AMERICAN EVALUATION ASSOCIATION
PROFESSIONAL DEVELOPMENT WORKSHOP SESSION 21

REALWORLD EVALUATION

WORKING UNDER BUDGET, TIME, DATA AND POLITICAL CONSTRAINTS

OVERVIEW

MICHAEL BAMBERGER
JIM RUGH

NOVEMBER 5, 2008

FOR MORE REALWORLD EVALUATION RESOURCES: www.RealWorldEvaluation.org

Contents

Figures

Tables [*at the end of the text*]

Appendices

BRINGING IT ALL TOGETHER

Applying RealWorld Evaluation Approaches to Each Stage of the Evaluation Process

Introduction to this stand-alone version

In this chapter of the book we discuss how RealWorld Evaluation (**RWE**) approaches can be applied at each stage of the design and implementation of a typical evaluation. We identify RWE issues that can come up—that is, where there are constraints related to funding, time, availability of data, and clients' preconceptions—and suggest how the RWE approach can help to address those constraints. Readers new to the evaluation field might also find this chapter useful as a general introduction to the planning, design, implementation, dissemination, and use of any evaluation.

This chapter is designed to be both an introduction to RWE as well as a useful condensation of many of the main points of the book. Figure 1 summarizes the seven steps of the RWE approach and this overview also includes references to other chapters of the book where more detailed coverage of particular issues can be found[1].

A few of the important tables from other chapters have been included at the end of this stand-alone version of Chapter 16.

1. Scoping the Evaluation

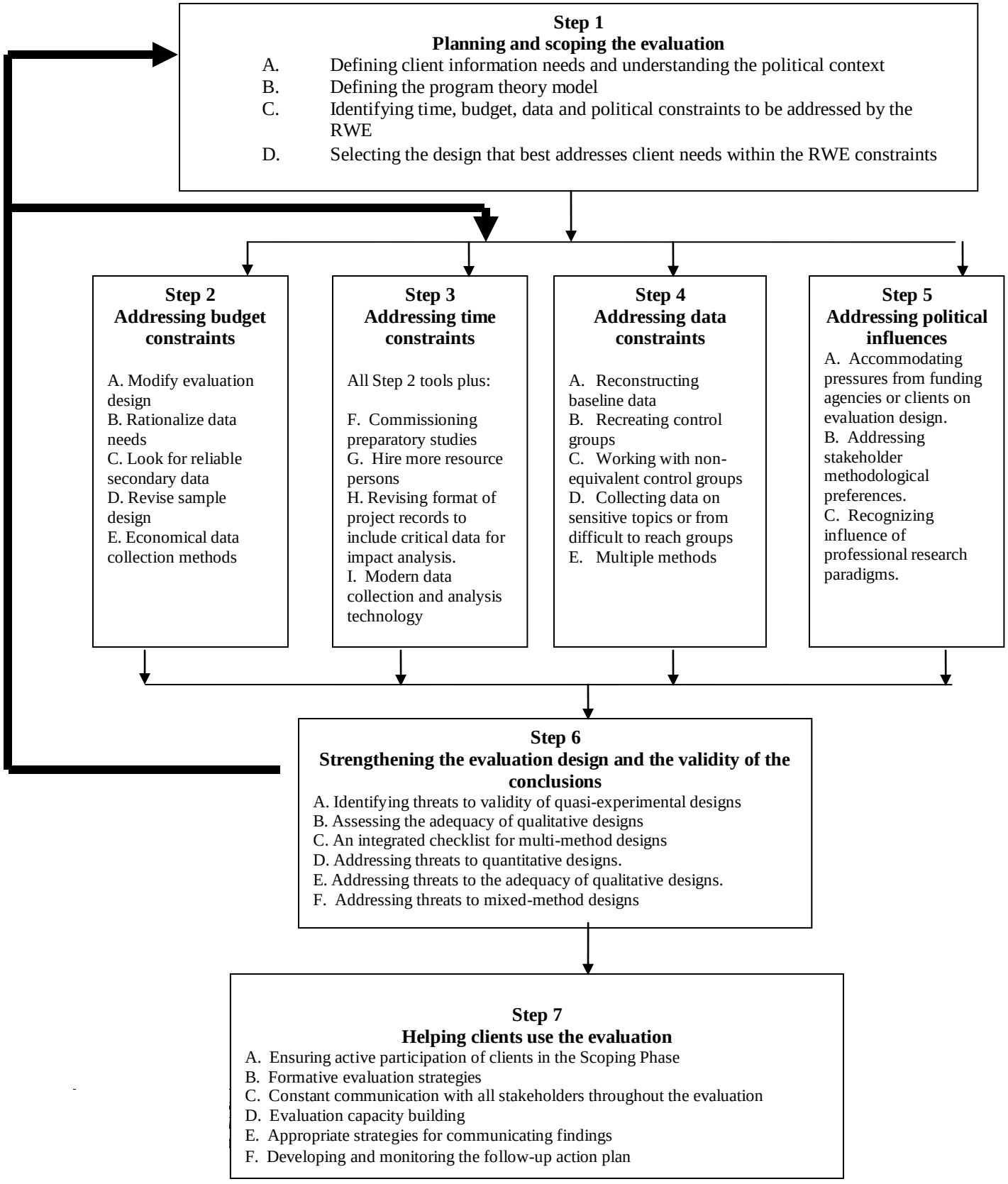
It is important that those charged with conducting an evaluation gain a clear understanding of what those asking for the evaluation (the **clients**[2] and **stakeholders**) are expecting—that is, the political setting within which the **project** and the evaluation will be implemented. It is also important to understand the policy and operational decisions to which the evaluation will contribute and the level of **precision** required in providing the information that will inform those decisions (see Chapter 2).

Understanding Client's Needs

An essential first step in preparing for any evaluation is to obtain a clear understanding of the priorities and information needs of the client (the agency or agencies commissioning the evaluation) and the key stakeholders (persons interested in or affected by the project). The timing, focus, and level of detail of the evaluation should be

[1] References to Chapters refer to *RealWorld Evaluation*
[2] The definitions of words in **bold** can be found in the Glossary.

Figure 1. The RealWorld Evaluation [RWE] Approach



determined by the client information needs and the types of decisions to which the evaluation must contribute.

The process of clarifying what questions need to be answered can help those planning the evaluation to identify ways to eliminate unnecessary data collection and analysis, hence reducing cost and time. The RealWorld evaluator must distinguish between (a) information that is essential to answer the key questions driving the evaluation and (b) additional questions that would be interesting to ask, if there were adequate time and resources, but that may have to be omitted given the limitations faced by the evaluation.

An additional decision relating to cost and time may concern who should be involved in data collection and review of the evaluation reports. Many development projects have a philosophy of promoting the empowerment of community members and other stakeholders (such as school administrators and teachers, health center staff, local government agencies, and non-governmental agencies) which includes inviting their participation in monitoring and evaluation activities. A question that can arise in the face of RWE constraints is, "How important is it to include participatory methods and adequate representation of project participants and other stakeholders in the evaluation?" Participatory data collection methods tend to be more expensive and time-consuming, because sufficient time must be allowed to develop rapport with the community and other stakeholders and to build trust. The cost-conscious RealWorld evaluator must determine whether the client and key stakeholders place a sufficiently high value on participatory approaches to allow for the time and budget required *to do them well.*

Another challenge for RWE is that it is often more time-consuming and expensive to reach the poorest and most vulnerable groups, so when time and budgets are constrained there will often be pressures to drop these groups from the consultations. "It would be really great to consult with the squatters who do not have land title, but unfortunately . . . we just don't have the money and/or the time."

An important function of the scoping phase is to understand whether the lack of consultation with the groups affected by the project, including the poorest and most vulnerable groups, is due to a lack of resources or to the low priority that the client assigns to their involvement. Often, lack of time and money may be used as an excuse, so it is important for the evaluator to fully understand the perspective of the client before deciding what approach to adopt.

Understanding the Political Environment

The political environment includes the priorities and perspectives of the client and other key stakeholders, the dynamics of power and relationships between them and the key players in the project being evaluated, and even the philosophical or methodological biases or preferences of those conducting the evaluation. Table 1 lists some of the ways in which political factors can affect evaluations when they are being designed, while they are being implemented and when the findings are being presented and disseminated (see also Chapter 6).

It is important to avoid the assumption that political influence is bad and that evaluators should be allowed to conduct the evaluation in the way that they know is "best" without interference from politicians and other "narrow-minded" stakeholders

trying to make sure that their concerns are introduced into the evaluation. The whole purpose of evaluation is to contribute to a better understanding of policies and **programs** about which people have strong and, often, opposed views. If an evaluation is not subject to any political pressures or influences, this probably means either that the topic being studied is of no consequence to anyone or that the evaluation is designed in such a way that the concerned groups are not able to express their views. Evaluators should never assume that they are right and that stakeholders who hold different views on the key issues, appropriate methodology, or interpretation of the **findings** are biased, misinformed, or just plain wrong. In the **Standard Checklist for Assessing the Adequacy and Validity of all Evaluation Designs** (Appendix 1), assessment Criteria C has to do with the **internal validity** and **authenticity** of the evaluation findings: "Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?"

If key groups do not find the analysis credible, then the evaluator may need to go back and check carefully on the methodology and underlying assumptions. It is never an appropriate response to sigh and think how difficult it is to get the client to "understand" the findings and recommendations.

One of the dimensions of contextual analysis used in developing the program theory model (see following section) is to examine the influence of political factors. Many of the contextual dimensions—economic, institutional, environmental, and sociocultural—influence the way that politically concerned groups will view the project and its evaluation. A full understanding of these contextual factors is essential to understanding the attitudes of key stakeholders to the program and to its evaluation. Once these concerns are understood, it may become easier to identify ways to address the pressures placed by these stakeholders on the evaluation.

Not surprisingly, many program evaluations are commissioned with political motives in mind. A client may plan to use the evaluation to bolster support for the program and may consequently resist the inclusion of anything but positive findings. On the other hand, the real but undisclosed purpose the client may have had for commissioning the evaluation may be to provide ammunition for firing a manager or closing down a project or a department. Seldom if ever are such purposes made explicit. Different stakeholders may also hold strongly divergent opinions about a program, its execution, its motives, its leaders, and how it is to be evaluated. Persons who are opposed to the evaluation being conducted may be able to preempt an evaluation or obstruct access to data, acceptance of evaluation results, or continuation of an evaluation contract.

Before the evaluation begins, the evaluator should anticipate these different kinds of potential political issues and try to explore them, directly or indirectly, with the client and key stakeholders. Chapter 6, Table 6.1[3]*, illustrates some of the many ways that the political context can affect how an evaluation is designed, implemented, disseminated, and used.

---

[3] Tables from RealWorld Evaluation marked with * are not included in this overview.

Political dimensions include not only clients and other stakeholders. They also include individual evaluators who have preferred approaches that resonate with their personal and professional views as to what constitutes competent, appropriate practice. Different evaluators, even those who have chosen to work together on a project, may take different stances regarding their public and ethical responsibilities. Evaluators, like everyone else, have their own personal values. However, for many evaluators, it may be more comfortable to think of the work of evaluation not as an imposition of the evaluator's values but, rather, as an impartial or objective data-based judgment about program merit, shortcomings, effectiveness, efficiency, and goal achievement. The evaluators must be aware of their own perspectives (and biases) and seek to ensure that these are acknowledged and taken into consideration. (See Section A of the Standard Adequacy and Validity Checklist in Appendix 1.)

Clients may base their selection of evaluators on their reputations for uncompromising honesty, counting on those reputations to ensure the **credibility** and acceptance of findings. Or the choice of evaluator may be based on ideological stances the evaluator has taken that are in agreement with the client's. These decisions may be so understated as to initially go unnoticed in friendly negotiations and enthusiastic statements about the strategic importance of the proposed evaluation. Chapter 6 discusses some of the options available to the evaluator when it is felt that some of the pressures from clients are ethically or professionally unacceptable.

Evaluators should also be alert to the fact that political orientations of clients and stakeholders can influence how evaluation findings are disseminated and used. Clients can sometimes ignore findings they do not like and can suppress distribution by circulating reports only to carefully selected readers, by sharing only abbreviated and softened summaries, and by taking responsibility for presenting reports to boards or funding agencies and then acting on that responsibility in manipulative ways. Clients have been known to give oral presentations and even testimony that distorted evaluation findings, to take follow-up activities not suggested and even contraindicated by evaluation reports, and to discredit evaluations and evaluators who threaten their programs and prestige.

The wise evaluator should be aware of such realities and be prepared to deal with them in appropriate ways. Chapter 6 suggests some RWE strategies for addressing political constraints such as these, as well as others, during the evaluation design, the implementation of the evaluation and in the presentation and use of the evaluation findings.

Defining the Program Theory

Before an evaluation can be conducted, it is necessary to identify the explicit or implicit theory or logic model that underlies the design upon which a project was based (see Chapter 9). An important function of an impact evaluation is to test the hypothesis that the project's interventions and outputs contributed to the desired outcomes, which, along with external factors that the project assumed would prevail, were to have led to sustainable impact.

Defining the program theory or logic model is good practice for any evaluation. It is especially useful in RWE, where, due to budget, time, and other constraints, it is

necessary to prioritize what the evaluation needs to focus on. An initial review of what a project did in light of its logic model could reveal missing data or information that is needed to verify whether the logic was sound and whether the project was able to do what was needed to achieve the desired impact.

If the logic model was clearly articulated in the project plan, it can be used to guide the evaluation. If not, the evaluator needs to construct it based on reviews of project documents and discussions with the project implementing agency, project participants, and other stakeholders (see Chapter 9). In many cases, this requires an iterative process in which the design of the logic model evolves as more is learned during the course of the evaluation.

In addition to articulating the internal cause-effect theory on which a project was designed, a logic model should also identify the socioeconomic characteristics of the affected population groups, as well as contextual factors such as the economic, political, organizational, psychological and environmental conditions that affect the target community.

The key phases or levels of a simple logic model can be summarized as follows (see Figure 16.1* and Chapter 2):

1. ***Design***. How was the project designed? Who designed this project? Was it only a few staff members of the donor or implementing agency, or an external consultant? Or was there extensive participation by a mixture of stakeholders, including the intended beneficiaries? Was the design based on a holistic diagnostic assessment of the conditions in the target communities? And was the design informed by lessons learned from evaluations of previous projects using similar approaches under similar conditions?

2. ***Inputs***[4]***.*** Inputs represent the financial, human, material, technological, and information resources used for the development intervention.

3. ***Implementation process***. This includes actions taken or work performed through which inputs such as funds, technical assistance, and other types of resources are mobilized to produce specific outputs. One of the critical factors is whether, and how, intended beneficiaries and other stakeholders were involved in the implementation process.

4. ***Outputs***. Outputs include the products and services that result from the completion of activities within a development intervention. Note that project implementers have direct control over outputs – although not over the external contextual factors that may affect the timely delivery or quality of the outputs..

5. ***Outcomes***. These are the intended or achieved short-term and medium-term effects of an intervention's outputs. Note that project implementers do not have direct control

[4] Many of these definitions are based on the OECD glossary (Organization for Economic Cooperation and Development 2002).

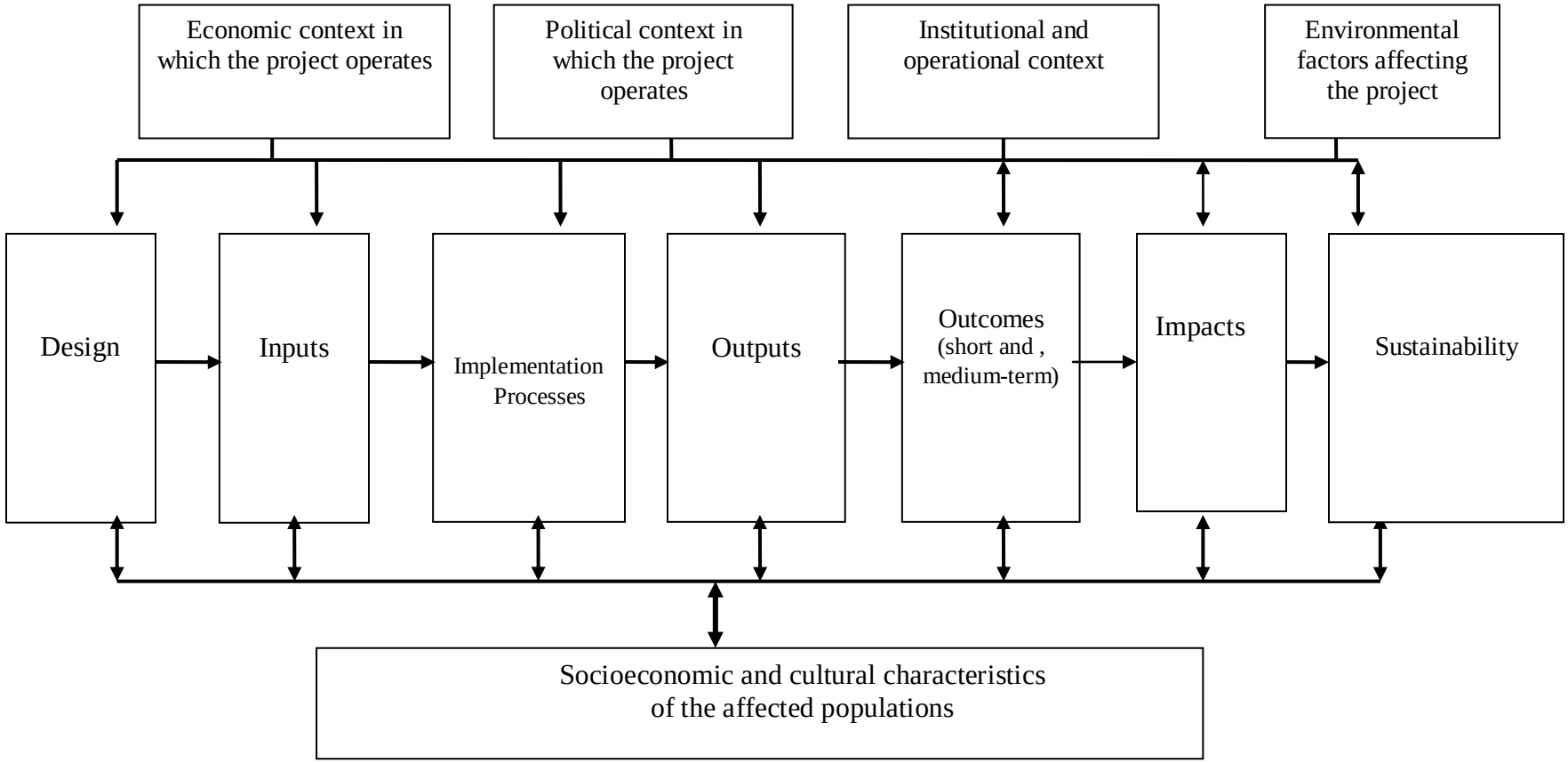
over outcomes. Outcomes are what others do on their own, albeit influenced by the project's outputs.

