APPENDICES

The appendices provide more detailed resource material for the chapters to which they refer. They include examples of checklists and indicators, evaluation tools and techniques, and also case studies providing examples of completed evaluations.

Many of the technical terms in these appendices are included in the Glossary in the book.

Contents

Number	Title
Chapter 2: First Clarify the Purpose: Scoping the Evaluation	
2.1	A Checklist to Assess the Evaluation Purpose and the Context Within Which It Will Be Implemented
2.2	Seven Basic Impact Evaluation Designs
2.3	Potential Methodological Weaknesses in Many Statistically Strong Evaluation Designs
2.4	Developing the Terms of Reference (Statement of Work) for the Evaluation
Chapter 3: Not Enough Money: Addressing Budget Constraints	
3.1	Factors Affecting the Sample Size
3.2	Threats to Adequacy and Validity Relating to Budget Constraints
Chapter 5: Critical Information Is Missing or Difficult to Collect: Addressing Data Constraints	
5.1	Working With Comparison Groups in Retrospective Evaluations
5.2	Using Recall Techniques in Retrospective Surveys
5.3	Challenges Collecting Baseline Data on a Comparison Group
5.4	Special Issues and Challenges Working With Comparison Groups
Chapter 7: Strengthening the Evaluation Design and the Validity of the Conclusions	
7.1	Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Quantitative (Experimental and Quasi-Experimental) Impact Evaluation Designs
7.2	Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Qualitative Impact Evaluation Designs
7.3	Integrated Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Impact Evaluation Designs (Standard Version)
7.4	Example of a Completed Threats-to-Validity Worksheet
7.5	Integrated Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Impact Evaluation Designs (Advanced Version)
7.6	Approaches for Assessing Validity of Mixed-Method Evaluations
7.7	Points During the RWE Cycle at Which Corrective Measures Can Be Taken
7.8	Factors Determining the Adequacy of an Evaluation Design and the Validity of the Findings
7.9	Examples of Other Checklists Used to Assess Evaluation Quality and Validity
Chapter 10: Theory-Based Evaluation and Theory of Change	
10.1	Results-Based Reporting and Logical Frameworks
10.2	The Two Components of a Program Theory Framework: Program Impact Models and Implementation Models
Chapter 11: Evaluation Designs: The RWE Strategy for Selecting the Appropriate Evaluation Design to Respond to the Purpose and Context of Each Evaluation	
11.1	A More Detailed Look at the RealWorld Evaluation Design Frameworks
11.2	The RWE Approach to the Classification of Factors Affecting the Choice of Evaluation Design

Number	Title
Chapter 11: Evaluation Designs: The RWE Strategy for Selecting the Appropriate Evaluation Design to Respond to the Purpose and Context of Each Evaluation	
11.3	The Strengths and Weaknesses of the Seven RWE Design Frameworks
11.4	Challenges Facing the Use of Experimental and Other Statistical Designs in RealWorld Evaluation Contexts
11.5	Examples of Randomized Control Trials
Chapter 12: Quantitative Evaluation Methods	
12.1	The Main Types of Questions That Can Be Included in Quantitative Surveys
12.2	Useful Sources of Secondary Data for QUANT Evaluations
12.3	Large-Scale Compilations of the Findings of Randomized Control Trials (RCTs)
12.4	Data Analysis for Quantitative Evaluations
Chapter 14: Mixed-Method Evaluation	
14.1	Two Case Studies Illustrating Different Ways in Which Mixed-Method Designs Can Strengthen Impact Evaluations
14.2	Characteristics of Quantitative and Qualitative Approaches to Different Stages of the Evaluation Process
14.3	Common Issues Affecting the Validity of Statistical Impact Evaluation Designs and How Mixed-Method (MM) Designs Can Help Address Them
14.4	Three Case Studies Illustrating the Use of Mixed-Method Evaluations
Chapter 15: Sampling Strategies for RealWorld Evaluation	
15.1	Using Power Analysis and Effect Size for Estimating the Appropriate Sample Size for an Impact Evaluation
Chapter 17: Gender Evaluation: Integrating Gender Analysis Into Evaluations	
Part 1: Specific Gender Evaluation Tools	
17.1	The Harvard Gender Analysis Framework
17.2	Tools for More In-Depth Gender-Responsive Evaluation (GRE) Designs
17.3	Two Examples of a Women's Empowerment Index
Part 2: General Approaches and Methodologies for Gender Analysis	
17.4	The World Bank/IEG Implementation Completion Report: Gender Flag Review
17.5	Recommended Structure for the Evaluation Cooperation Group GRE Reports
17.6	Widely Used Gender Indices and Checklists
17.7	Evaluation Approaches Used in Standard GRE Designs
17.8	Examples of the Application of the Different GRE Designs
17.9	Evaluating the Gender Dimensions of Complex Development Programs
17.10	The Contribution of Gender-Responsive Budgeting (GRB) to National Policy Dialogue
17.11	Feminist Critical Theory
17.12	Gender-Sensitive Data-Collection Methods and Applications Used in the Case Studies in Appendix 17.14 and the Additional Time and Cost Implications (Compared to Standard Evaluation Methods)

