I presents a one-page summary of the assessment and proposed follow-up actions for senior management and for other partners.

1.7. The Challenge of Institutionalization

In many countries, donors have been supporting impact evaluations for many years, and many of these evaluations have been of a high technical quality. However, much less attention has been given to helping governments build a *national monitoring and evaluation system* to ensure that central planning and finance agencies have a strategy for identifying each year the evaluation questions that have the greatest policy relevance, commissioning evaluations that will address these questions, and ensuring that the findings contribute to key budget planning and policy decisions. This requires a clearly defined strategy by donors and governments for promoting the institutionalization of impact evaluation at the national and sector levels. This remains a weakness in most donor approaches to evaluation.

1.8. The Importance of Competent Professional and Ethical Practice

In recent years, a number of important standards and best practice guidelines have been developed for evaluators and the agencies that commission and use evaluations (see Chapter 9). While these are widely, but not universally, used in many Western countries such as the United States, Canada, and Europe, they have been less frequently used in development evaluation. There is an even bigger gap with respect to ethical standards, and while the U.S. federal government and many other agencies have strict standards for research on human subjects (most research has to go through an institutional review board, also known as an independent ethics committee or ethical review board), this is not always the case in international evaluation. For example, many international development agencies do not have formal guidelines covering human subject research, and guidelines on how to implement a "do no harm" policy are often not clearly defined.

However, ethical issues affecting evaluation practice are often complex[4] and are even more challenging when working in a different culture.[5] They are made more difficult when having to accommodate different cultures and different national government policies. Some agencies try to follow government policies, but this can lead to further complications when working in countries that have different approaches to privacy, gender equality, rights of minorities, and dissemination of information. These are important issues that require greater attention and discussion.

1.9. Basing the Evaluation Design on a Program Theory Model

The RWE approach stresses the importance of basing the evaluation design on a program theory model (theory of change or logic model). While a conventional experimental design can assess whether a particular outcome has been achieved and whether it is reasonable to attribute this to the program intervention, this result tells us little about why and how the program contributed to the outcome, what were the key

[4] The wide range of topics covered in the *American Journal of Evaluation*'s Ethical Challenges section, as well as the complex and subtle nature of many of these issues, illustrates the difficulties of addressing these issues, even when the evaluator is working in her or his own country

[5] See the Cultural Competency Statement by American Evaluation Association (2011) at http://www.eval.org/ccstatement.asp.

elements of the approach, and under what circumstances would similar outcomes be achieved if the program were replicated in a different setting. Even more important, if outcomes were not achieved or were different than expected, an experimental design, when used in isolation, offers little to help us answer these questions.

A well-designed program theory will spell out all of the steps through which outcomes are to be achieved, the key assumptions and hypotheses to be tested, and how outcomes are likely to be affected by contextual factors, including assumptions concerning expected conditions and contributions by external sources. It is useful to spell out this process in more detail in a results chain, which can also identify potential negative outcomes at each stage if the model does not operate as planned. The theory model also helps define the key inputs, process, output, outcome, and impact indicators that should be measured and helps identify the hypothesis that should be tested. Although this is more controversial, many authors argue that a well-articulated program theory model can also help test causality.[6]

So a well-articulated program theory model can greatly strengthen the evaluation design and the interpretation of findings. However, despite this great potential, many program theory/logic models have proved to be of very limited practical utility. One of the most common reasons is that once they have been developed, often at the instigation of the funding agency prior to approval of funds for a proposed project, they are often subsequently put in a drawer and forgotten. In other cases, the level of detail (pages and pages of indicators) is so overwhelming that no one in the agency understands how to use them. Another common problem is that the process of designing the program theory is guided by the donor, and therefore implementation agency staff do not feel they have ownership of the process or may feel embarrassed to suggest the basic indicators that they would like to use when the planning workshop is directed by a highly paid international consultant who has flown halfway around the world to direct the workshop.[7]

1.10. The Importance of Context

How programs are formulated, designed, and implemented, as well as what benefits they produce and for whom, are all affected by the particular constellation of economic, political, organizational, sociocultural, and environmental (to name but a few) factors that operate in that particular context. So to explain why the program was not implemented exactly (or even approximately) as planned, why certain outcomes were or were not achieved, and why certain groups did or did not benefit, it is essential to understand how the program interacted with its setting and how it was influenced by the particular constellation of contextual factors.