6. ***Impacts***. Impacts are the positive and negative long-term effects on identifiable population groups produced by a development intervention, directly or indirectly, intended or unintended. These impacts can be economic, sociocultural, institutional, environmental, technological, or of other types.

7. ***Sustainability***. Sustainability refers to the continuation of benefits from a development ***intervention*** after major external assistance has been completed; it is the resilience to risk of net benefit flows over time. Many people either do not think about this at all, or they assume that impacts will be sustained. However, impacts may not be sustained for a number of reasons: The project may receive subsidies that will not continue; external conditions such as weather or political or economic conditions may change and thus threaten the stability of the project's outcomes; the underlying causes of the problem may not have been addressed so that the project interventions will not resolve *these* problems and will not be sustainable; and the project may introduce techniques or technologies that people or organizations are unable to continue implementing on their own.

Every project is designed and implemented within a unique *setting* or *context* that includes local and regional economic, political, institutional, and environmental factors as well as the socio-cultural characteristics of the communities or groups affected by the project. The program theory must incorporate all these factors through a *contextual analysis.* Where a project is implemented in a number of different locations, it will often be the case that performance and outcomes will differ significantly from one site to another because of the different configurations of **contextual variables**.

Figure 2 A Simple Program Theory Model



Customizing Plans for Evaluation

Those commissioning an evaluation need to consider a number of factors that should be included in the terms of reference (TOR). The client, and an evaluator being contracted to undertake this assignment, might find the following set of questions helpful to be sure these factors are taken into consideration as plans are made for conducting an evaluation. The answers to these questions can help to focus on important issues to be addressed by the evaluation, including ways to deal with RWE constraints.

- Who asked for the evaluation? Who are the key stakeholders? Do they have preconceived ideas regarding the purpose for the evaluation and expected findings?
- Who should be involved in planning/implementing the evaluation?
- What are the key questions to be answered?
- Will this be a formative or summative evaluation? Is its purpose primarily for learning and improving, accountability, or a combination of both?
- Will there be a next phase, or will other projects be designed based on the findings of this evaluation?
- What decisions will be made in response to the findings of this evaluation? By whom?
- What is the appropriate level of rigor needed to inform those decisions?
- What is the scope/scale of the evaluation?
- How much time will be needed/available?
- What financial resources are needed/available?
- What evaluation design would be required/is possible under the circumstances?
- Should the evaluation rely mainly on quantitative (QUANT) methods, qualitative (QUAL) methods, or a combination of the two?
- Should participatory methods be used?
- Can/should there be a survey of individuals, households, or other entities?
- Who should be interviewed?
- What sample design and size are required/feasible?
- What form of analysis will best answer the key questions?
- Who are the audiences for the report(s)? How will the findings be communicated to each audience?

2. Choosing the Best Design from the Available Options

Table 2 identifies five evaluation strategies (true experimental designs, randomized field designs, strong non-randomized designs, weaker non-randomized designs and non-experimental designs) and Table 3 summarizes eleven widely used RWE designs that are used to operationalize these strategies. Table 4 assesses the strengths and weaknesses of each design. Design 1 describes a comprehensive longitudinal design that collects data on the project and control groups at the start of the project, during the period of project implementation, at the end of the project and several years later when it is possible to measure impacts and sustainability. This design can be combined with any of Designs 2-6. However, it is not used very frequently due to the cost and time requirements. Designs 2-9 describe statistical impact designs that compare the project group with a matched control group. All of these designs include a counterfactual that, by addressing the question "What would have been the situation if the project had not taken place?" permits a statistical assessment of the proportion of the observed changes in the project population that can be attributed to the effect of the project intervention. These eight designs are ordered in descending order of statistical rigor, starting with Design 2, Randomized Control Trial that is statistically the strongest design, followed by Designs 3-6 that are relatively strong and Designs 7-9 that are weaker although they still include a control group. Designs 3-6 are distinguished by different procedures that are used for matching the project and control groups.

While many writers refer to designs 2-6 as being "strong" or "rigorous" impact evaluation designs, with Design 2, randomized control trials sometimes referred to as the "gold standard"; it is important to understand that the designs are technical only more rigorous with respect to statistical procedures to control for certain kinds of selection bias. These biases derive from how project beneficiaries are selected and how the control group sample is selected. Randomized control trials and strong quasi-experimental designs are powerful ways to control for these selection biases. However, these designs are not necessarily strong than other designs with respect to, among other things, their construct validity, the adequacy of the indicators and the reliability of the information that is collected. In some ways these designs are weaker than other designs, including non-experimental designs. For example, many statistical impact designs do not collect information on the process of project implementation (the "black box" approach), nor do they examine the context within which the project is implemented or the setting within which the interviews are conducted. For this reason we refer to these designs as being "Statistically strong" or "robust, rather than calling them "strong".

Designs 10-11 are non-experimental designs that do not include a control group. While many of these evaluations are not commissioned until late in the project and only involve a few weeks (or even a few days) in the field; some non-experimental designs can continue over a long period of time and involve the collection of extensive and rich qualitative data. Some practitioners of non-experimental use non-statistical approaches for assessing the counterfactual, for example: program theory models, theory of change and concept mapping among others. When designed creatively—using the **mixed-method** designs that draw on all of the available QUANT and QUAL design, data collection, and analysis approaches (see Chapter 14)—Designs 10 and 11 can provide

operationally useful estimates of the extent to which the project contributes to the desired effects. However, given the many threats to the validity of conclusions, it is important to review carefully the limitations on the kinds of conclusions that can be drawn from the analysis. Chapter 10 describes each of the evaluation designs, gives a case study illustrating how the design has been applied in the field, and assesses the major threats to validity and adequacy. Table 10.4* summarizes the conditions under which each design does and does not work well. This should be consulted at an early stage in the evaluation planning to help narrow down the options.

We describe below a number of strategies that can be used to strengthen most RWEs. Evaluators are strongly encouraged to consider using some of these strategies whenever appropriate and feasible. The choice of the most appropriate RWE design is determined by a number of factors, including the following:

- *When did the evaluation begin?* At the start of the project, while the project was being implemented, or after the project had ended?
- *When will the evaluation end?* Will this be a one-time evaluation conducted while the project is being implemented (most commonly for the midterm review), will it end at approximately the same time as the project (end-of-project evaluation and report), or will it continue after the project ends (longitudinal or ex-post evaluation)?
- *What type of comparison will be used?* There are three main options: (a) a randomized design in which individuals, families, groups, or communities are randomly assigned to the project and **control groups**; (b) a **comparison group** selected to match as closely as possible the characteristics of the project group; or (c) no type of control or comparison group.
- *Does the design include process evaluation?* Even if an evaluation is focused on measuring sustainable changes in the conditions of the target population, it needs to identify what most likely led to those changes. That includes an assessment of the quality of a project's implementation process and whether it made a plausible contribution to any measured impact.
- *Are there preferences for the use of QUANT, QUAL, or mixed-methods approaches?* See Chapters 11, 12, and 13.

Paring Down the Evaluation Design

RWE approaches are used because time, resources, the available data, and possibly, the political setting do not permit the use of stronger evaluation designs. Under these circumstances, the evaluator must work with the client to agree on how the resource and time requirements as well as the data needs can be pared down while still ensuring an acceptable level of precision. Chapters 3 and 4 discuss options for working within a tight budget or under time constraints, and Chapter 5 describes ways to make the most effective use of the available data. Table 5 summarizes ways to reduce the costs of data collection and analysis and Table 6 estimates the potential cost savings from using

simplified evaluation designs). The following are recommended steps in defining the best and most acceptable design under given constraints:

- Spend the time needed to fully understand the client's priority information needs and the political and other constraints under which they are operating:
    - What does the client really need from the evaluation?
    - Is it essential to have rigorous QUANT analysis to ensure the credibility of the evaluation, or is an in-depth QUAL analysis more important and credible to clients?
    - What is the nonnegotiable "bottom line" in terms of the minimum information needs and the real deadlines for producing a first draft and a final report?
    - Who are the stakeholders to whom the evaluation is directed and whose opinions are critical?

- Review the options for reducing costs (Chapter 3) and time (Chapter 4) and for strengthening the available database (Chapter 5).
- Use these options to prepare several possible scenarios for achieving the evaluation objectives within the resource constraints. Review the "Standard Checklist for Assessing the Adequacy and Validity of Quantitative, Qualitative, and Mixed-Method Designs" (Appendix 1) and the Checklist for Assessing Threats to Validity of Quantitative Evaluation Designs (Appendix 2) and assess strengths and weaknesses of each option from the perspective of the client(s) and other key stakeholders.
- If none of the available scenarios can satisfy the client's bottom-line, prepare two additional scenarios:
    - Scenario 1: Estimate what would be the additional budget or time requirements to satisfy the bottom-line (e.g., an extra $25,000 would be required to include a comparison group in the evaluation design or the deadline for the submission of the draft report would need to be extended by 3 months).
    - Scenario 2: Indicate what modifications would be required in the evaluation design to stay within the available resources.

Under Scenario 2 it might be possible to produce aggregate estimates of project impact at the national level but not to provide disaggregated estimates of the impact of different combinations of services or the impact at different project sites or on different socioeconomic groups (e.g., men and women, wage earners and the self-employed, different ethnic groups).

Alternatively, or in addition, it might be necessary to lower the statistical confidence level (see Chapter 14) so that statistically significant differences between the project and comparison groups might be assessed at only the 0.10 (10%) confidence level rather than the conventional 0.05 (5%) level. Table 7 lists factors affecting the sample size.

Some of the information might have to be obtained from, for example, **focus group** interviews and **PRA** (the term is now used to refer to a wide range of participatory appraisal methods, although it originally meant **participatory rural appraisal**) group techniques rather than from a household sample survey.

All the options should be discussed with the client as well as the implications of each option in terms of the level of precision, the types of analysis that can be conducted, and the credibility of the findings to different stakeholders. It is essential to ensure that the client fully understands the options and trade-offs before a decision is made on how to proceed. Sometimes the client will ask the evaluator, "What would you advise is the best approach, because you are the expert." If asked this question, the evaluator must explain that this is a policy decision to be made in consultation with the client and that the role of the evaluator is simply to provide advice on the technical implications of each option with respect to precision, types of analysis, and professional credibility.

Identify the Kinds of Analysis and Comparisons that are Critical to the Evaluation

A key factor in the choice of the evaluation design, and for determining the size and structure of the sample, has to do with the kinds of analysis required and the levels of statistical precision needed. It is useful to think of three kinds of evaluation:

1. *Exploratory or research evaluations* in which the purpose is to assess whether the basic project concept and approach "works." This is often used when a new type of service is being piloted or when an existing service is to be provided in a new way or to reach new target groups. Examples of the key evaluation questions include the following:

   a. Are farmers willing to experiment with new kinds of seed?
   b. Do the new teaching methods get a positive response from the schools, students, and parents and is there initial **evidence** of improved performance?
   c. If poor women are given loans, are they able to use the money to start or expand a small business?
   d. Which groups benefit most and least, and why?

2. *Small-scale quasi-experimental or* QUAL *designs* to assess whether there is evidence that the project is producing significant effects on the target population. Some designs include a comparison group, whereas others use a more general comparison with similar communities through PRA techniques or focus groups. Questions of attribution (what would have been the condition of the project group if the project had

not taken place?) are addressed but in a less rigorous way than for large-scale impact assessments (see below). Some of the critical questions might include the following:

a. Are the intended project beneficiaries (e.g. individuals, families, schools, communities) better off as a result of the project?
b. How confident are we that the observed changes were caused by the project and not be external factors such as improvements in the local economy?
c. Would the project be likely to have similar effects in other areas if it were replicated? Where would it work well and less well, and why?
d. Which contextual factors (economic, political, institutional, environmental, and cultural) affect success? (See Chapter 9 for a discussion of contextual factors and mediator variables).
e. Who did and who did not benefit from the project?

3. *Large-scale impact assessment* where the purpose is to assess, with greater statistical rigor, how large an effect (defined numerically in terms of percentage or quantitative change) has been produced, and who does and does not benefit. Ideally, the evaluation should use a mixed-method approach integrating QUANT and QUAL methods. Critical questions might include the following:

a. What QUANT impacts (high-level sustainable effects) has the project produced? The emphasis is on "how much" and not just "what."
b. What is the quality of the services (compared with other programs, to expected standards, and in the opinion of the target groups)?
c. Are the project effects statistically significant "beyond a reasonable doubt"?
d. Who has benefited most and least, and are there any groups that have not benefited at all or who are worse off?
e. What are the **intervening variables** (e.g., socioeconomic characteristics of the project groups, cultural factors affecting participation) that influence the magnitude of impacts?

For *exploratory evaluations*, a descriptive analysis using techniques such as observation, interviews of at least a few selected members of the target population, key informant interviews, and perhaps focus groups would probably suffice. A simple and rapid survey might also be used to collect basic information on the project population. It would probably not be necessary to use a formal comparison group, although similar communities or areas might be visited to assess how similar or different they are to the project areas.

For *small-scale QUANT and QUAL evaluations of outcomes and impacts* it is useful, although not always possible, to identify a comparison group to help estimate what the situation of the project group would have been in the absence of the project (the **counterfactual**). Ideally, a mixed-method approach is used, assessing both quality and quantity of services and impacts. The design is also much stronger if baseline data (in whatever form available) can be obtained. Simple statistical comparisons such as

difference of means or proportions should be made between the project and comparison groups.

For *large-scale impact assessments*, relatively large samples (often requiring hundreds of observations in both project and comparison groups) are required so that multivariate analysis can be used to statistically control for differences between the project and comparison groups and estimate the quantitative influence of the intervening variables. Again a mixed-method approach should be used so that QUANT estimates are complemented by QUAL descriptions of the project context, the process of project implementation, the quality of services, and the opinions and experiences of beneficiaries and agencies and staff responsible for project implementation.

Assessing Threats to Validity and Adequacy of Different Designs

The RWE approach assesses the strengths and weaknesses of the different stages of the evaluation design, to identify factors affecting the validity of the conclusions and recommendations. This is important for any evaluation, but particularly so for RWE where conventional methodological procedures often have to be relaxed due to time or budget constraints or because some of the required data is not available. Several factors affect the adequacy of the evaluation design and findings (see Table 11 and Chapter 7) including the following:

- The appropriateness of the evaluation focus, approach, and methods for obtaining the types of information required
- The availability of data and data sources
- How well the data collected will support interpretations about program performance and impacts
- The qualifications of the evaluation team in terms of both evaluation methodology and the specific fields of the program

For QUANT evaluations, four sets of generally accepted "threats to validity" were established by Cook and Campbell in the 1960s (see Shadish, Cook, and Campbell 2002 for an updated version). There is much less agreement among QUAL researchers as to how to address threats to the adequacy or validity of an evaluation, but writers such as Guba and Lincoln (1989) and Miles and Huberman (1994) have proposed a framework that has been used by a number of authors and that we have followed.

Annex 1 presents a Threats to Validity Checklist that combines seven dimensions proposed by QUANT and QUAL evaluators. These are:

- *A. Objectivity*. Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?
- *B. Reliability*. Is the process of the study consistent, coherent and reasonably stable over time and across researchers and methods?
- *C. Internal validity*. Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying? Are there reasons why the assumed causal relationship between two variables (e.g. project treatment and

outcomes or impacts) may not be valid? Many internal validity problems arise from the way in which project participants were selected (they are different in important ways from the comparison group), because the characteristics of the project group changes over the life of the project because of people dropping out, or because experiences during the project influence the way people respond.