(Continued)

(Continued)

Number	Title
Part 3: Case Studies Illustrating Different Approaches to Gender Analysis	
17.13	Summary of Design and Data-Collection Methods Used in the Seven Case Studies Described in Appendix 17.14
17.14	Case Studies Illustrating Different Gender Impact Evaluation Methodologies
Chapter 19: Managing Evaluations	
19.1	The Evaluation Framework and Scope of Work
19.2	The Actors Involved in an Evaluation of How Gender Equality Issues Were Addressed in an International Development Program
19.3	Procedures for Contracting Evaluation Consultants
19.4	Guidelines for Strengthening Terms of Reference
19.5	Building in Quality Assurance Procedures
19.6	Evaluation Capacity Development

APPENDICES FOR CHAPTER 2 FIRST CLARIFY THE PURPOSE: SCOPING THE EVALUATION

2.1 A Checklist to Assess the Evaluation Purpose and the Context Within Which It Will Be Implemented

2.2 Seven Basic Impact Evaluation Designs

2.3 Potential Methodological Weaknesses in Many Statistically Strong Evaluation Designs

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

Chapter 2 provides guidelines for scoping the evaluation. This addresses stakeholder expectations concerning how the evaluation will be designed and used, and how to identify the information that will be required to address these expectations. Program theory is proposed as a tool for helping design the evaluation. The chapter also identifies real-world constraints on the range of available evaluation designs, and the range of available designs that can be used within these constraints. The final section provides guidelines for managers on how to prepare the terms of reference. The four appendices provide a checklist for helping define the evaluation purpose and the context within which it operates and within which it will be evaluated (Appendix 2.1); the seven basic evaluation designs from among which to select the most appropriate design (Appendix 2.2); the potential methodological weaknesses of randomized control trials (RCTs) and other "strong" statistical designs (Appendix 2.3); and guidelines for developing the terms of reference that will be used in the process of contracting consultants or for defining for an internal evaluation team the purpose of the evaluation, and the proposed evaluation design or the criteria to be used in designing the evaluation (Appendix 2.4).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 2.1 A CHECKLIST TO ASSESS THE EVALUATION PURPOSE AND THE CONTEXT WITHIN WHICH IT WILL BE IMPLEMENTED

This table identifies some of the characteristics and dimensions of the evaluand (thing to be evaluated) that need to be taken into consideration as decisions about the design of the evaluation are made.