Many programs consist of multiple subprojects that operate in quite a few different settings, and in most cases, the characteristics of many of these contextual factors will vary from location to location. Consequently, their individual and combined influence on program implementation and outcomes is likely to vary from location to location. Thus, the identification and analysis of these contextual factors

[6] If implementation proceeds according to the program theory and if the expected outcomes are achieved, this gives some credibility to the claim that that the program contributed to the outcomes. However, the claims are more credible if alternative models are developed to test rival hypotheses.

[7] Michael Patton (personal communication) talks about program theory workshops being seen as a "guessing game" where participants believe that the instructor knows the correct answer and that they are being asked to guess this answer—which makes participants feel they are being treated as schoolchildren.

are critical for explaining variations in project performance and outcomes in different locations, and the evaluation design should include a framework for the analysis of contextual factors and for assessing their influence.

Unfortunately, most QUANT impact evaluations do not include a systematic analysis of contextual factors. Often discussion will be limited to the presentation of anecdotal evidence to explain one particular variation in outcome. While QUAL evaluations will usually include a discussion of context, it is not always done systematically.

1.11. The Importance of Process

As we discussed earlier, projects are rarely implemented exactly as planned, and sometimes how a project is actually implemented can be significantly different from the original design (sometimes for good reasons). Also, as we've already mentioned, when projects operate in different locations, implementation will often vary from location to location. Sometimes the differences are due to unanticipated problems, poor management, or because more effective ways were found to implement the intended program. There are other cases where the original objectives may have changed due to changed circumstances or new government policies or because the intended beneficiaries were able to adapt the program to their needs. Ray Pawson's (2006) *realist evaluation* approach also argues that every beneficiary influences how a program evolves. People tell their neighbors and friends what they did and did not like about the program or perhaps how they can manipulate the rules to obtain more or different benefits.

Even when the stated purpose of the evaluation is to assess outcomes, it is clearly important to understand these processes as they can dramatically affect the achievement of outcomes. Many impact evaluations either ignore implementation processes or only study them through project reports or interviews with project staff, both of which can be very misleading.

1.12. The Evaluation of Complicated and Complex Programs

As discussed earlier, during recent years, international donor agencies have moved toward broad program support for packages that involve many different activities and often involve a number of different donors and government agencies. These interventions are defined as complicated or complex because they have multiple and often not clearly defined objectives, no clearly defined target populations, no clear start and end dates, and often multiple sources of funding.

Given these complexities, it is usually impossible to apply conventional experimental and quasi-experimental impact evaluation designs. Many donor agencies believe that it is not possible to conduct a rigorous assessment of such complex interventions, and often they have come to rely on commissioning consultants to apply a set of rating scales, sometimes adapted from the Organization for Economic Cooperation/Development Advisory Committee (OECD/DAC), where it is often difficult to judge the basis on which the assessments were made.

One of the messages of the RWE approach (see Chapter 16) is that a wide range of promising approaches can provide a credible level of evidence in many contexts. However, there are currently very few widely accepted and tested approaches, so creativity is required, with strong emphasis on the use of multiple methods and triangulation to assess consistency of the estimates from different sources.

2. The Road Ahead

In this section, we identify some of the key challenges that RWE, and program impact evaluation in general, must address over the next few years.

2.1. Developing Standardized Methodologies for the Evaluation of Complex Programs

As we have discussed, there are few standardized approaches for the evaluation of complex development interventions, such as country assistance strategies, generalized budget support, thematic evaluations (such as gender mainstreaming or strengthening local capacity for the planning and management of evaluations), or post-conflict reconstruction. In Chapter 16, we presented a number of promising approaches, and a goal over the next few years will be to test these approaches in different contexts and to develop some guidelines on how they can be incorporated into the evaluation toolkit of development agencies and national partners.

2.2. Creative Approaches for the Definition and Use of Counterfactuals

There remains a widely held perception that a statistically matched (randomly chosen) comparison group is the only acceptable form of counterfactual and that where such a counterfactual cannot be used, it is not possible to conduct methodologically sound impact evaluation. We totally agree that the experimental counterfactual is a powerful tool whose use should be considered where feasible and appropriate, usually as one of several complementary evaluation techniques. However, as Leeuw, Ling, and Rieper (2010) and Scriven (2009) point out, many if not most social and natural sciences (including qualitative evaluation; criminal justice; forensics; many branches of economics, including economic history; geology and astronomy) rarely if ever have access to experimental designs requiring randomly assigning subjects to treatment and control groups. Instead, they use other logical and reasoning approaches to present credible evidence that supports conclusions "beyond a reasonable doubt." A challenge for RealWorld evaluations is to draw on these approaches, as well as the many similar approaches already being used within the evaluation field to broaden the range of what can be considered "credible evidence" to support a broader range of logically defensible counterfactuals.