- *D. Statistical conclusion validity*. Incorrect inferences/conclusions about the effects of project interventions on the intended outcomes and impacts: The problems may result from the incorrect application of a statistical test or because of limitations of the sample (e.g., the sample may not cover all the population). Another common problem is that because of the small sample size (usually due to budget or time constraints), the analysis may conclude that the project did not have a statistically significant effect, when in fact the sample was too small to have found this effect, even if it did exist. There is often a trade-off between reducing sample size to cut costs and the need to ensure the sample is large enough to find the effects if they do exist. Also, many well-designed and -executed projects can expect to produce only small effects, which makes it more difficult to detect the effects. The issues of Type II errors (wrongly concluding there was no effect), the **power of the test**, **effect size,** and the estimation of sample size are discussed in Chapter 14.
- *E. Construct validity* the degree to which inferences can legitimately be made from the theoretical constructs (definition of key concepts) on which the program theory is based: Many of the key constructs are difficult to define (e.g., poverty, vulnerability, well-being, hostile work environment) and even harder to measure. A lack of precision and clarity in the definition and measurement of these key constructs will undermine the ability of the evaluation to understand and interpret how the project has operated and what it has achieved.
- *F. External validity.* Incorrect inferences about whether evaluation findings would apply to different persons, times, or settings: Most QUANT evaluations are intended to determine the extent to which the evaluation findings can be generalized to a broader population (e.g., all low-income communities, all unskilled women workers, all secondary schoolchildren). There are a number of ways in which the characteristics of the project population may not be typical of the broader population so that generalization of findings may be misleading. For example, an adult literacy program may have been successful, at least in part, because of the enthusiastic support of the local chamber of commerce, which provided transport, free exercise books, and snacks for participants. Consequently, the program's success might not justify the conclusion that it would be similarly successful in other cities where it did not enjoy this strong local support.
- *G. Utilization*. How useful were the findings to clients, researchers and the communities studied?

The threats to validity checklist can be used to identify and assess potential weaknesses in all of the designs described in Table 3. Readers who are not specialists in statistical analysis and QUANT evaluation may find some of the elements of components C,D,E and F difficult to follow. While it is worth glancing through the categories to get an idea of the wide range of factors that can affect the validity of QUANT evaluation

designs, it is always possible to ask the advice of a specialist when rigorous QUANT evaluations must be designed and assessed.

Appendix 2 presents a simplified threats to validity checklist that can be used to assess the validity generated through the different techniques for reconstructing baseline data discussed in Chapter 4. For non-specialist readers, this simplified checklist may provide sufficient guidance for assessing the validity of most evaluations.

Table 12 identifies some of the threats to validity that can affect even the strongest evaluation designs.

When and How to Use the Threats to Validity and Adequacy Checklist

The checklist should be referred to throughout the evaluation to ensure that things are on the right track and to rapidly identify and address problems affecting the validity of the methodology and the conclusions. The most important times to use the checklist are these:

- *When the evaluation is being designed and data collection and analysis methods are being planned.* The checklist can be used to identify potential problems and to consider ways to address them. It is particularly useful in RWE for identifying the potential threats to validity resulting from the methods proposed to reduce costs and time and to work with limited databases. For example, what additional threats to validity arise if the baseline comparison group is cut or if income data are collected through focus groups rather than through household sample surveys?
- *When most of the data collection has been completed.* The checklist can help identify any potential threats that have arisen during data collection (e.g., unexpectedly high nonresponse rates or confusion about the concept of unemployment). It should be applied as soon as possible after data collection is completed (or even when it is still underway) so that there is still time to take corrective measures.
- *When the draft evaluation report has been completed.* Ideally, there may still be time to take some corrective measures, but if this is no longer possible, the evaluation report should use the checklist to identify and clearly state the potential threats to validity and how these might affect the conclusions and recommendations.

Appendix 2* of RealWorld Evaluation (Not included in this overview) presents an "RWE Project Worksheet for Identifying and Addressing Threats to Validity and Adequacy" that can be used at any stage of the evaluation to identify threats to validity and adequacy, to assess the seriousness of each threat for the purposes of the evaluation, and to identify actions that can be taken to correct or at least address the most important

threats. Appendix 3* gives a worked example illustrating how the worksheet could be applied to the assessment of an evaluation of a housing project.

3. Determining the Appropriate Methodology

QUANT and QUAL approaches and methods are, by and large, designed for different purposes. Chapter 11 and 12 discuss the strengths and potential weaknesses of QUANT and QUAL approaches, respectively. Recognizing these strengths and weaknesses, the RWE approach considers that in most situations the strongest and most robust evaluation design will probably combine both QUANT and QUAL approaches. Chapter 13 is devoted to a detailed discussion of how mixed methods, systematically combining QUANT and QUAL approaches, can be useful in RWEs. It is argued that mixed-method approaches are particularly valuable for RWEs because the combination of different data collection and analysis methods can detect and overcome, or at least reduce, some of the threats to validity resulting from the compromises that have to be made in the light of budget, time, and data constraints. For example, reducing sample size increases the sampling error, making it more difficult to detect significant differences. If PRA and other QUAL techniques are used to obtain independent estimates of project impacts and if the findings are consistent with the statistical analysis of the sample surveys, then confidence in the findings may be increased.

Many authors argue that when mixed-method approaches are used, every evaluation has either a predominant QUANT or QUAL focus ("theoretical drive") and the other approach is used to complement this. The theoretical drive will be determined by the professional orientation of the researcher or the preference of the client. However, other authors propose an integrated approach that does not give primacy to either approach. Whichever position is taken on this issue, some programs lend themselves more naturally to QUANT evaluation methods (e.g., very large programs affecting many thousands of people), whereas in other cases QUAL methods may seem better suited (e.g., a program whose goal is to improve the quality of teaching practices or one that is trying to introduce vegetable growing into several fairly small villages). However, in all cases, the choice of methods will also be influenced by the preferences of the evaluators and the clients.

Throughout this book, we strongly urge evaluators to select the data collection and analysis tools best suited to the needs of the client and the nature of the program being evaluated and to avoid selecting methods simply because they are qualitative or because they are quantitative.

4. Ways to Strengthen RealWorld Evaluation Designs

Because of the contexts within which RWEs are implemented, the conventional quasi-experimental QUANT designs, when used in isolation, are subject to a number of threats to validity that are likely to weaken the quality of the data and the validity of the conclusions. (This is why RWE strongly encourages the use of mixed-method designs.)

These issues are addressed in more detail in Chapter 10. The limitations of the conventional QUANT designs include the following:

- Problems can arise concerning the reliability of measurement of key indicators, particularly when these relate to sensitive issues such as illegal drug use, control of resources, domestic violence, and social constraints on women's economic activities or mobility.
- There can be difficulties in capturing variations in project implementation, the quality of services, and the access of different groups to the services and benefits.
- Conventional designs do not analyze contextual factors affecting the outcomes and impacts of the project in different locations.
- Important differences (nonequivalency) between the project and comparison groups, particularly those that are difficult to quantify (e.g., motivation, community organization) are difficult to capture.

The following are a number of procedures that can be used to strengthen these designs (see also Box 7.2* Chapter 7 for examples of ways to address common threats to statistical, internal, construct and external validity). Evaluators should consider incorporating some of the following procedures into RWE designs where appropriate.

Basing the Evaluation Design on a Program Theory Model

As we saw earlier in this chapter, the formulation of a program theory model helps identify the key issues and hypotheses on which the limited evaluation resources should focus (see Chapters 2 and 9). A theory model can also be used to complement a quasi-experimental design by describing the project implementation process and analyzing the contextual factors that affect implementation and outcomes, and it helps interpret the evaluation findings and the assessment of whether a project should continue or be replicated.

Complementing the Quantitative (Summative) Evaluation with Process Evaluation

A **process evaluation** uses QUANT, QUAL, and mixed methods to observe and assess the process of project implementation and to make recommendations for ways to strengthen subsequent phases of an ongoing program. It addresses questions such as these:

- How were the different components of the project implemented, and how closely did implementation on the ground conform to the project plan or operational manual?
- For ongoing projects, how could the quality of the services be assessed and be improved? Is there evidence that they are leading to desired outcomes?
- Who has access to and/or uses the services and who does not? Why do certain groups not use the services?

- Was the design and organization of the project participatory, managed by a small group or top-down? Who is involved in decision making during implementation?
- What proportion of the community (intended beneficiaries) know about the project? Is their information correct? What do they think about the project?
- What are the relations between the project organizers and the community?
- What do governmental and other organizations know about this project, and what impressions do they have of the quality of services and effectiveness of this project?

Incorporating Contextual Analysis

Contextual analysis assesses the influence of economic, political, organizational and environmental factors on the implementation and outcome of projects. These are defined in program theory models as **mediators** (see Chapter 9). It also examines the influence of the preexisting sociocultural characteristics of the target populations on how different groups respond to the project.

Most contextual analysis is qualitative (e.g., interviews with key informants, review of project documents, and **participant observation**), but it may also include QUANT analysis of data from household surveys. **Contextual variables** can also be transformed into numerical variables (e.g., "**dummy variables**") and incorporated into multivariate analysis (see Chapter 13, note 1).

Reconstructing Baseline Conditions

Many evaluations do not begin until the project has been underway for some time or perhaps is even nearing completion. It is very common under these circumstances to find that no baseline data have been collected at the beginning of the project. This is most commonly the case for the comparison group, but it is also often true for the project group. The absence of baseline data is usually one of the most serious threats to validity, and therefore, RWE proposes a number of different ways to *reconstruct* baseline data. Some of these approaches, described in Table 8 (see also Chapter 5), include the following:

- Using secondary data (see following section)
- Using individual recall (respondents are asked to recall the situation of their family or community at around the time the project began)
- Using PRA and other participatory techniques to *reconstruct* the history of the community and to assess the changes that have been produced by the project
- Interviews with key informants, preferably persons who know the target community, as well as other communities, and therefore have a perspective on relative changes occurring over time

Although all these methods provide potentially valuable information, there are significant threats to validity inherent in any recall method. These result from lack of precise memory, the tendency to confuse the precise time period (so that events that took place earlier may be reported as having occurred since the project began or vice versa), and in some cases, deliberate distortion. Consequently, it is important to treat all recall data with caution and to always use mixed-method approaches to triangulate independent estimates of the reported information from different sources.

Using Secondary Data

Administrative Planning and Monitoring Data Collected by the Organization Being Evaluated. Data collected as part of the pre-project diagnostic assessment, and data collected by the project implementers' monitoring system during the life of a project are important but often underutilized sources for reconstructing baseline conditions (see Chapter 5). Most projects collect a lot of data for administrative and monitoring purposes, and frequently these records contain information that can be useful for reconstructing information on the conditions of the project population at the time the project started (baseline data). Some of the kinds of data typically collected include the following:

- Planning and feasibility studies before the project began
- Socioeconomic characteristics of individuals or families who apply for or receive services
- Attendance at community meetings and, possibly, reports on the meetings (e.g., minutes)
- Activity reports by the agency staff or others involved in the implementation of the project. At a minimum these provide information on the process; ideally, they also mention changes observed in clients' knowledge, attitude, and practices.

Although these kinds of administrative data can be extremely useful as surrogate baseline data, it is important to be aware that the data were not collected for the purpose of evaluation and they may have some limitations (such as being incomplete or poorly kept or not including all the information required for the purposes of evaluation). The second section of Chapter 5 lists some of the questions that should be asked when assessing the quality and utility of this information for the reconstruction of baseline data.

Records from outside the organization. Records from other programs or projects in the same area can often provide information on conditions before the current project began. For example, surveys are often conducted to estimate the number of children not attending school, sources and costs of water supply, or availability of microcredit. More general statistical data may also be available on, for example, school enrollment rates, infant mortality, agricultural prices, microcredit lending, and transport patterns. It is important to assess the strengths and weaknesses of these records with respect to the following:

- Time differences between the start of the project (when data are required for the baseline) and the time when the secondary data were collected. Time differences are particularly critical when general economic conditions may have changed between the survey date and the project launch.
- Differences in the population covered. For example, did the surveys include employment in the informal as well as the formal sectors? Did it cover pedestrian as well as vehicular means of transport?
- Was information collected on key project variables and potential impacts? Are the secondary data statistically valid for the particular target population addressed by the project being evaluated?
- Does information cover both men and women? Or was all information obtained from a single person (usually the "household head") and aggregated for all household members (see Box 5.2)?

The Use of Mixed-Method Approaches to Strengthen Validity of Indicators and to Improve Interpretation of Findings

All of the seven evaluation designs described in Chapter 10 can be strengthened using mixed-method designs (see Chapter 13) that combine QUANT and QUAL approaches in one of the following ways:

- Exploratory studies to understand the context and to identify key issues and hypotheses to be tested. This is particularly important in the construction of the program theory model (see Chapter 9).
- Analysis of the quality of the services provided by the project
- Analysis of the accessibility of the project to different sectors of the target population
- Analysis of the contextual factors (the economic, political, organizational, and natural environmental conditions) within which each project site operates
- Understanding the cultural characteristics of the affected populations and how these influence project implementation and outcomes
- Using **triangulation** to provide two or more independent estimates of key process and outcome indicators (see below).

Estimates are always stronger if they can be independently confirmed from two or more independent sources. This can be done by the following:

- Using independent estimates of change in impact variables obtained from surveys through the use of observation, focus groups, and secondary data. Compare the estimates through triangulation. If estimates from different sources are consistent, greater confidence can be given to the findings.

- If estimates are inconsistent, there should be a follow-up strategy to determine reasons and make adjustments to estimates.

Chapter 10 illustrates the use of triangulation to compare three independent estimates (survey, observation, and key informants) of household income. In one example, the three estimates are *converging* (consistent), whereas in the second example, the estimates are *diverging* (inconsistent). In this second case, it is necessary to follow up to determine the reason for the inconsistencies and to decide how to select the most credible value (or values) to use in the analysis. Ideally, the evaluation design should allow time and resources to return to the field to follow up, either during the interview supervision phase or during the analysis phase (when the discrepancies tend to be discovered). Unfortunately, this is usually not possible, particularly on the RWE budget, so other options should be considered:

- **Enumerators** should be instructed to note inconsistencies between reported information and their direct observation. They should indicate how they interpret the discrepancies and possibly what they think is the best estimate. They may also be instructed to ask some follow-up questions (see Chapter 11).
- Inconsistencies should be identified during the interview supervision phase and follow-up through post-interview discussion with interviewers and possibly by revisits to a sample of respondents.
- Rules for survey instrument coding and analysis of how to address inconsistencies should be defined (e.g., should more weight be given to one source of data, should QUANT estimates be adjusted in a certain defined way?).
- For critical variables, it is possible to create two different indicators giving upper and lower estimates (e.g., income, school enrollment, unemployment), one in which the survey information is not adjusted and the other in which it is. Both estimates will be presented separately in the analysis.

Table 13 summarizes the characteristics of quantitative and qualitative approaches to different stages of the evaluation process and Table 14 describes the elements of an integrated, multidisciplinary research design.

Adjusting for Differences between the Project and Comparison Groups

Multivariate Analysis

When large sample surveys are conducted, multivariate analysis is often used to statistically control for differences between the project and comparison groups to improve the estimates of project impact (see Chapters 10 and 11). The analysis statistically matches subjects (e.g., individuals, households) on variables such as age, education, and income and determines whether there is still a significant difference

between the project and comparison groups on the impact variable (e.g., proportion of children attending school, number of adults unemployed). If there is still a significant difference after this statistical matching, this gives greater confidence that the project is really contributing to the difference. However, if there is no longer a difference, this suggests that school attendance or unemployment may be determined more by household characteristics than by participation in the project.

Using QUAL Methods to Analyze Differences between Project and Comparison Groups

Key informants, focus groups, and participant observation are some of the methods that can be used to identify potentially important differences between project and comparison groups and to assess how they might affect project outcomes and the estimation of project impacts. (See Chapter 5 on collecting data from difficult-to-reach groups and rapid methods for comparing project and comparison groups.)

5. Staffing the Evaluation Economically

In this section, we address issues concerning external experts (either from another country or from a different part of the country), content area specialists, and locally available data collectors. The ideal is to compose an evaluation team that includes a good combination of persons with different experiences, skill sets, and perspectives. Where RWE constraints are faced, especially funding, compromises may have to be made in the composition of the evaluation team. Although we address each of these categories of persons separately, it is important to consider the overall combination and the effectiveness of the full evaluation team in meeting the requirements of an evaluation.

Use External Consultants Wisely

External consultants are usually contracted: (a) because of lack of local technical expertise (inside the organization or in the local research community), (b) to build up local capacity, (c) to save time, (d) to ensure independence and objectivity, (e) to ensure expert credibility, and/or (f) because of a requirement by the funding agency. While, if well selected and used, external consultants can significantly improve the quality of the present and future evaluations, they are also expensive and sometimes disruptive, so they should be selected and used wisely. Under RWE constraints, the goal should be to limit the use of external consultants to those areas where they are essential. Although many of the following points refer to the use of national or international evaluation consultants in developing countries, the same general principles apply in developed countries. For example, the cost and time implications of a consultant flying from Washington D.C. to work on an evaluation on the West Coast of the United States (a distance of almost 3,000 miles) are similar to someone flying from England to Nigeria. Similarly, there are many situations in the United States where English is not the first language in the project areas, so language and cultural competence and sensitivity are similarly important.

Here are a few general rules for selecting and using consultants:

- Ensure that local agencies and the client are actively involved in defining the requirements for the external consultant and in the selection process.
- Consider carefully the merits of an international as opposed to a national consultant. There is often a trade-off between greater technical expertise of the international consultant and the local knowledge (and of course language ability) of the national consultant. Not using any national consultants can also antagonize the local professional community who may be reluctant to cooperate with the international expert.
- If an international consultant is used, give priority to candidates who have experience in the particular country and with local language skills (if required).
- For evaluations with an operational focus, avoid selecting consultants with impressive academic credentials but limited field experience in conducting program evaluations. The requirements are different than for an academically oriented research project.