A. Characteristics of the evaluand and the purpose and nature of the evaluation	
1. Basic purpose of evaluation[a]	a. Developmental (i.e., support innovative exploration of evolving approaches for addressing problems) b. Formative (i.e., learning and improvement of planned intervention during the implementation process in order to improve the process itself) c. Summative (i.e., accountability and judgment of the overall merit, worth, value, and significance of completed program; though this can feel like a postmortem, summative evaluation can inform major decisions about future programming)

(Continued)

(Continued)

A. Characteristics of the evaluand and the purpose and nature of the evaluation	
2. Other purposes of evaluation	a. Compliance with stated program design b. Impact: existing or potential achievement of higher-level outcomes (e.g., improved quality of life of intended beneficiaries) c. Adapting an intervention to a new context d. Adapting an existing program to a major change e. To help make resource allocation decisions on competing or best alternatives f. To help identify emerging problems and build consensus on the causes of a problem and how to respond g. To support public sector reform and innovation
3. Complexity of the evaluand[b]	a. Simple project: few intervention components, defined timeline b. Complicated program: sector program with various components, often combining several individual projects c. Complex program (e.g., general budget support or multiprogram interventions often involving several funding agencies and operating at the national or cross-country level or evolving situations such as natural disasters, violent conflict, or other dramatic changes) d. Overall assessment of the level of complexity (see Chapter 16)
4. The local and national context within which the evaluation will be implemented	a. Economic context b. Political context c. Policy, legal, and administrative context d. Organizations and agencies involved in the project e. Natural environment f. Characteristics and culture of the target population, politics, history, socioeconomic context, values, relative peace or conflict, needs, and interests of stakeholders
5. Geographic scale	a. Program or sector level (which could involve multiple countries) b. Multinational/regional (several countries) c. National (one country) d. Subnational region (e.g., district or province) e. One or a few local communities
6. Scale of intervention[c]	a. Small (e.g., less than 5,000 individuals or households) b. Medium (e.g., up to 50,000 units) c. Large (e.g., more than 50,000 units)
7. Size of the evaluation budget[d]	a. Small (e.g., less than 5% of program budget) b. Moderate (e.g., up to 15% of program budget) c. Generous (e.g., more than 15%—for example, a major purpose is research to test a new intervention)
8. When evaluation commissioned	a. Start of intervention (baseline/pretest) b. Midterm c. End of intervention (posttest) d. After intervention completed (ex-post)
9. Duration of the evaluation	a. Continues throughout intervention cycle b. Evaluation commissioned late in the intervention cycle but sufficient time is budgeted to conduct required data collection and analysis c. Great time pressure (the evaluation must be completed in weeks or a few months)

A. Characteristics of the evaluand and the purpose and nature of the evaluation	
10. Client	a. Donor agency b. Planning ministry c. Implementing agency d. Civil society or other
11. Who conducts the evaluation	a. Internal evaluator (or evaluation team) b. External consultant(s) (individual or team) c. Mixed team combining external and internal members
12. Definition of boundaries	a. Were evaluation boundaries clearly defined b. The level of stakeholder involvement in boundary definition c. Level of consensus on the definition of boundaries
B. Methodological dimensions	
13. Statistical rigor (client preference and what is feasible)	a. Statistically strong evaluation design b. Qualitative design c. Less rigorous, more descriptively focused design
14. Choice of the most suitable evaluation design (more than one design can be combined)	a. Experimental and quasi-experimental designs b. Theory-based designs c. Case study designs, including qualitative comparative analysis (QCA) d. Qualitative designs e. Systematic reviews f. Statistical designs g. Complexity-responsive design h. Big data analytic design
15. QUANT–QUAL preference	a. All or mostly quantitative methods and data b. All or mostly qualitative methods and evidence c. Appropriate mix of QUANT and QUAL
16. Data source(s)	a. Direct collection from units of study b. Secondary sources c. Appropriate mix of primary and secondary sources d. Incorporating big data and ICT-generated data

a. Adapted from Patton (2011). As summarized in Exhibit 2.2, pp. 44–46, and Exhibit 10.1, pp. 308–313) and Morra-Imas & Rist (2009, Box 1.1, p. 15), with additional categories added by the present authors to reflect other purposes of funding agencies and clients.

b. The concept of complexity is discussed in Chapter 15.

c. The concepts of "large" and "small" with respect to cost and scale are relative. What might be considered "small-scale" or "low-cost" by a major donor might be considered very large by a nongovernmental organization (NGO).

d. Though we give relative budget percentages for illustrative purposes, obviously the actual amount available for evaluation makes a significant difference on the kind of evaluation that can be undertaken.