Government and donor agency decision makers are almost never presented with statistically significant tests proving that a particular intervention was responsible for an observed outcome. When making a decision on whether to continue or modify a program, they normally have to weigh evidence and recommendations from many different sources, and what they seek is analysis that compares alternative possible explanations of what policies, programs, or external events contributed to particular outcomes. What they seek is credible evidence that the program being studied has made a significant contribution to the desired outcomes beyond a reasonable doubt. Strengthening the generation and assessment of credible evidence is one of the important challenges for RWE over the next few years.

2.3. Strengthening Quality Assurance and Threats to Validity Analysis

We have discussed the RWE approach being used to strengthen quality assurance through the use of threats to validity analysis that can be operationalized by using the

threats to validity worksheets described in Chapter 7 and in Appendixes A to E. This is still very much a work in progress, and it is hoped that we and others will work with agencies that are interested in testing out the approach on the evaluations they support, to test and refine the approach.

2.4. Defining Minimum Acceptable Quality Standards for Conducting Evaluations Under Constraints

The RWE approach has now developed sets of guidelines for addressing budget, time, data, political, organizational, and administrative constraints and challenges. A number of observers have pointed out that these approaches do not provide clear guidance for defining minimum acceptable standards. There is a danger that evaluators following this approach could produce evaluations that result in findings and recommendations that are based on "we did the best we could under the circumstances" but are methodologically questionable. While the threats to validity checklists provide general pointers, the RWE approach would benefit from more specific guidelines and examples on how to determine what minimum acceptable standards might be. For example, how do we judge what is the minimum time required in the field to conduct methodologically sound data collection, how do we assess whether data collected through reconstructing baselines are of an acceptable validity, and what is the minimum time and level of financial and professional resources required for the analysis of different kinds of data?

Separate but complementary guidelines will be required for evaluations that use predominantly QUANT, QUAL, and mixed-method approaches.

2.5. Further Refinements to Program Theory

Program theory is a key building block of the RWE approach, and a goal will be to introduce further refinements to strengthen applications to RWE. This will include but not be limited to the following:

- Strengthening contextual analysis, including methods for transforming descriptive assessments into dummy variables or ordinal scales that can be integrated into regression analysis
- Strengthening methodologies for process analysis and for assessing the influence of contextual factors on the efficiency of project implementation
- Using results chain analysis to help generate hypotheses, identify unexpected outcomes, and identify key data to be collected
- Further work on the use and limitations of program theory for causal analysis
- Further work on trajectory analysis to help identify the appropriate time horizon for estimating outcomes and impacts.

2.6. Further Refinements to Mixed-Method Designs

The use of mixed methods is another key component of the RWE approach, and the goal will be to develop and test further refinements with respect to, among others, the following:

- Guidelines for mixed-method sampling to strengthen the representativity of findings from small samples of case studies or in-depth interviews. A key question will be to develop guidelines for estimating the required sample size for the qualitative samples.

- Ways to combine multiple data collection and analysis methods to strengthen the validity of data obtained under budget and time constraints
- Guidelines for strengthening the conceptual framework by integrating QUANT deductive hypotheses with QUAL inductive hypotheses

2.7. Further Work on Sampling to Broaden the Use of Small Samples

We introduced into this edition Lot Quality Acceptance Sampling (LQAS) as one strategy that permits the assessment of many types of program performance with small samples, but we plan to explore other options for obtaining statistically robust estimates of outcomes and impacts through small samples.

2.8. Feedback Is Welcome

In this concluding chapter, we have hinted at some of the ongoing exploratory work that is needed to make the RWE approach even more practical and useful to a broad number of institutions and practitioners. We invite you, as a colleague who has taken the time to read this summary of the book, to share your experiences as you experiment with these and similar approaches. We invite you to contribute to the continued expansion of our networks of colleagues who collectively share in this journey. One way to do so is to join the RealWorld Evaluation listserv by sending an e-message to RealWorldEval-subscribe@yahoogroups.com. Also, note that we periodically add more materials to the www.RealWorldEvaluation.org website.

We look forward to your company on the road ahead!

Note: the 2nd edition of the *RealWorld Evaluation* book was published by Sage in December, 2011. To order your copy of the full book, go to www.sagepub.com/rwe.