External consultants are often not used in the most cost-effective way, either because they are doing many things that could be done as well or better by local staff, or because they are brought in at the wrong time. Here are some suggestions on ways to ensure the effective use of external consultants:

- Define carefully what the consultant is being asked to do and consider whether all these activities are necessary.
- Even when the budget is tight, try to plan sufficient time for the consultant to become familiar with the organization, the project, and settings in which it is being implemented. A consultant who does not understand the project, has not spent some time in the communities, or has not built up rapport with project staff, clients, and other stakeholders will be of very little use.
- Plan carefully at what points the consultant should be involved and coordinate ahead of time to ensure that he or she will be available when required. Get tough with consultants who wish to change the timing, particularly at short notice, to suit their own convenience. Some of the critical times to involve a consultant are these:
    - During the scoping phase when critical decisions are being made on objectives, design, and data collection methods and when agreement is being reached with the client on options for addressing time, budget, and data constraints
    - When decisions are being made on sample size and design
    - When the results of the initial round of data collection are being reviewed

  - When the draft evaluation report is being prepared
  - When the findings of the evaluation are being presented to the different stakeholders

- Arrange for a briefing document (preparatory study) to be prepared, by agency staff or local consultants, before the external consultant starts work. This should summarize important information about the project (including compilation of key documents), key partner agencies, and the settings where the project is located. The document, which should be prepared in coordination with the consultant (for example through an exchange of e-mail or phone calls), might also include rapid diagnostic studies in a few communities. A well-prepared document of this kind can save a great deal of time for the consultant and can initiate dialogue on key issues and priorities among clients, local researchers and stakeholders before the consultant even arrives.
- Consider the use of video or phone conferences so that the consultant can maintain more frequent contact with others involved in planning and implementing the evaluation. This enables the consultant to contribute at critical stages of the evaluation without having to always be physically present. In this way, the consultant can make suggestions about the sample or other stages of the design at a sufficiently early stage for it to be possible to make changes based on these recommendations. Video and phone conferences also have the advantage of flexibility, thus avoiding the extremely costly situation where, for example, a consultant flies from Europe to West Africa to participate in the project design phase, only to discover that everything has been delayed for several weeks.

Think about Content Area Specialists

In addition to expertise in the relevant evaluation areas (e.g., qualitative interviewing, questionnaire construction, sample design, and data analysis), it is also essential to include at least one team member with the necessary experience in the content area of the evaluation (e.g., agricultural extension, secondary education, microcredit). Ideally, if resources permit, the team should include both a sector expert with experience in many different countries or programs as well as someone with local knowledge The school or health system in Chicago or Dhaka will probably have many unique features (cultural, organizational, and political) that it is important to incorporate into the evaluation.

Chapter 14 makes the same point with respect to sample design and statistical analysis. The effectiveness of sample design or the application of the appropriate statistical tests is often compromised because the statistician is not familiar with the literature on practices in a particular area (such as educational testing or a branch of psychological research).

Be Creative about Data Collectors

Creative options are sometimes available for reducing the cost of contracting data collectors. In a health evaluation, it may be possible to contract student nurses; in an agricultural evaluation, to contract agricultural extension workers; and for many types of evaluation, to contract graduate students as interviewers or enumerators. Arrangements can often be made with the teaching hospital, the ministry of agriculture, or a university professor to contract students or staff at a rate of pay that is satisfactory to them but well below the market rate. Although these options can be attractive in terms of potential cost savings or for the opportunity to develop local evaluation capacity, there are obvious dangers from the perspective of quality. The interviewers may not take the assignment very seriously; it may be politically difficult to select only the most promising interviewers or to take action against people producing poor-quality work. Supervision and training costs may also be high, and the time required to complete data collection may increase. However, experience shows that these kinds of cooperation can work very well if there is a serious commitment on the part of the agency or university faculty.

Another creative option is to employ data collectors from the community. Sometimes a local high school can conduct a community needs assessment study, or a community organization can conduct baseline studies or monitor project progress. A number of self-reporting techniques can also be used. For example, individuals or families can keep diaries of income and expenditures, daily time use, or time, mode, and destination of travel. Community groups can be given cameras, tape recorders, or video cameras and asked to make recordings on issues such as problems facing young people, community needs, or the state of community infrastructure. Although all these techniques pose potential validity questions, they are valuable ways to understand the perspective of the community on the issues being studied.

6. Collect Data Efficiently

Simplifying the Plans to Collect Data

Data collection tends to be one of the most expensive and time-consuming items in an evaluation. Consequently, any efforts to reduce costs or time will almost inevitably involve simplifying plans for data collection. This involves three main approaches:

1. Discuss with the client what information is really required for the evaluation and eliminate other information in the TOR or mentioned in subsequent discussions that is not essential in answering the key questions driving this evaluation.
2. Review data collection instruments to eliminate unnecessary information. Data collection instruments tend to grow in length as different people suggest additional items that it would be "interesting" to include, even though not directly related to the purpose of the evaluation.
3. Streamline the process of data collection to reduce costs and time. Table 3 (see also Chapter 3) summarizes a number of different strategies to reduce the costs of data collection. These include the following:

- Simplifying the evaluation design (e.g., eliminating the collection of baseline data or cutting out the comparison group)
- Clarifying client information needs (discussed above)
- Look for reliable secondary data (discussed above)
- Reducing sample size (see Chapter 14)
- Reducing the costs of data collection, input, and analysis (e.g., use of self-administered questionnaires, using direct observation instead of surveys, using focus groups and community fora instead of household surveys, and finding cheaper data collectors)

Table 8 lists ways to reduce the time required for data collection and analysis and Table 9 describes rapid data collection techniques

Commission Preparatory Studies

It is sometimes possible to achieve considerable cost and time savings by commissioning an agency staff person or local consultant to prepare a preparatory study. This can cover these points:

- A description of the different components of the project being evaluated and how they are organized
- Basic information on the implementing agency
- Rapid diagnostic studies of the project communities and possible comparison communities
- Information on government agencies, **NGOs** and other organizations involved in or familiar with the project
- Recommendations on community leaders and other key informants with whom the international consultant should meet and preparation of background information on them

Look for Reliable Secondary Data

A great deal of time and expense can be saved if reliable and relevant secondary data can be obtained. Depending on the country and subjects, it may be possible to find records maintained by government statistical agencies or planning departments, university or other research organizations, schools, commercial banks or credit programs, mass media, and many sectors of civil society. Indeed, the evaluator should make use of any relevant records such as monitoring data and annual reports produced by the implementing agency itself. These records may be produced for planning purposes, administrative and financial control, assessing progress, or communicating with the different groups whose authorization, financial support, or general approval are critical to the success of the organization and its activities. Some of the important points to check when assessing the strengths and weaknesses of various kinds of secondary data were discussed earlier in this chapter

Caution: never accept secondary data at face value without checking its reliability.

While many problems with secondary data concern differences in the time period covered, inadequate coverage of some sectors of the target population, or poor quality of data collection and reporting, there are sometimes more fundamental weaknesses in the data. When the quality of supervision is poor, a significant number of surveys may have been completely falsified by the interviewers or important sections may have been left out or poorly recorded.

Collect Only the Necessary Data

It is important to ensure that only essential information is collected. The collection of unnecessary information increases costs and time and also reduces the quality of the information required because respondents become tired if they have to answer large numbers of questions. Therefore, we recommend that all data collection instruments be carefully scrutinized to cut out information that is not relevant and essential to the purpose of the evaluation and that will never be analyzed or used.

Similarly, the data analysis plan should be reviewed to determine what kinds of disaggregated data analysis are actually required. If it is found that certain kinds of proposed disaggregation are not needed (e.g., comparing the impacts of the project on participants in different locations), then it will often be possible to reduce the size of the sample.

Find Simple Ways to Collect Data on Sensitive Topics and from Difficult-to-Reach Populations

Another challenge to evaluators, although not unique to RWE, regards the collection of data on sensitive topics such as domestic violence, contraceptive usage, or teenage violence; or from difficult to reach groups such as commercial sex workers, drug users, ethnic minorities, migrants, the homeless, or, in some cultures, women. A number of methods can help to address such topics and reach such groups. However, RWE constraints such as budget, time, or political prejudices could create pressures to ignore these sensitive topics or leave out groups of people who are difficult to reach. There are at least three strategies for addressing sensitive topics:

1. Identify a wide range of informants who can provide different perspectives.
2. Select a number of culturally appropriate strategies for studying sensitive topics.
3. Systematically triangulate.

Some of the culturally appropriate methodologies that can be used include the following:

- Participant observation
- Non-participant observation (observation of persons or groups as an outsider without being involved in their activities)
- Focus groups
- Case studies

- Key informants
- PRA techniques

Difficult-to-reach groups include commercial sex workers, drug or alcohol users, criminals, informal and unregistered small businesses, squatters and illegal residents, ethnic or religious minorities, boyfriends or absent fathers, indentured laborers and slaves, informal water sellers, girls attending boys' schools, migrant workers, and persons with HIV/AIDS, particularly those who have not been tested.

The evaluator may face one of two scenarios. In the first scenario, the groups may be known to exist, but members are difficult to find and reach. In the second scenario, the clients and, at least initially, the evaluator may not even be aware of the existence of such marginalized or "invisible" groups. The techniques for identifying and studying difficult-to-reach groups are similar to those used for addressing sensitive topics and include the following:

- *Participant observation*. This is one of the most common ways to become familiar with and accepted into the milieu where the groups operate or are believed to operate. Often, initial contacts or introductions will be made through friends, family, clients, or in some cases, the official organizations with whom the groups interact.
- *Key informants.* Schedule interviews with persons who are particularly familiar with and well informed about the target groups.
- *Tracer studies.* Neighbors, relatives, friends, work colleagues, and so on are used to help locate people who have moved.
- *Snowball samples*. With this technique, efforts are made to locate a few members of the difficult-to-locate group by whatever means are available. These members are then asked to identify other members of the group so that if the approach is successful, the size of the sample will increase. This technique is often used in the study of sexually transmitted diseases.
- *Sociometric techniques.* Respondents are asked to identify to whom they go for advice or help on particular topics (e.g., advice on family planning, traditional medicine, or for the purchase of illegal substances). A sociometric map is then drawn with arrows linking informants to the opinion leaders, informants, or resource persons.

7. Analyze the Data Efficiently

Look for Ways to Manage the Data Efficiently

Before data can be analyzed, they must be input into an electronic or manual format. If this is not done properly, the quality and reliability of the data can be compromised or time, money, or both can be wasted. Furthermore, if data are not properly managed, there is the risk that significant amounts of information will be lost.

The following are some of the main steps in the development and implementation of an analysis plan:

- *Drafting an analysis plan* (see Table 11.4*, Chapter 11). This must specify for each proposed type of analysis, the objectives of the analysis, the hypothesis to be tested, the variables included in the analysis, and the types of analysis to be conducted.
- *Developing and testing the codebook.* If there are open-ended questions, the responses must be reviewed to define the categories that will be used. If any of the numerical data have been classified into categories ("More than once a week," "Once a week," etc.), the responses should be reviewed to identify any problems or inconsistencies.
- *Ensuring reliable coding.* This involves both ensuring that the codebook is comprehensive and logically consistent and also monitoring the data-coding process to ensure accuracy and consistency between coders.
- *Reviewing surveys for missing data and deciding how to treat missing data* (see Chapter 14) In some cases, it will be possible to return to the field or mail the questionnaires back to respondents, but in most cases, this will not be practical. Missing data are often not random, so the treatment of these cases is important to avoid bias. For example, there may be differences between sexes, age, and economic or education groups in their willingness to respond to certain questions. There may also be differences between ethnic or religious groups or between landowners and squatters. One of the first steps in the analysis should be to prepare frequency distributions of missing data for key variables and, when necessary, to conduct an exploratory analysis to determine whether there are significant differences in missing data rates for the key population groups mentioned above.
- *Entering the data into the computer or manual data analysis system.*

- *Cleaning the data.* This involves the following:

    - Doing exploratory data analysis to identify missing data and to identify potential problems such as outliers. These are survey variables where a few scores on a particular variable fall far above or below the normal range. A few outliers can seriously affect the analysis by making it much more difficult to find statistically significant results (because the standard deviation is dramatically increased). Consequently, the data cleaning process must include clear rules on how to treat outliers (see Chapter 11).
    - Deciding how to treat missing data and the application of the policies
    - Identifying any variables that may require recoding

- *Providing full documentation of how data were cleaned, how missing data were treated and how any indices were created.*[2]

While RWE follows most of the standard data analysis procedures, a number of approaches may be required when time or budget are constraints. When *time* is the main constraint and where additional resources may be available to speed up the process, the following approaches can be considered:

- Direct inputting of survey data into handheld computers
- Use of electronic scanning to read questionnaires
- Subcontracting data analysis to a university or commercial research organization
- Hiring more, or more experienced, data coders and analysts

When *money* is the main constraint, one or more of the following options can be considered:

- Limiting the kinds of statistical analysis to reduce expensive computer time
- Consider acquiring and using popular statistical packages such as SPSS or SAS so that the analysis can be conducted in-house rather than subcontracting. Needless to say .this option requires the availability of statistical expertise in-house.

Focus Analysis on Answering Key Questions

It is sage advice for any evaluation to focus on the key questions that relate to the main purpose of undertaking an assessment. This is especially important for RWE, because choices need to be made as to what can be dropped as a consequence of limitations of time and funding. By being reminded of what the major questions are and what is required to adequately answer them, those planning a RWE can be sure to focus on those issues and not others. Typically, the clients and stakeholders, as well as the evaluators themselves, would like to collect additional information. However, when faced with RWE constraints, what would be "interesting to find out" must be culled from "what is essential" to respond to those key questions that drive the evaluation.

As we saw in Chapter 2 and earlier in this chapter, examples of typical key evaluative questions include the following:

- Is there evidence that the project achieved (or will achieve) its objectives? Are there measurable changes in the characteristics of the target population with respect to the impacts the project was trying to produce? Which objectives were (and were not) achieved? Why? Is it reasonable to assume that the changes were due in a significant measure to the project rather than to external factors (not controlled by the project implementers)?

- Did the project aim for the right objectives? Was it based on an adequate diagnosis of the underlying causes of the problem(s) to be addressed?
- What impact has the project had on different sectors of the target population—including the poorest and most vulnerable groups? Are there different impacts on men and women? Are there ethnic, religious, or similar groups who do not benefit or who are affected negatively?
- Are the outcomes sustainable and are benefits likely to continue? Were the target communities or groups reasonably typical of broader populations (such as all poor farmers or all urban slum dwellers) and is it likely that the same impacts could be achieved if the project were replicated on a larger scale?
- What are the contextual and external factors determining the degree of success or failure of the project?

The real-world evaluator must understand which are the critical issues that must be explored in depth and which are less critical and can be studied less intensively or eliminated completely. It is also essential to understand when the client needs rigorous (and expensive) statistical analysis to legitimize the evaluation findings to members of congress or parliament, or to funding agencies critical of the program, and when more general analysis and findings would be acceptable. The answer to these questions can have a major impact on the evaluation budget and time required, particularly on the required sample design and size.

8. Report Findings Efficiently and Effectively

As we mentioned in the section above titled "Customizing Plans for Evaluation", an evaluation should focus on the key questions that relate to the main purpose of its being undertaken. This is especially important for RWE, because choices need to be made as to what can be dropped as a consequence of limitations of time and funding. Those key questions need to be kept in mind not only during the planning for the evaluation, data collection and analysis, but also when the report(s) are being written. There is a temptation to report on all sorts of "interesting findings," but the evaluator(s) need to keep the report focused on answering the key questions that the client(s) and stakeholders wanted answered.

One of the most effective ways to increase the likelihood that evaluation findings are used is to ensure that they are of direct practical utility to the different stakeholders. Some of the factors affecting utilization include the following:

- Timing of the evaluation
- Recognizing that the evaluation is only one of several sources of information and influence on decision makers and ensuring that the evaluation complements these other sources

- Building an ongoing relationship with key stakeholders, listening carefully to their needs, understanding their perception of the political context, and keeping them informed of the progress of the evaluation. There should be "no surprises" when the evaluation report is presented. (Operations Evaluation Department 2005; Patton 1997)

Some steps in the presentation of evaluation findings include these:

- Understand the evaluation stakeholders and how they like to receive information.
- Use visual presentation to complement written reports or verbal presentations. Where appropriate and feasible, make use of presentation tools such as PowerPoint, but do not become a slave to the technology and be prepared to work without this if the logistics become too complicated. Visual presentations are particularly useful when the presentation is not made in the first language of many people in the audience.
- Share the evaluation results through oral presentations. Many stakeholders are not comfortable with written reports or slide presentations, so talking about the findings can be important.
- Plan the written report to make it simple, attractive, and user-friendly. Consider presenting different versions of the findings in ways that are most understandable and useful to different audiences. (Well say more about this below.)
- Involve the mass media. When a goal is to reach and influence a wide audience (e.g. public opinion, all parents of secondary-school-age children, lawmakers), the press can be a valuable ally. However, working with the media requires time and preparation and if their involvement is important, it may be worth hiring a consultant who "knows the ropes."