APPENDIX 2.2 SEVEN BASIC IMPACT EVALUATION DESIGNS

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

7. Posttest (only) analysis of project group (no baseline or statistical comparison group)		X		P_1	

Notes:

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

(2) In this simplified table, the point at which data are first collected on the project group (P_1) is also the time at which the evaluation begins. In Table 11.3 (Chapter 11), we distinguish between evaluations that start at the beginning of the project (and in which baseline data are collected through primary data collection) and evaluations that start late in the project but in which baseline data are obtained from secondary sources or through the baseline reconstruction techniques discussed in Chapter 5.

(3) The project intervention is usually a process that occurs over time, that is, past the midterm to the end of the life of the project.

APPENDIX 2.3 POTENTIAL METHODOLOGICAL WEAKNESSES IN MANY STATISTICALLY STRONG EVALUATION DESIGNS

Many evaluation designs that are commonly referred to in the evaluation literature are, in fact, only strong with respect to their ability to control for sources of statistical selection bias. The reasons that they are statistically strong (e.g., randomization or statistical matching of samples, strict and inflexible rules concerning how data are collected, and administration of the same survey instrument to the same or equivalent samples before and after the project implementation) make these quantitative designs potentially weak in other respects, including the following:

- *Weak construct validity:* Many statistical designs are not based on a program theory model (although it is perfectly possible to incorporate a program theory model).
- *Analysis of project process:* There is no or insufficient analysis of the project implementation process.
- *No consideration of contextual variables:* Contextual variables that can explain differences in outcomes in different project locations are not discussed.
- *Mono-method bias:* Many quantitative designs collect all of their data from a single instrument, most commonly a structured questionnaire. This increases the risk of bias or incomplete information, as it is not possible to compare estimates obtained from different independent sources.
- *Difficulties in collecting information on sensitive topics:* Many quantitative approaches use data-collection methods such as structured questionnaires.
- *Inflexible evaluation design:* Inflexibility and difficulty in adapting the design to changes in the project design or the context in which it is implemented.

Two conclusions result from these potential methodological weaknesses in strong statistical designs. First, it is important, when discussing the merits of different evaluation designs, to always distinguish between "strong statistical designs" and "methodologically strong designs." While statistical evaluations can be designed to ensure all-round methodological strength, this is frequently not done, so that many statistically strong designs can be vulnerable in other ways. Similarly, it is possible to have qualitative or mixed-method designs that might be considered weak in terms of conventional quantitative terms but that may use designs that are methodologically sound in other ways. Second, it is almost always possible to strengthen the methodology of all evaluation designs—quantitative, qualitative, and mixed method—by incorporating the essential evaluation design components discussed in Table 2.3 in Chapter 2.

APPENDIX 2.4 DEVELOPING THE TERMS OF REFERENCE (STATEMENT OF WORK) FOR THE EVALUATION

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

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

APPENDICES FOR CHAPTER 3 NOT ENOUGH MONEY: ADDRESSING BUDGET CONSTRAINTS

3.1 Factors Affecting the Sample Size

3.2 Threats to Adequacy and Validity Relating to Budget Constraints

Chapter 3 discusses the challenges when conducting evaluations under budget constraints. Five strategies are presented for addressing these constraints while maintaining an acceptable level of methodological validity. The chapter concludes with a reference to the checklists presented in Chapter 7 and Appendices 7.1 to 7.6 for assessing the importance of different threats to validity and how they can be applied to assessing the validity of the different strategies proposed for working under budget constraints.