Succinct Report to Primary Clients

The impact of many evaluations is reduced because the findings and recommendations do not reach the primary clients in a form they like and understand. There is no one best way to report evaluation findings, which depends on the clients and the nature of the evaluation. A good starting point is to ask clients which previous reports they found most useful and why.

A general rule, particularly for RWE where time tends to be a constraint, is to keep the presentation short and succinct. It is a good idea to have a physically short document that can be widely distributed; although the executive summary at the start of a large report may be well written, some clients and stakeholders may be intimidated by the size of the document and may not get round to opening the summary.

Vaughan and Buss (1998) present some useful guidelines for figuring out what to say to busy policymakers and how to say it. They point out that many policymakers have

the intellectual capacity to read and understand complicated analysis, but most do not have the time. Consequently, many will want to be given a flavor of the complexities of the analysis (they do not wish to be talked down to) but without getting lost in details. Other policymakers may not have the technical background and will want a simpler presentation. So there is a delicate balance between keeping the respect and interest of the more technical while not losing the less technical. However, everyone is short of time. Therefore the presentation must be short, even if not necessarily simple. Vaughan and Buss's rules for figuring out what to say are the following:

- Analyze policy but not politics. Evaluators are hired to provide technical expertise, not to advise on political strategies.
- Keep it simple.
- Communicate reasoning as well as bottom lines. Many policymakers will want to know how the evaluator arrived at the conclusions so that they can assess how much weight to give to the findings.
- Use numbers sparingly.
- Elucidate, don't advocate. If evaluators advocate particular policies they risk losing the trust of the policymaker.
- Identify winners and losers. Decision makers are concerned with how policies affect their constituencies, particularly in the short run. Consequently, if evaluators and analysts want policymakers to listen to them, they must identify winners and losers. For example, one of the most effective selling points of the study on why the very expensive but politically sensitive wheat flour ration program in Pakistan should be terminated was the analysis of who were the potential losers (the distributors of wheat flour and the retail store owners) and how their losses could be mitigated (Operations Evaluation Department 2005, chap. 6).
- Don't overlook unintended consequences. People will often respond to new policies and programs in unexpected ways, particularly to take advantage of new resources or opportunities. Sometimes unexpected reactions can destroy a potentially good program, and in other cases unanticipated outcomes may add to the program's success. Policymakers are sensitive to the unexpected because they understand the potentially high political or economic costs. Consequently, if the evaluation can identify some important consequences of which policymakers were not aware, this will catch the attention of the audience and raise the credibility of the evaluation.

Practical, Understandable, and Useful Reports to Other Audiences

In addition to the client and other primary stakeholders (e.g., concerned government ministries and the funding agency), there are often other stakeholders who are interested in the evaluation for different reasons. Some groups, such as members of the target population, are directly affected by the evaluation; others are involved in

advocacy and either wish to use the findings to support their arguments or to criticize the report because it does not support them; and others are interested in the practical applications of the findings. Often the client does not wish to have the evaluation findings too widely disseminated, particularly if they are critical or might raise sensitive issues. In these cases, evaluators may face sensitive ethical and professional concerns about whether they have the ethical and perhaps professional obligation to disseminate the evaluation findings to all groups affected by the project, despite the instruction of the client to limit distribution. These ethical issues are discussed in Chapters 6 and 7.

Assuming that these ethical issues are satisfactorily resolved, a dissemination strategy has to be defined to reach groups with different areas of interest, levels of expertise in reading evaluation reports, and preferences in terms of how they like to receive information. In some cases, different groups may also require the report in different languages. The evaluation team must decide which stakeholders are sufficiently important to merit the preparation of a different version of the report (perhaps even translation into a different language) or the organization of separate presentations and discussions.

These issues are particularly important for RWE because reaching the different audiences, particularly the poorest, least educated, and least accessible has significant cost and time implications. There is a danger that when there are budget or time constraints, the evaluation will reach only the principle clients, and many of the groups whose lives are most affected (such as the indigenous groups whose way of life is threatened, the urban squatters who may be forcibly relocated, or the low-income communities who may or may not benefit from the new water and sanitation technology) may never see the evaluation and may never be consulted on the conclusions and recommendations.

An important purpose of the scoping exercise (Step 1 of the RWE approach) is to agree with the client who will receive and have the opportunity to express opinions about the evaluation report. If the client shows little interest in wider dissemination, but is not actively opposed, then the evaluator can propose cost-effective strategies for reaching a wider audience. If, on the other hand, the client is actively opposed to wider consultation or dissemination, then the evaluator must consider the options—one of which would be to not accept the evaluation contract.

Assuming the main constraints to wider dissemination are time and budget, the following are some of the options:

- Enlist the support of the mass media. It will often be necessary to invest considerable time in cultivating relationships with television, radio, and print journalists. They can be invited to join in field visits or community meetings and they can be sent interesting news stories from time to time.
- Enlist the support of NGOs and civil society organizations. They will often be willing to help disseminate but may wish to present the findings from their own perspective (which might be quite different from the evaluation team's findings), so it is important to get to know

different organizations before inviting them to help with dissemination.

- Meetings can be arranged with organizations in the target communities to present the findings and obtain feedback. It is important that these meetings are organized sufficiently early in the report preparation process so that the opinions and additional information can be incorporated into the final report.

9. Help Clients Use the Findings Well

Unfortunately, it is all too common for an evaluation to be completed, a formal report written and handed over to the client, and then nothing more done about it. Following the above advice, including involving the client and other key stakeholders throughout the evaluation process, one would hope that the findings of an evaluation are relevant and taken seriously. However, if there is no follow-up, one can be left with the impression that the evaluation had no value. There are examples where major donor agencies, noting the limited use of evaluation reports, have decided to simply stop commissioning routine evaluations. Wouldn't it be better for more effort to be put into making sure evaluations are focused on answering key questions, well done, and then more fully utilized?

A major purpose of RWE is to help those involved focus on what is most important and to be as efficient as possible in conducting evaluations that add value and are useful. The final step—utilization—must be a part of that efficiency formula. If information is not used to inform decisions that lead to improved program quality and effectiveness, it is wasted. The point here is that those conducting evaluations need to see that the follow-through is an important part of the evaluation process.

One way to do this is to help the client develop an action plan that outlines steps that will be taken in response to the recommendations of an evaluation and then to monitor implementation of that action plan. Doing this is obvious if this was a **formative evaluation**, where the findings are used to improve subsequent implementation of an ongoing project. Even in the case of a summative evaluation (where the purpose was to estimate the degree to which project outcomes and impacts had been achieved) or where the project that was evaluated has now ended, follow-up should include helping to utilize the lessons learned to inform future strategy and in the design of future projects. At a minimum, those responsible for an evaluation need to do whatever can be done to be sure that the findings and recommendations are documented and communicated in helpful ways to present and future decision makers.

Further reading

A short reading list is given at the end of this chapter. Readers should consult the reading list at the end of each chapter of the book and the extensive references at the end of the book. Appendix 4 also presents electronic resources covering many of the issues discussed in this chapter.

Table 1: Some of the Ways that Political Influences Affect Evaluations

	Examples
During evaluation design	
The criteria for selecting evaluators	Evaluators may be selected: • for their impartiality or their professional expertise • for their sympathy towards the program • for their known criticisms of the program (in cases where the client wishes to use the evaluation to curtail the program) • for the ease with which they can be controlled • because of their citizenship in the country or state of the program's funding agency
The choice of evaluation design and data collection methods	The decision to use either a quantitative or qualitative approach or to collect data that can be put into a certain kind of analytical model (e.g., collecting student achievement or econometric data on an education program) can predetermine what the evaluation will and will not address.
Example of a specific design choice: Whether to use control groups (i.e., quasi-experimental design)	Control groups may be excluded for ethical rather than methodological reasons such as: • to avoid creating expectations of compensation • to avoid denial of needed benefits to parts of a community • to avoid pressures to expand the project to the control areas • to avoid covering politically sensitive or volatile groups. On the other hand evaluators may insist on including control groups in the evaluation design in order to follow conventional practice in their profession even when they contribute little to addressing evaluation questions.
The choice of indicators and instruments	The decision to only use quantitative indicators can lead (intentionally or otherwise) to certain kinds of findings and exclude the analysis of other, potentially sensitive topics. For example, issues of domestic violence or sexual harassment on public transport will probably not be mentioned if only structured questionnaires are used.
The choice of stakeholders to involve or consult	The design of the evaluation and the issues addressed may be quite different if only government officials are consulted, compared to an evaluation of the same program in which community organizations, male and female household heads and NGOs are consulted. The evaluator may be formally or informally discouraged from collecting data from certain sensitive groups, for example by limiting the available time or budget, a subtle way to exclude difficult to reach groups.
Professional orientation of the evaluators	The choice of, for example, economists, sociologists, political scientists or anthropologists to conduct an evaluation will have a major impact on design and outcomes.

The selection of internal or external evaluation	Evaluations conducted internally by project or agency staff have a different kind of political dynamic and are subject to different political pressures compared to evaluations conducted by external consultants, generally believed to be more independent. The use of national versus international evaluators also changes the dynamic of the evaluation. For example, while national evaluators are likely to be more familiar with the history and context of the program, they may be less willing to be too critical of programs administered by their regular clients.
Allocations of budget and time	While budget and time constraints are beyond the total control of some clients, others may try to limit time and resources to discourage addressing certain issues or to preclude thorough, critical analysis.
During implementation	
The changing role of the evaluator	The evaluator may have to negotiate between the roles of guide, publicist, advocate, confidante, hanging judge, and critical friend.
The selection of audiences for progress reports and initial findings	A subtle way for the client to avoid criticism is to exclude potential critics from the distribution list for progress reports . Distribution to managers only, excluding program staff or to engineers and architects, excluding social workers and extension agents will shape the nature of findings and the kinds of feedback to which the evaluation is exposed.
Evolving social dynamics	Often at the start of the evaluation relations are cordial, but they can quickly sour when negative findings begin to emerge or the evaluator does not follow the client's advice on how to conduct the evaluation (e.g., from whom to collect data).
Dissemination and use	
Selection of reviewers	If only people with a stake in the continuation of the project are asked to review the evaluation the feedback is likely to be more positive than if known critics are involved. Short deadlines, innocent or not, may leave insufficient time for some groups to make any significant comments or to include their comments, introducing a systematic bias against these groups.
Choice of language	In developing countries, few evaluation reports are translated into local languages, excluding significant stakeholders. Budget is usually given as the reason, suggesting that informing stakeholders is not what the client considers valuable and needed. Language is also an issue in the U.S., Canada and Europe where many evaluations concern immigrant populations.
Report distribution	Often, an effective way to avoid criticism is to not share the report with critics. Public interest may be at stake, as when clients have a clear and narrow view of how the evaluation results should be disseminated or used and will not consider other possible uses.

Source: *RealWorld Evaluation* Table 6.1

Table 2: Five evaluation strategies and the corresponding designs						
	Methodological strength of the evaluation design [See Table 2 for description of the designs]					
Evaluation strategy	Strongest	Strong	Sound	Weaker	Weakest	Example
1. True experimental design: *Randomized assignment of subjects and strict control of project setting*	Design 1					Testing a new drug under laboratory conditions
2. Randomized field design: *Randomized assignment of subjects but only limited control over project setting*		Design 1				Using a lottery to select villages to participate in self-help water supply project when demand exceeds supply.
3. Strong non-randomized (quasi-experimental) design) *Pre-and post-test project and control groups*			Designs 1 and 2			Low-cost housing project where project participants and comparison groups from types of communities where participants previously lived are interviewed at start and end (5 years later) of project.
4. Weaker non-randomized designs: *Baseline or comparison group eliminated*				Designs 4 and 5		Post-test comparison of communes where rural roads constructed and similar communes without roads
5. Non-experimental designs [only post-test project group]: *No baseline or control group so it is difficult to establish a logically sound counterfactual*					Design 6	Analysis of communities where health centers are operating. There is no baseline survey and no comparison group.

Table 3. A typology of impact evaluation and effects assessment designs

Key: T = Time 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 = Intervention (An intervention is usually an on-going process, but could be a discrete event.)	Start of project [baseline / pre-test]	Intervention (more likely on-going process, rather than one-off event)	Midterm evaluation	End of project evaluation [endline]	Post-project evaluation (some time after intervention ended) [ex-post]	Randomized allocation of subjects to project and control groups (Note 2)	Designs where baseline reconstruction on strategies could be used. (Note 3)	The stage of the project cycle at which each evaluation design begins (Note 4)
	T_1		**T_2**	**T_3**	**T_4**			
LONGITUDINAL DESIGN [*When time and resources permit, this design can be used to strengthen any of the other designs*]								
1. ***Comprehensive longitudinal design*** with pre, mid-term, post- and ex-post observations of both groups.	P_1 C_1	X	P_2 C_2	P_3 C_3	P_4 C_4	*Sometimes*	QUAL	Start.
STATISTICAL DESIGNS [using a matched control group to define the counterfactual]. *While these designs are "rigorous" in their ability to control for statistical selection bias, they are not necessarily stronger than other designs with respect to construct validity and instrument development, process and contextual analysis, giving voice to affected populations or the use of mixed methods.*								
EXPERIMENTAL (RANDOMIZED) DESIGN								
2. ***Randomized control trials***. Subjects are randomly assigned to the project (treatment) and control groups.	P_1 C_1	X		P_2 C_2		*Always*	QUAL	Start.
QUASI-EXPERIMENTAL DESIGNS								
Relatively strong statistical designs with pretest + posttest project treatment and control groups								
Different methods for selecting the project and comparison groups								
3. ***Pre-test + post-test comparison group design with statistical matching of the two groups.*** Participants self-selected or selected by the project agency. Statistical techniques, such as propensity score matching, use secondary data to match both groups on relevant variables.	P_1 C_1	X		P_2 C_2		*Sometimes*	QUAL	Start
4. ***Regression discontinuity.*** A clearly defined cut-off point is used to define project eligibility. Groups above and below the cut-off are compared.								
5. ***Pre-test + post-test with comparison group design with judgmental matching of the two groups.*** Comparison areas are selected judgmentally with subjects randomly selected from within these areas.								

6. ***Pipeline control group design***. When a project is implemented in phases, subjects in Phase 2 (i.e., who will not receive benefits until some later point in time, e.g. mid-term) can be used as the control group for Phase 1 subjects.	P^a_1 C^b_1	X	P^a_2 P^b_1 C^c_2	P^a_3 P^b_2 C^c_3		*Sometimes*	QUAL	Start.
Statistically weaker quasi-experimental designs where baseline data has not been collected on the project and/or control group but where it is still possible to use a statistical counterfactual								
7. ***Pre-test + post-test comparison where the baseline study is not conducted until the project has been underway for some time*** (most commonly this is at time of the mid-term evaluation)		X	P_1 C_1	P_2 C_2		*Sometimes*	QUANT & QUAL	During project implementation - often at mid-term
8. ***Pre-test + post-test comparison of project group combined with post-test (only) comparison of project and control group***	P_1	X		P_2 C_1		*Sometimes*	QUANT & QUAL	Start
9. ***Post-test comparison of project and control groups***		X		P_1 C_1		*Sometimes*	QUANT & QUAL	End
NON-EXPERIMENTAL DESIGNS *The most widely used approaches for assessing program effects. Attribution/contribution is sometimes assessed through non-statistical counterfactuals using approaches such as reference to secondary data, program theory (logic models), theory of change and concept mapping*								
10. ***Pre-test + post-test comparison of project group (****no statistical counterfactual)*	P_1	X	*Sometimes*	P_2	*Sometimes*	*Never*	QUANT & QUAL	Start
11. ***Post-test analysis of project group (****no baseline nor statistical counterfactual)*		X	*Sometime*	P_1	*Sometimes*	*Never*	QUANT & QUAL	End

Notes:
(1) Although there is a technical difference between a control group (used in experimental designs) and a comparison group (used in quasi-experimental designs where a different selection procedures are used for the non-treatment group), we will follow the practice of using *control group* as shorthand, except when we wish to specifically indicate that randomization was not used in which case we will use the term "comparison group" (sometimes called a "non-equivalent control group).
(2) Although the randomized control trial is the only design specifically built around randomized assignment of subjects to the project and control groups, randomization is sometimes incorporated into other design – though sometimes in a more ad hoc way.
(3) All of the designs could consider the use of baseline reconstruction strategies. Designs 1 through 6 include primary collection of baseline data for both groups. However, this normally only includes QUANT data (e.g., survey, structured observation, anthropometric measures). Consequently baseline reconstruction, if used, would mainly focus on QUAL data to complement the already available QUANT data. On the other hand, in the situations represented by designs 4-8 no baseline data had been collected for one or both groups, and consequently baseline reconstruction techniques could be used to collect QUANT as well as QUAL data.
(4) Baseline data may be obtained either from the collection of new data through surveys or other data collection instruments, or it may be obtained from secondary data – census or survey data that has already been collected. When secondary data sources are used, the evaluation may not actually start until late in the project cycle but the design is classified as a pretest + posttest comparison design.

Table 4 The Strengths and Weaknesses of the Nine Project and Control Group Impact Evaluation Designs		
Design	Advantages	Disadvantages
1. Comprehensive longitudinal design with pre-, midterm, post-, and ex-post observations on the project and comparison groups.	This is the strongest design, studying both the implementation process and sustainability. May be required for research testing new project innovation that, if impact can be proven, will be expanded to much greater scale.	The disadvantage is that it is the most expensive, the most time-consuming and the most difficult to implement.
2. Randomized control trial (RCT)	This is the only design that can statically control for sample selection bias as subjects are randomly assigned to the treatment and control groups.	▪ It is estimated that RCTs can probably only be applied in less than 5% of impact evaluations. ▪ Most RCTs do not analyze the process of project implementation and consequently cannot determine whether failure to achieve intended impacts is due to design failure or implementation failure. ▪ Inflexible and unable to adapt to changes in project design, implementation on the local context
3. Pretest-posttest project and comparison groups with statistical matching.	This is the strongest general-purpose QED and for many purposes provides statistically strong estimates of project impact.	▪ As for Design 2 ▪ Can only be used with good quality secondary data is available for statistical matching of samples
4. Regression discontinuity	Can provide unbiased estimates of project impact even when project beneficiaries are non-randomly selected	▪ Requires a clearly defined criterion for project eligibility ▪ The criteria must be strictly and uniformly applied and this is often difficult to ensure.
5. Pretest-posttest project and comparison groups with judgmental matching.	The design is flexible and can be used in a wider range of real-world contexts. Provides reasonably good estimates of project impact when satisfactory matching criteria can be established.	▪ Assumes the comparison group is reasonably similar to project group and willing to participate in two surveys even though they receive no benefits. ▪ Does not assess project implementation
6. Pipeline control group design	▪ Does not require an external control group so the design is cheaper and easier to use	▪ Requires that the Phase 2 project group used as a control and the Phase 1 group are similar. This

		is often not the case ▪ Requires that the Phase 2 group does not have access to Phase 1 benefits and this is often not the case.
7. Truncated longitudinal pretest-posttest project and comparison group design	▪ Observes implementation process as well as impacts. ▪ Reasonably robust model, particularly for projects where implementation begins slowly so that not too much is missed by starting the evaluation late.	▪ Does not begin until around project mid-term, so the project startup and initial implementation period is not captured.
8. Pretest-posttest project group combined with posttest analysis of project and comparison groups.	▪ Assesses if the project model works and produces the intended outputs. ▪ Assesses similarities and differences between project and control areas. ▪ Assesses the extent to which the project could potentially be replicated.	▪ Does not assess whether observed end-of-project differences between the project and comparison groups are due to the project or to preexisting differences between the two groups. ▪ Does not control for local history that might affect outcomes.
9. Posttest project and comparison groups	▪ Evaluates projects that implement well-tested interventions or that operate in isolated areas where there is no interference from other outside interventions.	▪ Does not estimate the exact magnitude of project impacts ▪ Does not control for local history ▪ Does not assess potential for replication on a larger scale ▪ Does not study the project implementation process

Note: The strength of all of these models can be increased by combining them with the impact evaluation framework and analysis of contextual factors discussed in Chapter 9 (pp 175-77) and with some of the RWE techniques discussed in Chapter 10. For Designs 1,2,3,4 and 5, which use comparison groups, the analysis can be greatly strengthened by using multiple regression to statistically control for differences in the characteristics of the project and comparison groups. Where appropriate secondary data is available, these designs can also be strengthened through statistical matching techniques such as propensity score matching and instrumental variables (See World Bank 2006).

Source: Bamberger, Rugh and Mabry. 2006. RealWorld Evaluation Table 10.3

Table 5. Reducing Costs of Data Collection and Analysis for Quantitative and Qualitative Evaluations	
Quantitative Evaluations	**Qualitative Evaluations**
A. Simplifying the evaluation design	
• Pre-test post-test comparison of project group with post-test comparison of project and control groups (Design 3): eliminates baseline control group • Pre-test post-test comparison of project group (Design 4): eliminates pre-test and post-test control group. • Post-test comparison of project and control group (Design 5): eliminates baseline • Evaluation based on post-test data from project group (Design 7): eliminates control group and baseline project group	• Prioritize and focus on critical issues. • Reduce the number of site visits or the time period over which observations are made. • Reduce the amount and cost of data collection. • Reduce the number of persons or groups studied.
B. Clarifying client information needs	
Prioritize questions and data needs with the client to try to eliminate the collection of data not actually required for the evaluation objectives.	
C. Use existing data	
• Census or surveys covering project areas • Data from project records • Records from schools, health centers and other public service agencies	• Newspapers and other mass media • Records from community organizations • Dissertations and other university studies [for both QUAL and QUANT]
D. Reducing sample size	
• Lower the level of required precision (lower precision = small sample) • Reduce types of disaggregation required (less disaggregation = smaller sample • Stratified sample designs (less interviews) • Use cluster sampling (lower travel costs)	• Consider critical or quota sampling rather than comprehensive or representative sampling • Reduce the number of persons or groups studied.

E. Reducing costs of data collection, input and analysis	
• Self-administered questionnaires (with literate populations) • Direct observation (instead of surveys) (sometimes saves money but not always) • Automatic counters and other non-obtrusive methods • Direct inputting of survey data through hand-held devices. • Optical scanning of survey forms and electronic surveys	• Decrease the number or period of observations • Prioritize informants • Employ and train university students, student nurses, and community residents to collect data (for both QUAL and QUANT) • Data Input through hand-held devices.
Mixed method designs • triangulation to compensate for reduced sample size. • focus groups and community forums instead of household surveys • Participatory Rapid Appraisal (PRA) and other participatory methods	

Source: Bamberger, Rugh and Mabry. 2006. Table 3.1.

Table 6 Estimated Cost Savings for Less Robust RWE Designs Compared with Design 2

Design		*Estimated Cost Saving Compared with Design 2*
3	Truncated longitudinal design	5–10%
4	No comparison group baseline study	10–20%
5	No baseline study for either group	30–40%
6	No comparison group	40–50%
7	Only posttest project group	60–80%

NOTE: The estimated cost savings are based on the percentage reduction in the total number of interviews, but taking into account that there are fixed costs, such as questionnaire design and training.

Table 7 Factors Affecting the Sample Size

Factor	Explanation	Influence on sample size
1. The purpose of the evaluation	Is this an exploratory study or are very precise statistical estimates required?	The more precise the required results the larger the sample
2. Will a one or two-tailed test be used? (.Is the direction of the expected change known?)	If the purpose of the evaluation is to test whether positive outcomes have increased, or negative ones have declined. Then a one-tailed test can be used. If the purpose is to test whether there has been "a significant change" without knowing the direction. Then a two-tailed test is required	The sample size will be approximately 40% larger for a two-tailed test.
3. Is only the project group interviewed?	In some evaluation designs only subjects from the project group are interviewed. This is the case if information on the total population is available from previous studies or secondary data. In other cases a comparison group must also be selected and interviewed.	The sample size will be doubled if the same number of people have to be interviewed in both the project and comparison groups.
4. Homogeneity of the group	If there is little variation among the population with respect to the outcome variable, then the standard deviation will be small.	The smaller the standard deviation the smaller the sample.
5. The effect size	Effect size is the amount of increase the project is expected to produce	The smaller the effect size the larger the sample
6. The efficiency with which the project is implemented	While some projects are implemented in a very efficient way with all subjects receiving exactly the same package of services, in other cases the administration is poorer and different subjects receive different combinations of services. The quality of the services can also vary.	The poorer the quality and efficiency of the project, the larger the sample.
7. The required level of disaggregation.	In some cases the client only requires global estimates of impact for the total project population. In other cases it is necessary to provide disaggregated results for different project sites, variations in the package of services provided or for different socio-economic groups (sex, age, ethnicity etc)	The greater the required disaggregation the larger the sample
8. The sample design	Sampling procedures such as stratification can often reduce the variance of the estimates and increase	Well designed stratification may reduce sample size.

	precision.	
9. The level of statistical precision	When estimating whether the project had an impact "beyond a reasonable doubt", this is normally defined as "less than a 1 in 20 chance that an impact as large as this could have occurred by chance". This is defined at the 0.05 confidence level. If more precise results are required the 0.01 level may be used (less than 1 in a 100). For an exploratory study the 0.10 level may be used (a 1 in 10 chance).	The higher the confidence level the larger the sample
10. The power of the test	The statistical power of the test refers to the probability that when a project has "real" effect, this will be rejected by the statistical significance test. The conventional power level is 0.8 meaning that there is only a 20% chance that a real effect would be rejected. Where a higher level of precision is required the Power can be raised to 0.9 or higher	The higher the power level the larger the sample.
11. Finite population correction factor	If the sample represents more than say 5% of the total population it is possible to reduce the sample size through the finite population correction factor	The greater the proportion the sample represents of the total population the smaller the sample.

Source: Bamberger, Rugh and Mabry. 2006. Table 3.3

Table 8. Reducing the time required for data collection and analysis in quantitative and qualitative evaluations.

Approaches also used to reduce costs		
1. Simplifying the evaluation design		
2. Clarifying and prioritizing client information needs		
3. Using existing documentary data		
4. Reducing sample size		
5. Using cheaper and faster methods of data collection		
Additional approaches that save time but may not save money and often increase costs		
	Quantitative	**Qualitative**
6. Reducing time constraints on external (often foreign) consultants or sub-contractors a. Commissioning the advance collection and organization of available data by consultants b. Commissioning exploratory <u>studies by a local consultant to identify some of the key issues and the characteristics of the population prior to the arrival of the external consultant.</u> c. Video-conferences involving external and local consultant prior to the visit of the external consultant can advance planning and save time.	a. Compilation of secondary data and initial assessment of quality and relevance for the study. (Note: QUAL studies would use the same type of existing documents.) b <u>Rapid surveys</u> to obtain demographic, economic or other relevant data on the target populations to help develop the sample design and the preparation of the sampling frame (list or map with the location of all families or other subjects in the population studied). Rapid studies can also be used to obtain preliminary estimates of, for example education or literacy scores. c. Establish rapport with the community and local leaders and officials to facilitate the smooth implementation of the study and to avoid bureaucratic delays (for example obtaining documents required to start the study.	a. Compilation of research literature and sources such mass media materials, photographs b. <u>Rapid Ethnographic studies</u>, focusing on key concepts and issues to be covered in the study and to lay the groundwork for the external consultants. c. photos, videos and tape-recordings that can be sent to external consultants to document the conditions of the communities during different times of year (for example the monsoons and the dry season) . This can be important if consultants are not able to visit the region in every season. c. Establish rapport (as for QUANT)

7. Hiring more data collectors a. Increasing the number of interviewers and supervisors b. Hiring more experienced interviewers and supervisors. This can reduce the time required for training and can increase the efficiency and speed of the regular data collectors. c. Sub-contracting data collection or analysis.		
8. Revising format of project records to include critical data for impact analysis	Include indicators on access and use of services as well as relevant household or individual characteristics – particularly impact indicators.	Encourage/enable project staff to record more than project activities, if relevant to the evaluation; to document observations of changes in conditions in beneficiaries' households.
9. Modern data collection and analysis technology.	• Hand-held computers for data input • Optical scanning • Automatic counters • QUANT computer software for data analysis and presentation (note clear whether this does save time as the main purpose is to permit more comprehensive analysis)	• E-mail surveys (these can be used for both QUAL and QUANT studies) • Video-cameras and tape recorders • photography • GPS mapping and aerial photography can be used to observe demographic patterns, agricultural practices and the conditions of infrastructure over a large which can save the considerable amounts of time required to reach remote villages and areas, • QUAL computer software for data analysis and presentation

Table 9 Rapid Data Collection Methods

		Savings of elapsed time, effort, or both		
	Ways to reduce time requirements	Elapsed	Effort	Both
A. Mainly qualitative methods				
Key informant interviews	Key informants can save time either by providing data (agricultural prices, people leaving and joining the community, school attendance and absenteeism) or by helping researchers focus on key issues or pointing out faster ways to obtain information. Ways to reduce time of key informant interviews • Reduce the number of informants • Limit the number of issues covered • Hire more researchers to conduct the interviews or to tape interviews for the researcher to review. Do this with caution as it is important for the researcher to maintain personal contact with key people in the community			
Focus groups and community interviews	• Sub-contract to focus group specialists such as market research firm • Conduct several focus groups simultaneously instead of sequentially • Collecting information from meetings rather than surveys. Information on topics such as access to and use of water and sanitation; agricultural practices and gender division of labor in farming can be obtained in group interviews possibly combined with the distribution of self-administered surveys. • Important to use techniques to ensure views of all participants are captured (time pressures mean that more vulnerable and harder to access groups may be left out).	√		
Structured observation	• Observation can sometimes, but not always be faster than surveys. For example: observation of the gender division of labor in different kinds of agricultural production, who attends meetings and participates in discussions, types of conflict observed in public places in the community.			
Use of preexisting documents and artifacts	• Many kinds of pre-existing data can be collected and reviewed more rapidly than new data can be collected. For example, school attendance records, newspapers and other mass media, minutes of community meetings, health center records, surveys in target communities conducted by research institutions.	√		

Using community groups to collect information	• Organization of rapid community studies (QUAL and QUANT) using community interviewers (local school teachers often cooperate with this)			
Photos and videos	• Giving disposable cameras or camcorders to community informants to take photos (or make videos) illustrating, for example, community problems.	√		
Triangulation	• Having several interviewers simultaneously interview and separately record their observations on the same key respondents rather than having separate interviews. This can save elapsed time if it replaces several separate interviews with the same person	√		
B. Mainly quantitative methods				
Rapid surveys with short questionnaires and small samples	• Reducing the number of questions and the size of the sample can significantly reduce the time required to conduct a survey. • Increasing the number of interviewers			
Reduce sample sizes	• There are specialized sampling techniques such as Lot Quality Acceptance Sampling (Valadez and Bamberger 1994) designed to provide estimates of the utilization or quality of public services such as health and education with very small samples. Samples of 14-28 households may be sufficient to assess utilization or quality of a single health center.			
Triangulation (used also in QUAL and Mixed methods)	• Obtaining independent estimates from different sources (e.g., survey and observation) sometimes makes it possible to obtain estimates from smaller samples hence saving both elapsed time and effort.	√		
Rapid exit surveys	• People leaving a meeting or exiting a service facility can be asked to write their views on the meeting or service on an index card which is put on the wall. • Often only one key question will be asked. For example: "Would you recommend a neighbor to come to the next meeting or use this center)?"	√		
Use of preexisting data	• Previous surveys or other data sources may eliminate the need to collect certain data • Previous survey findings can reduce the time required for sample design or by providing information on the *standard deviation* (how narrowly or widely subjects are distributed around the mean) of key variables may make it possible to reduce sample size or to save time through more efficient *stratification* or *cluster sampling*. (These terms are defined in Chapter 16).	√		

Observation checklists	• Observation checklists can often eliminate the need for certain surveys (for example pedestrian and vehicular traffic flows, use of community facilities, time required to collect water and fuel).			
Automatic counters	• Recording people entering buildings or using services such as water.			
C. Mixed methods				
Triangulation (used also in QUAL and QUANT methods)	• Triangulating data from several quantitative and qualitative methods may sometimes make it possible to obtain estimates from smaller samples hence saving effort and elapsed time. Note: not always the case as use of more data collection methods has obvious time/cost implications			
Rapid quantification of *participatory assessment methods* and focus groups	• Short and rapid sample surveys can be combined with numerical estimates obtained from community interviews and focus groups to provide estimates of, for example, service usage, unemployment rates, *time-use* for a community or other population group			
Note: It is often difficult to differentiate between saving time and reducing effort. It is also important to stress that saving time by increasing the size of the team will usually increase the budget. Hence the need to clarify with the client whether the major constraint is time, budget or both.				

Table 10 Strategies for Addressing Data Constraints

Reconstructing Baseline Data[a]		
Approaches	*Sources/Methods*	*Comments/Issues*
Using existing documents (secondary data)	• Project records • Data from public service agencies (health, education, etc.) • Government household and related surveys	
Assessing the reliability and validity of secondary data (see Chapter 8 for a discussion of these concepts)	• School enrollment and attendance records • Patient records in local health centers • Savings and loans cooperatives records of loans and repayment • Vehicle registrations (to estimate changes in the volume of traffic) • Records of local farmers markets (prices and volume of sales)	All data must be assessed to determine their adequacy in terms of • Reference period • Population coverage • Inclusion of required indicators • Documentation on methodologies used • Completeness • Accuracy • Freedom from bias
Using recall: asking people to provide numerical (income, crop production, how many hours a day they spent traveling, school fees) or qualitative (the level of violence in the community, the level of consultation of local government officials with the community) at the time the project was beginning	• Key informants • PRA (participatory rural appraisal) and other participatory methods	Recall can be used for • School attendance • Sickness/use of health facilities • Income/earnings • Community/individual knowledge and skills • Social cohesion and conflict • Water usage and cost • Major or routine household expenditures • Periods of stress • Travel patterns and transport of produce
Improving the reliability/validity of recall	• Conduct small pretest-posttest studies to compare recall with original information • Identify and try to control for potential bias (underestimation of small expenditures, truncating large expenditures by including	

	some expenditures made before the recall period, distortion to conform to accepted behavior, intention to mislead) • Clarifying the context (time period, specific types of behavior, reasons for collecting the information) • Link recall to important reference points in community or personal history • Triangulation (key informants, secondary sources, PRA)	
Key informants	• Community leaders • Religious leaders • Teachers • Doctors and nurses • Store owners • Police • Journalists	• Use to triangulate (test for consistency) data from other sources
Special Issues and Challenges When Working with Comparison Groups		
Approach	*Sources*	*Comments/ Issues*
Identifying and reconstructing comparison groups	Government, statistics, earlier surveys, records of schools, health centers and other public service agencies.	Challenges and issues include • Political pressures • Ethical issues in using comparison groups • Using previous surveys as sampling frame • Rapid pilot studies to test variance etc. • Judgmental matching • Use later phases of project as comparison • Internal comparison groups when different participants receive different combinations of services • Appropriateness of potential comparison groups • Statistical creation of control (cluster analysis, analysis of
Special issues in reconstructing data on comparison groups	• Econometric posttest comparison of project and comparison areas cannot	Methodological issues • Self-selection of participants (issues: difficult

	control for historical differences between the two groups (see Chapter 10)	to match a comparison group on factors such as motivation) • Projects selected to represent either groups with the greatest potential to succeed or the groups facing the greatest challenge (issues: in both cases, difficult to find comparison group with similar characteristics)
Collecting sensitive data (e.g., domestic violence, fertility behavior, household decision making and resource control, information from or about women, and information on the physically or mentally handicapped)	• Participant observation • Focus groups • Unstructured interviews • Observation • PRA techniques • Case studies • Key informants	These issues also exist with project participants, but they tend to be more difficult to address with comparison groups because the researcher does not have the same contacts or access to the community.
Collecting data on difficult-to-reach groups (e.g., sex workers, drug or alcohol users, criminals, informal small businesses, squatters and illegal residents, ethnic or religious minorities, and in some cultures, women.)	• Observation (participant and non-participant) • Informants from the groups • Self-reporting • Tracer studies and snowball samples • Key informants • Existing documents (secondary data) • Symbols of group identification (clothing, tattoos, graffiti)	As for previous point

a. Similar approaches can be used for project and comparison areas, but there is often greater access to information for project populations.