Two appendices are included for this chapter: a discussion of factors that influence decisions on the sample size (Appendix 3.1) and the identification of the threats to validity that can directly affect validity when adopting strategies to address budget constraints (Appendix 3.2).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 3.1 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, a two-tailed test is required (see Appendix 15.1).	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. Normally, a comparison group must also be selected and interviewed.	The sample size will be doubled if the same number of people must 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 (i.e., variability), the smaller the required sample.
5. The effect size	Effect size is the amount of increase the project is expected to produce (see Chapter 15, Section 4.2).	The smaller the effect size, the larger the required sample.
6. The efficiency with which the project is implemented	When project administration is poor, different individuals or groups may receive different combinations of services. The quality of the services can also vary. This makes it difficult to determine if lower-than-expected outcomes are due to poor project design or to the fact that many subjects are not receiving all intended services.	The poorer the quality and efficiency of the project, the larger the required sample.
7. The required level of disaggregation	In some cases, the client requires only global estimates of impact for the total project population. In other cases, it is necessary to provide disaggregated results for different project sites, for variations in the package of services provided, or for different socioeconomic groups (sex, age, ethnicity, etc.).	The greater the required disaggregation, the larger the sample.

(Continued)

(Continued)

Factor	Explanation	Influence on Sample Size
8. The sample design	Sampling procedures such as stratification can often reduce the variance of the estimates and increase precision.	Well-designed stratification may reduce sample size.
9. The level of statistical precision	"Beyond a reasonable doubt" is usually defined as meaning there is less than a 1 in 20 possibility that an impact as large as this could have occurred by chance (defined as the "0.05 confidence level"). If less precise results are acceptable, it is possible to reduce sample size by accepting a lower confidence level—for example, a 1 in 10 possibility that the result occurred by 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 (see Chapter 15, Section 4.4 and Appendix 15.1).	The higher the required power level, the larger the sample.
11. Finite population correction factor	The finite population correction factor reduces the required sample size by the proportion that the sample represents of the population (see Appendix 15.1).	The greater the proportion the sample represents of the total population, the smaller the sample.

APPENDIX 3.2 THREATS TO ADEQUACY AND VALIDITY RELATING TO BUDGET CONSTRAINTS

NOTE: Appendices 7.1, 7.2, and 7.3 present sets of checklists for assessing threats to validity for quantitative, qualitative, and mixed-method designs. In this appendix we use Appendix 7.3 to illustrate how budget constraints may affect the validity of mixed-method designs, as this is the most general case. Interested readers can refer to Appendices 7.1 (quantitative designs) and Appendix 7.2 (qualitative designs). The present appendix only includes the items in the checklists that are most likely to be affected by budget constraints.

Source: Appendix 7.3 Integrated Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Impact Evaluation Designs (Standard Version)

Part 3 Checklists Used to Assess the Different Components of the Threats to Validity

Checklist 2. Internal Design Validity (Reliability and Dependability)

1. *How context rich and meaningful ("thick") are the descriptions?* Budget pressures often reduce the richness of the data collected.
3. *Did triangulation among complementary methods and data sources produce generally converging conclusions?* Budget constraints often reduce the use of triangulation because the application of different data-collection methods usually increases costs.
5. *Are areas of uncertainty identified?* Was negative evidence sought, found? Budget pressures can reduce the search for negative evidence.
8. *Were data collected across the full range of appropriate settings, times, respondents, and so on?* Budget pressures frequently result in the elimination of some groups—often, the most difficult to reach.
16. *History.* Budget pressures often constrain ability to control for historical differences between project and comparison areas.
23. Use of less rigorous designs due to budget and time constraints.

Checklist 3. Statistical Conclusion Validity

1. *The sample is too small to detect program effects.* Budget pressures often result in the sample size being reduced below the minimum size required to satisfy power analysis criteria (see Chapter 15, Section 4.6).
5. *Restriction of range and extrapolation from truncated/incomplete database.* Time pressures sometimes result in samples or secondary data with more limited coverage.

Checklist 4. Construct Validity

3. *Use of a single method to measure a complex construct (monomethod bias).* Budget pressures may limit the number of data-collection methods or the number of independent indicators of key variables.
12. *Using indicators and constructs developed in other countries without pretesting in the local context.* Budget pressures often result in inadequate testing and customization of instruments.