Table 11 Factors Determining the Adequacy of the Evaluation Design and of the Findings

1. **How well suited are the evaluation focus, approach, and methods for obtaining the information needed regarding, for example:**
a. Managerial decisions
b. Stakeholder perspectives on program adequacy
2. How available are data and data sources, for example:
a. Whether appropriate data exist or can be generated to address information needs
b. Whether stakeholders and documentary data sources are accessible to evaluators
3. How well the data will support valid interpretations about the program regarding, for example:
a. Achievement of program goals, extent of delivery of program benefits
b. Cost-effectiveness of the program
d. The adequacy of resources affecting goal attainment
e. Unintended consequences
4. How adequate the evaluation team is, for example in terms of:
a. Evaluation methodology
b. The specific field of the program
c. Sufficiency of evaluation resources for the scope of the program

Table 12 Some Threats to Validity can Affect even the Strongest Quantitative Designs (Designs 1 and 2)

Threats to Validity[a]	
F. **Threats to statistical conclusion validity**	
1. *Low statistical power*	The sample is too small to be able to detect statistically significant effects (see Chapter 14).
4. *Unreliability of measures*	The indicators do not adequately measure key variables.
5. *Restriction of range*	The sample does not cover the whole population. For example, the lowest or highest income groups are excluded or the sample only covers enterprises employing more than 10 people.
6. *Unreliability of treatment implementation*	The treatments were not applied uniformly to all subjects, and often there is no documentation of the differences in application. For example, some mothers received malaria tablets and guidance from the nurse, others received only the tablets.
G. **Threats to internal validity**	
2. *Selection bias*	Differences between project and comparison groups with respect to factors affecting outcomes.
6. *Attrition*	While the project group is initially representative of the total population, certain subgroups (e.g., the less educated, women with small children, the self-employed) have higher dropout rates, so the people who are actually exposed to the project are no longer representative of the whole population.
H. **Threats to construct validity**	
1. *Inadequate explanation of constructs and program theory model*	The basic concepts of the model are not clearly explained or defined.
8. *Reactivity to the data collection instruments*	Responses may be affected by how subjects react to the interview or other data collection methods. For example, respondents may report that they are poorer than they really are or that the project has not produced benefits because they are hoping the agency will provide new services or reduce the cost of current services.
I. **Threats to external validity**	
6. *Influence of policymakers on program outcomes*	Support or opposition of policymakers in particular locations may affect program outcomes in ways that might be difficult to assess.
7. *Seasonal cycles*	Many surveys are conducted at only one time in the year and may not adequately capture important seasonal variations.

a. See Integrated Checklist for Assessing Threats to Validity of Quantitative, Qualitative, and Mixed-Methods Designs (Appendix 1) for the full list of threats. The numbers in the left column correspond to that checklist.

Table 13 Characteristics of QUANT and QUAL Approaches to Different Stages of the Evaluation Process

Evaluation Activity	*Quantitative Approach*	*Qualitative Approach*
The conceptual framework and the formulation of hypotheses	• Evaluations are usually, but not always, based on a theoretical framework derived from a review of the literature and usually involve testable hypotheses. • Hypotheses are often *deductive* (based on testable hypotheses derived from theory). • Hypotheses are usually quantitative and can be evaluated with statistical significance tests. • The framework often starts from the macro rather than the micro level.	• Although some evaluations define and test hypotheses, many do not. • Many evaluations emphasize the uniqueness of each situation, and the conceptual framework may be defined through a process of iteration with the framework being continuously updated as new information is obtained. • Hypotheses, if used, are often *inductive* (derived from information gathered during the course of the study).
Selection of subjects or units of analysis	• Random sampling means that findings can be generalized and permits statistical testing of differences between groups. • Requires a sampling frame that lists all the members of the target population(s) to be studied. • Selection methods are usually defined in advance, clearly documented and unchanging throughout the study. • Typically, a fairly large sample is selected from which to collect a finite set of quantitative data.	• Choice of selection procedures varies according to the purpose of the study. • Purposive sampling is used to collect the most useful and interesting data related to the purpose of the study. • Although this is not usually done for QUAL evaluations, sometimes for mixed-method approaches, the sample may be selected using the same master sampling frame as for the QUANT component of the research. For example, a subsample of the villages in which samples of households (or other units) are selected for the QUANT survey may be selected for the QUAL analysis (although the types of data collection and the subjects, groups or organizations to be studied in the QUAL analysis will usually be different). • Usually a smaller number of people interviewed in more depth.
Evaluation design	• Normally, one of the quasi-experimental designs described in Chapter 10 is used. A randomly selected sample that represents the project participants and, possibly, a control or comparison group is interviewed at one or more points	• The researcher(s) become immersed in the community over a long period of time. • The effects of the program are studied through collecting information on the many different elements of the community

	during the project. • Where possible, outcomes (impacts) are estimated by comparing data collected before and after (and possibly during) the implementation of the project.	and its economic, political, cultural, ecological, and psychological setting. • Normally, the evaluation does not try to establish a direct cause and effect or linear relationship.
Data collection and recording methods	• Data are usually recorded in structured questionnaires that are followed consistently throughout the study. • There is extensive use of pre-coded, closed-ended questions. • The study mainly uses numerical values (integer variables) or closed-ended (ordinal or nominal) variables that can be subjected to statistical analysis. • Observational checklists with pre-coded responses may be used.	• Interview protocols are the most common instrument, often semi-structured. • The data collection instrument may be modified during the course of the study as understanding grows. • Interview data are sometimes recorded verbatim (audiotape, videotape) and sometimes in written notes. • Study may use analysis of existing documents (document analysis). Textual data from documents are often highlighted in a copy of the original, which is kept as part of the data set. • Study may use focus groups (usually fewer than 10 people) and meetings with larger community groups. • Study may use participant and non-participant observation. • Study may use photography.
Triangulation	• Consistency checks are built into questionnaires to provide independent estimates of key variables (e.g., data on income may be compared with data on expenditures). • Direct observation (a QUAL technique) can be used as a consistency check on answers given by the respondent (e.g., information on income can be compared with evidence of the number and quality of consumer durables in evidence inside or outside the house). • Information from earlier surveys with the same respondents is sometimes used as a consistency check on information given in a later survey. • Secondary data (census data, national household surveys, information from government agencies) can be used to check	• Several qualitative methods are used for multiple perspectives and triangulation. • Triangulation by observation: A monitor can observe a focus group or group meeting both to identify any potential bias resulting from how the session was conducted and also to provide an independent perspective (e.g., reporting on the interactions between group members, observing how certain people respond to the comments or behavior of others).

	estimates from the evaluation survey.	
Data analysis	See Chapter 11 for more details on these and other methods. • Study may use analysis of non-response to determine if there are any systematic biases (e.g., higher non-response rates among lower- or higher-income participants, different response rates for men and women). • Study may use analysis of missing data and "outliers"; decisions must be made about whether to exclude these values (with possible resulting biases in the estimations) or to make statistical adjustments. • Study may use descriptive statistics—indicators of central tendency and dispersion and central tendency (see Chapter 11) • Study may use multivariate analysis to examine factors contributing to the magnitude and direction of change. • Study may use significance tests for differences between groups.	See Chapter 12 for a discussion of most of these methods, • Study may use inductive analysis from data to analyze patterns from which understandings are developed and interpretations constructed. • Study may use thematic content analysis of interviews, observation reports, documents, and other sources. • Unique case analysis assumes that each case has unique attributes and that much can be learned by concentrating on a single case. • Study may use cross-case analysis that helps assess the extent to which findings from individual cases can be generalized and helps see processes and outcomes across many cases. • Study may use constant-comparative method in which new data and preliminary interpretations are constantly compared. • QUAL analysis is holistic, and the program being studied is viewed as a complex tapestry of interwoven threads. Analysis emphasizes the context (setting) and how this affects the operation of the program being evaluated. • QUAL analysis involves intuitive understanding. • Case studies may be conducted to study outliers or to help explain and understand the program in its full scope.

Table 14 Elements of an Integrated, Multidisciplinary Research Approach

Composition of the Research Team

- Include primary researchers from different disciplines. Allow time for researchers to develop an understanding and respect for each other's disciplines and work. Each should be familiar with the basic literature and current debates in the other field.
- Ensure similar linkages between local researchers from the city, state, or region where the project is being implemented.

Integrated Approaches during the Evaluation Design

- Ensure that the evaluation framework draws on theories and approaches from all the disciplinary teams involved in the evaluation (e.g., anthropology, medicine, law, sociology, economics, demography) and frameworks from predominantly qualitative and quantitative perspectives, with each being used to enrich and broaden the other.
- Ensure that hypotheses and research approaches draw from all disciplines.
- The research framework should formulate linkages between different levels of analysis (e.g., both quantitative survey and qualitative interviews of households, students, farmers; qualitative holistic analysis of the program setting).
- Ensure that concepts and methods are not taken out of context but draw on the intellectual debates and approaches within the respective disciplines.
- Consider using behavioral models that combine economic and other quantitative modeling with in-depth understanding of the cultural context within which the study is being conducted.

Data Collection and the Use of Triangulation

- Conduct exploratory analysis to assist in hypothesis development and definition of indicators.
- Select quantitative and qualitative data collection methods designed to complement each other, and specify the complementarities and how they will be used in the fieldwork and analysis.
- Select at least two independent estimating methods for key indicators and hypotheses.
- Ensure full documentation of all sample selection, data collection, and analysis methods.

Data Analysis and Interpretation and Possible Field Follow-Up

- Conduct and present separate analyses of quantitative and qualitative findings to highlight different interpretations and findings from different methods and then prepare an integrated report drawing on all of the data.
- Use systematic triangulation procedures to check on inconsistencies or differing interpretations. Follow up on differences, where necessary, with a return to the field.
- Budget resources and time for follow-up visits to the field (not just for mixed methods).
- Highlight different interpretations and findings from different methods and discuss how these enrich the interpretation of the study. Different outcomes should be considered a major strength of the integrated approach rather than an annoyance.
- Present cases and qualitative material to illustrate or test quantitative findings.

Presentation and Dissemination of Findings

- Combine conventional forms of presentation with written reports complemented by PowerPoint presentations with some of the more participatory presentation methods used in some qualitative evaluations. Recognizing lack of receptivity by many stakeholders to long, technical reports, the team may also develop more innovative and user-friendly reports.
- Broaden the range of stakeholders invited to the presentation and review of findings to include some of the community and civil society groups that qualitative evaluators often work with but many of whom may not be consulted in many quantitative evaluations.

Appendix 1

CHECKLIST FOR ASSESSING THREATS TO THE VALIDITY OF AN IMPACT EVALUATION[5]

Part I Cover Sheet

1. Name of project/program
2. Who conducted this validity assessment? (indicate organizational affiliation)
3. When did the evaluation begin? A. Start of the project ___ B. Mid-term ___ C. Towards the end of the project ___ D. When the project has been operating for several years ___
4. At what stage of the evaluation was this assessment conducted? A. Proposed evaluation design ___ B. Progress report on the evaluation ___ C. Draft final evaluation report ___ D. After the evaluation has been completed ___
5. Reason for conducting the threats to validity assessment
6. Summary of findings of the assessment
7. Recommended follow-up actions (if any)

[5] **Source**: Adapted by the authors from Miles and Huberman (1994) Chapter 10 Section 1; Guba and Lincoln (1989); Shadish, Cook and Campbell (2002) Tables 2.2, 2.4, 3.1 and 3.2; Bambergr, Rugh and Mabry (2006) Chapter 7 and Appendix 1 and Bamberger (2007). The present authors are solely responsible for the adaptation and interpretation of the data.

Part II SUMMARY ASSESSMENT FOR EACH COMPONENT **[***see attachments for more detailed assessments***]**						
	Very strong				Serious problems	Not applicable
	1	2	3	4	5	N/A
Component A. Objectivity (*Confirmability*) **[Attachment A]:** *Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						
Component B. Reliability [Attachment B]: *Is the process of the study consistent, coherent and reasonably stable over time and across researchers and methods? If emergent designs are used are the processes through which the design evolves clearly documented?*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						
Component C. **Internal validity** (Credibility) [Attachment C]: *Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying? Are there reasons why the assumed causal relationship between two variables (e.g. project treatment and outcome or impact) may not be valid?*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						

Component D. **Statistical conclusion validity** [Attachment D]: *Reasons why inferences about statistical association (e.g between treatments and outcome/impact or the differences between project and control group) may not be valid.*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						
Component E. Construct Validity [Attachment E]. *The adequacy and comprehensiveness of the constructs used to define processes, outcomes and impacts, contextual and intervening variables (moderators and mediators).*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						
Component F. External Validity (transferability) [Attachment F]: *Reasons why inferences about how study results would hold over variations in persons, settings, treatments and outcomes may not be correct.*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						
Component G. Utilization [Attachment G]: *How useful were the findings to clients, researchers and the communities studied?*						
Summary assessment and recommendations						
Overall rating of this component of the evaluation						
Number of issues/problems identified [indicate no. of 4 and 5 ratings]						

ATTACHMENTS A - G

Checklists used to assess 7 categories of potential threats to the adequacy and validity of an impact evaluation

Attachment A. OBJECTIVITY *(Confirmability)*	
Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?	R
1. Are the conclusions and recommendations presented in the executive summary consistent with, and supported by, the information and findings in the main report.	
2. Are the study's methods and procedures adequately described? Are study data retained and available for re-analysis?	
3. Is data presented to support the conclusions? Is evidence presented to support all findings.	
4. Has the researcher been as explicit and self-aware as possible about personal assumptions, values and biases?	
5. Were the methods used to control for bias adequate?	
6. Were competing hypotheses or rival conclusions considered?	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Attachment B. RELIABILITY (*dependability*)	
Is the process of the study consistent, coherent and reasonably stable over time and across researchers and methods? If emergent designs are used are the processes through which the design evolves clearly documented?	
1. Are findings trustworthy, consistent and replicable across data sources and over time? Did methods of data collection and interpretation vary over time as researchers increased there understanding of the phenomena being studied, and if so were adequate measures taken to ensure the reliability and consistency of the data and interpretation?	
2. Were data collected across the full range of appropriate settings, times, respondents, etc.?	
3. Did all fieldworkers have comparable data collection protocols?	
4. Were coding and quality checks made, and did they show adequate agreement?	
5. Do the accounts of different observers converge? If they do not (which is often the case in qualitative studies) is this recognized and addressed?	
6. Were peer or colleague reviews used?	
7. Was the evaluation conducted under budget, time or data constraints? Did this affect the quality of the data or the validity of the evaluation design, and if so what measures were taken to compensate for these limitations.	
8. Were the rules used for confirmation of propositions, hypotheses, etc. made explicit?	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Attachment C. **INTERNAL VALIDITY (***dependability***)**	
Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying? Are there reasons why the assumed causal relationship between two variables may not be valid?	
1. How context-rich and meaningful ("thick") are the descriptions? Is there sufficient information to provide a credible/valid description of the subjects or the situation being studied?	
2. Does the account ring true, make sense, seem convincing? Does it reflect the local context?	
3. Did triangulation among complementary methods and data sources produce generally converging conclusions? If expansionist qualitative methods are used where interpretations do not necessarily converge, are the differences in interpretations and conclusions noted and discussed?	
4. Are the presented data well linked to the categories of prior or emerging theory? Are the findings internally coherent, and are the concepts systematically related?	
5. Are areas of uncertainty identified? Was negative evidence sought, found? How was it used? Have rival explanations been actively considered?	
6. Were conclusions considered accurate by the researchers responsible for data collection?	
7. **Temporal precedence of interventions and effects**. Was it clearly established that the intervention actually occurred before the effect that it was predicted to influence? A cause must precede its effect. However, it is often difficult to know the order of events in a project. Many projects (for example, urban development programs) do not have a precise starting date but get going over periods of months or even years	
8. Project selection bias. Were potential project selection biases identified and were measures taken to address them in the analysis? Project participants are often different from comparison groups either because they are self-selected or because the project administrator selects people with certain characteristics (the poorest farmers or the best-organized communities).	
9. History. Were the effects of history identified and addressed in the analysis? Participation in a project may produce other experiences unrelated to the project treatment that might distinguish the project and control groups. For example, entrepreneurs who are known to have received loans may be more likely to be robbed or pressured by politicians to make donations, or girls enrolled in high school may be more likely to get pregnant.	
10. Attrition. Was there significant attrition over the life of the project and did this have different effects on the composition of the project and comparison groups? Even when project participants originally had characteristics similar to the total population, selective drop-out over time may change the characteristics of the project population (for example the poorest or least educated might drop out)	
11. Testing. . Being interviewed or tested may affect behavior or responses. For example, being asked about expenditures may encourage people to cut down or socially disapproved expenditures (cigarettes and alcohol) and spend more on acceptable items.	
12. Potential biases or distortion during the process of recall. Respondents may deliberately or unintentionally distort their recall of past events. Opposition politicians may exaggerate community problems while community elders may romanticize the past.	
13. Information is not collected from the right people, or some categories of informants not interviewed Sometimes information is only collected from, and about certain sectors of the target population (men but not women, teachers but not students) in which case estimates for the total population may be biased.	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Attachment D. STATISTICAL CONCLUSION VALIDITY	
Reasons why inferences about statistical association (e.g. between treatments and outcomes/impacts or the differences between project and control group) may not be valid	
1. The sample is too small to detect program effects (Low Statistical Power): The sample is not large enough to detect statistically significant differences between project and control groups even if they do exist. Particularly important when effect sizes are small.	
2. Sample size for group and community level variables is too small to permit statistical significance testing. When the unit of analysis is the group, organization or community the sample size tends to be significantly reduced (compared to data collected from household sample surveys) and the power of the test is lowered so that it may not be possible conduct statistical significance testing. This is frequently the case when data is collected at the group level to save time or money.	
3. Unreliability of measures of change of outcome indicators. Unreliable measures of, for example, rates of change in income, literacy, infant mortality, always reduce the likelihood of finding a significant effect.	
4. Unreliability of treatment implementation. If the treatment is not administered in an identical way to all subjects the probability of finding a significant effect is reduced.	
5. Diversity (heterogeneity) of the population If subjects have widely different characteristics, this may increase the variance of results and make it more difficult to detect significant effects.	
6. Extrapolation from a Truncated or Incomplete Data Base. If the sample only covers part of the population (for example only the poorest families, or only people working in the formal sector) this can affect the conclusions of the analysis and can bias generalizations to the total population.	
7. **Project and comparison group samples do not cover the same populations.** It is often the case that the comparison group sample is not drawn from exactly the same population as the project sample. In these cases differences in outcomes may be due to the differences in the characteristics of the two samples and not to the effects of the project.	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Attachment E. CONSTRUCT VALIDITY	
The adequacy and comprehensiveness of the constructs used to define processes, outcomes and impacts, contextual and intervening variables (moderators and mediators).	
1. Inadequate explanation of constructs Constructs (e.g. implementation processes, effects/outcomes) being studied are defined in terms that are too general or are confusing or ambiguous thus making it impossible to have precise measurement. Examples of ambiguous constructs include: quality of life, unemployed, aggressive, hostile work environment, sex discrimination.	
2. Indicators do not adequately measure constructs (Construct confounding) the operational definition may not adequately capture the desired construct. For example, defining the unemployed as those who have registered with an employment center ignores people not working but who do not use these centers. Similarly, defining domestic violence as cases reported to the police significantly under-represents the real number of incidents	
3. Use of single indicator to measure a complex construct (Mono-operation bias) using a single indicator to define and measure a complex construct (such as poverty, well-being, domestic violence) will usually produce bias.	
4. Use of a single method to measure a construct (Mono-method bias). If only one method is used to measure a construct this will produce a narrow and often biased measure (for example, observing communities in formal meetings will produce different results than observing social events or communal work projects	
5. Only one level of the treatment is studied. Often a treatment is only administered at one, usually low, level of intensity (only small business loans are given) and the results are used to make general conclusions about the effectiveness (or lack of effectiveness) of the construct. This is misleading as a higher level of treatment might have produced a more significant effect.	
6. The implicit program theory model on which the project is based is not well documented. This makes it difficult to identify how the key constructs were understood by program planners	
7. Program participants and comparison group respond differently to some questions Program participants may respond in a more nuanced way to questions. For example, they may distinguish between different types and intensities of domestic violence or racial prejudice, whereas the comparison group may have broader, less discriminated responses.	
8. Participants assess themselves and their situation differently than comparison group People selected for programs may self-report differently from those not selected even before the program begins. They may wish to make themselves seem more in need of the program (poorer, sicker) or they may wish to appear more meritorious if that is a criterion for selection.	
9. Reactivity to the experimental situation. Project participants try to interpret the project situation and this may affect their behavior. If they believe the program is being run by a religious organization they may respond differently than if they believe it is run by a political group	
10. Experimenter expectancies Experimenters have expectations (e.g. about how men and women or different socio-economic groups will react to the program), and this may affect how they react to different groups.	
11. Novelty and disruption effects. Novel programs can generate excitement and produce a big effect. If a similar program is replicated the effect may be less as novelty has worn off.	

Attachment E (continued)	
12. Compensatory effects and rivalry. Programs create a dynamic that can affect outcomes in different ways. There may be pressures) to provide benefits to non-participants; comparison groups may become motivated to show what they can achieve on their own, or those receiving no treatment or a less attractive treatment may become demoralized	
13. Using indicators and constructs developed in other countries without pre-testing in the local context Many evaluations important theories and constructs from other countries and these may not adequately capture the local project situation. For example, many evaluations of the impacts of micro-credit on women's empowerment in countries such as Bangladesh have used international definitions of empowerment that may not be appropriate for Bangladeshi women.	
13. The process of "quantizing" (transforming qualitative variables into interval or ordinal variables) or "qualitizing" (transforming quantitative variables into qualitative) changes the nature or meaning of a variable in a way that can be misleading. One example of "quantizing" is to convert contextual variables (the local economic, political or organization context affecting each project location) into dummy variables to be incorporated into regression analysis.	
14. Does multi-level mixed-method analysis accurately reflect how the project operates and interacts with its environment.	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Attachment F. EXTERNAL VALIDITY [*Transferability*]	
Reasons why inferences about how study results would hold over variations in persons, settings, treatments and outcomes may be incorrect.	
1. Sample does not cover the whole population of interest subjects may come from one sex or from certain ethnic or economic groups or they may have certain personality characteristics (e.g. depressed, self-confident). Consequently it may be different to generalize from the study findings to the whole population.	
2. Different settings affect program outcomes. Treatments may be implemented in different settings which may affect outcomes. If pressure to reduce class size forces schools to construct extra temporary and inadequate classrooms the outcomes may be very different than having smaller classes in suitable classroom settings.	
3. Different outcome measures give different assessments of project effectiveness. Different outcome measures can produce different conclusions on project effectiveness. Micro-credit programs for women may increase household income and expenditure on children's education but may not increase women's political empowerment.	
4. Program outcomes vary in different settings. Program success may be different in rural and urban settings or in different kinds of communities. So it may not be appropriate to generalize findings findings from one setting to different settings	
5. Programs operate differently in different settings. programs may operate in different ways and have different intermediate and final outcomes in different settings. The implementation of community-managed schools may operate very differently and have different outcomes when managed by religious organizations, government agencies and non-governmental organizations.	
6. The attitude of policy makers and politicians to the program identical programs will operate differently and have different outcomes in situations where they have the support of policy makers or politicians than in situations where they face opposition or indifference. When the party in power or the agency head changes it is common to find that support for programs can vanish or be increased.	
7. Seasonal and other cycles. many projects will operate differently in different seasons, at different stages of the business cycle or according to the terms of trade for key exports and imports. Attempts to generalize findings from pilot programs must take these cycles into account.	
8. Are the characteristics of the sample of persons, settings, processes, etc. described in enough detail to permit comparisons with other samples?	
9. Does the sample design theoretically permit generalization to other populations?	
10. Does the researcher define the scope and boundaries of reasonable generalization from the study?	
11. Do the findings include enough "thick description" for readers to assess the potential transferability?	
12. Does a range of readers report the findings to be consistent with their own experience?	
13. Do the findings confirm or are they congruent with existing theory? Is the transferable theory made explicit?	
14. Are the processes and findings generic enough to be applicable in other settings?	
15. Have narrative sequences been preserved? Has a general cross-case theory using the sequences been developed?	
16. Does the report suggest settings where the findings could fruitfully be tested further?	
17 Have the findings been replicated in other studies to assess their robustness. If not, could replication efforts be mounted easily?	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Attachment G: UTILIZATION	
How useful were the findings to clients, researchers and the communities studied?	
1. Are the findings intellectually and physically accessible to potential users?	
2. Were any predictions made in the study and, if so, how accurate were they?	
3. Do the findings provide guidance for future action?	
4. Do the findings have a catalyzing effect leading to specific actions?	
5. Do the actions taken actually help solve local problems?	
6. Have users of the findings experienced any sense of empowerment or increased control over their lives? Have they developed new capacities?	
7. Are value-based or ethical concerns raised explicitly in the report? If not do some exist that the researcher is not attending to?	
8. Did the evaluation report reach the key stakeholder groups in a form that they could understand and use [Note: this question can be asked separately for each of the main stakeholders]?	
9. Is there evidence that the evaluation had a significant influence on future project design?	
10. Is there evidence that the evaluation influenced policy decisions?	
General comments on this component	
Ratings: 1 = Evaluation design or analysis is very strong; 5 = design or analysis has serious problems	

Appendix 2 Checklist for assessing the validity of reconstructed baseline data						
	The adequacy of reconstructed baseline data					
	Very strong	Quite strong	Adequate	Weak	Serious problems	Not applicable
A. Objectivity. ***Are the data and the assessments supported by the evidence and is the data collection relatively free of researcher bias?*** [*For example, if a rating scale indicates that when the project began "most community leaders have limited experience in running meetings", is this based on an analysis of a large number of communities or only a small number that the person completing the report has herself visited?*]						
B. Reliability. ***Were the data collection procedures consistent over time? Were the same procedures used in a consistent manner by all researchers?*** [*For example, was the definition of "low-income household" used in the same way in the baseline and each successive reporting period, and by all people preparing the monitoring reports?]*						
C. Credibility. ***Were the data collection methods, the data and interpretations used for reconstructing baseline and follow-up reporting credible to clients, stakeholders and the people studied?*** [*For example, did project management and other stakeholders find the estimates of male and female school enrolment rates at the start of the project to be credible?*]						
D. Construct validity. ***Did the indicators and measurements used to describe the impacts, outcomes, outputs and key contextual factors that were reconstructed adequately capture the complexity and multi-dimensionality of these constructs?*** [*For example, did the survey used to estimate household income at the start of the project include data on income from self-employment and rent as well as formal labor market earnings?*]						
E. Soundness of sampling and statistical computation procedures. ***Did the sample selection procedures adequately cover the target population, were the samples large enough for the required analysis, and were appropriate computation procedures used?*** [*For example, in cases where communities outside the official target areas took advantage of the village wells, did the reconstructed baseline sample for an impact evaluation adequately cover all potential beneficiaries (and not just those in the designated villages?*]						

F. [For impact evaluation design] The soundness of the sampling procedures for the reconstruction of the baseline control group. ***How well did the sample selection address the special challenges of control group reconstruction discussed in Chapter 5?*** [*For example, did the design address the issue of unobservables and try to find alternative ways to estimate their potential influence on project outcomes?*]						
G. Sustainability of monitoring data collection. ***Did the procedures used to reconstruct baseline data continue to generate data of the same quality for subsequent time periods for which data was missing?*** [*For example, do project staff have incentives to continue producing high quality monitoring reports? Are there quality control procedures in place? Do staff receive feedback and rewards for conscientious reporting?*]						

References

[A more comprehensive reference list is included in Bamberger, Rugh and Mabry. 2006. RealWorld Evaluation. Sage Publications]

Ahmed, Nizam. *Study on gender dimensions of the Second Bangladesh Rural Roads and Markets Improvement and Maintenance Project.* Available at www.worldbank.org/genddr/transport Click on "Grants and Pilots".

Aron, A & Aron, E. 2002. *Statistics for the Behavioral and Social Sciences*. Prentice Hall.

Bamberger, M (editor). 2000 *Integrating Quantitative and Qualitative Research in Development Projects*. Directions in Development Series. World Bank.

Brewer, J & Hunter, A. 2006. (eds) *Foundations of Multimethod Research. Synthesizing Styles*. Sage Publications.

Brown, J. 2000. "Evaluating the impact of water supply projects in Indonesia". In Bamberger, M (editor). *Integrating Quantitative and Qualitative Research in Development Projects*. Directions in Development Series. World Bank.

Creswell,J; C, V.L Clark; Guttman, M.L & Hanson, W. 2003. Advanced Mixed Method Research Designs. In Tashakkori, A & Teddue, C (Eds) *Handbook of Mixed Methods in Social and Behavioral Science*. (pp. 209-240). Thousand Oaks, CA: Sage

Meyers, L; Gamst, G & Guarino,A. 2006.*Applied Multivariate Research: Design and Interpretation.* Sage Publications.

Glewwe, P., Kremer, M., Sylvie Moulin & Zitzewitz, E. 2004. Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics* 74: 251-268). http://www.povertyactionlab.com/projects/project.php?pid=26

Kertzer, D & Fricke, T. 1997. "Toward and anthropological demography" In *Anthropological demography" Towards a new synthesis*. (ed) Kertzer and Fricke. University of Chicago Press.

Khandker, S. 1998. *Fighting Poverty with Microcredit: Experience in Bangladesh.* Oxford University Press.

Kozel, V & Parker, B. 2000. "Integrated approaches to poverty assessment in India" in Bamberger, M.(Ed) (2000). Integrating Quantitative and Qualitative Research in Development Projects. *Directions in Development*. (pp. 59-68). Washington D.C: The World Bank

Kumar, S. (2002). Methods for Community Participation. A Complete Guide for Practitioners. London. ITDG Publishing.

Mabry, l. 1998. "Case study methods." Pp. 155-70 in *Evaluation Research for Educational Productivity*. Edited by H.J. Walberg and A.J. Reynolds. Greenwich. CT. JAI Press.

Newman, C. 2001. Gender, time use, and change: impacts of agricultural export employment in Ecuador. *Policy Research Report on Gender and Development Working Paper Series No. 18*. Poverty Reduction and Economic Management Network/ Development Research Group. The World Bank. February 2001. Available on the web at www.worldbank.org/gender/prr.

Patton, M. Q. (1997). (Third Edition). *Utilization-focused evaluation*. Thousand Oaks. California: Sage Publications.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks. California: Sage Publication

Pradhan, Menno and Laura Rawlings. 2000. "The impact of targeting of social infrastructure investments: lessons from the Nicaraguan Social Fund." *World Bank Economic Review*. 16(2): 275-295.

Presser, P; Couper, M; Lessler, J; Martin, M; Martin, J; Rothgeb, J & Singer, E. Methods for Testing and Evaluating Survey Questionnaires. New York: Wiley

Presser, P; Couper, M; Lessler, J; Martin, M; Martin, J; Rothgeb, J & Singer, E. Methods for Testing and Evaluating Survey Questions (2004). Public Opinion Quarterly, Vol 68 No. 1 pp. 109-130, 2004.

Ravallion, M 2006. Evaluating anti-poverty programs. Handbook for Agricultural Economics (edited by Robert Evenson and Paul Schulz) Volume 4. North-Holland

Rietberger-McCracken, J. & Narayan, D. (1997). Participatory Rural Appraisal. Module III of *Participatory tools and techniques: a resource kit for participation and social assessment.* Environment Department. Washington D.C: The World Bank.

Roche, C. 1999. Impact Assessment for Development Agencies. Learning to Value Change. OXFAM**.**

Rugh, J. 1986. *Self-Evaluation: Ideas for Participatory Evaluation of Rural Community Development Projects*. Oklahoma City, OK: World Neighbors.

Schwarz, N & Oyserman, D. (2001). Asking Questions about Behavior: Cognition, Communication, and Questionnaire Construction. *American Journal of Evaluation*. Volume 22. No. 2 pp. 127-160.

Sirken, R. (1999). Statistics for the Social Sciences. Thousand Oaks. Sage Publications.

Tashakkori, A & Teddue, C (eds) 2003. *Handbook of Mixed Methods in Social and Behavioral Research.* Thousand Oaks, CA. Sage.

Vaughan, R. & Buss, T. 1994. *Communicating Social Science Research to Policymakers*. Applied Social Science Research Methods Series No. 48. Sage Publications.

Valadez, J. & Bamberger, M. 1994. Monitoring and evaluating social programs in developing countries: a handbook for policymakers, managers and researchers. Washington D.C. World Bank.

World Bank. Operations Evaluation Department. 2004. *Influential Evaluations.* World Bank. Operations Evaluation Department. 2005. *Influential Evaluations: Detailed Case Studies*. Available free at www.worldbank.org/ieg/ecd

World Bank. Independent Evaluation Group. 2006. Conducting quality impact evaluations under budget, time and data constraints. Available free at www.worldbank.org/ieg/ecd