Checklist 5. External Validity, Transferability, and Fittingness

1. *Sample does not cover the whole population of interest.*
7. *Do seasonal and other cycles affect implementation or outcomes.* These are often not adequately addressed when budget is a factor.
9. *Does the sample design theoretically permit generalization to other populations?* Simplifying sample design to save time can sometimes reduce representativity of the sample.

APPENDICES FOR CHAPTER 5 CRITICAL INFORMATION IS MISSING OR DIFFICULT TO COLLECT: ADDRESSING DATA CONSTRAINTS

5.1 Working With Comparison Groups in Retrospective Evaluations

5.2 Using Recall Techniques in Retrospective Surveys

5.3 Challenges Collecting Baseline Data on a Comparison Group

5.4 Special Issues and Challenges Working With Comparison Groups

Chapter 5 discusses the challenges of conducting evaluations when there are constraints on access to data. Strategies are discussed for addressing a number of scenarios: when the evaluation begins and the start of the project or evaluations that are commissioned toward the end of project implementation or ex-post, usually two to five years after the implementation is completed; collecting information on sensitive topics or difficult-to-reach groups; or when it is difficult to find data to construct a comparison group. Reference is also made to the challenges of ensuring that vulnerable groups, who are more expensive and difficult to reach, are included in the evaluations—something that is important for the evaluation of the Sustainable Development Goals.

Four appendices are included: working with comparison groups in retrospective evaluations (Appendix 5.1), using recall techniques in retrospective surveys (Appendix 5.2), challenges collecting baseline data on a comparison group (Appendix 5.3), and special issues and challenges working with comparison groups (Appendix 5.4).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 5.1 WORKING WITH COMPARISON GROUPS IN RETROSPECTIVE EVALUATIONS

Case 1. An evaluation was conducted in Guayaquil, Ecuador, to assess the impact of the cut-flower export industry (which employs a high proportion of women and pays women well above average wages) on women's income and employment and on the division of domestic tasks between husband and wife. Families living in another valley about 100 miles away and without access to the cut-flower industry were selected as a comparison group. This was a nonequivalent control group because families were not randomly assigned to the project and comparison groups. The project and comparison groups were interviewed after the flower industry had been operating for some time and, consequently, no baseline data were available. Multivariate analysis was used to determine whether there were differences in the dependent variables (women's employment and earnings and the number of hours spent by husband and wife on domestic chores) in the project and comparison areas after controlling for household attributes such as educational level of both spouses, family size, and so on. Significant differences were found between the two groups on each of these dependent variables, and it was concluded that there was evidence that access to higher-paid employment (in the cut-flower industry) did affect the dependent variables (the distribution of domestic chores between men and women and the hours women and men devoted to paid and nonpaid work). Although multivariate analysis matched the project and comparison groups more closely, it was not able to examine differences in the initial conditions of the two groups before the project began. For example, it is possible that the cut-flower industry decided to locate in this particular valley because it was known that a high proportion of women already worked outside the home and that husbands were prepared to assume more household chores, thus allowing their wives to work longer hours. The analytical model used in the study was not able to examine this alternative explanation.

Source: Newman (2001).

Case 2. An ex-post evaluation was conducted of the impact of microcredit on women's savings, household consumption and investment, and fertility behavior in Bangladesh. The evaluation used household survey data from communities that did not have access to credit programs as a nonequivalent control group (comparison group). Multivariate analysis was used to control for household attributes, and it was found that women's access to microcredit programs was significantly associated with most of the dependent variables. As in the case of the Ecuador study, this design did not control for existing differences between the project and comparison groups with respect to important explanatory variables such as women's participation in small-business training programs or prior experience with microcredit.

Source: Khandker (1998).

APPENDIX 5.2 USING RECALL TECHNIQUES IN RETROSPECTIVE SURVEYS

Information on baseline conditions of the project and comparison groups is often collected by conducting a retrospective survey when the project is nearing completion or has ended. The survey asks respondents to provide information on their situation, attitudes, or knowledge at the time the project began or at some other relevant point in the past.

Recall techniques are widely used in research areas such as poverty analysis, demography, and income and expenditure studies. The challenge with recall studies is that reference data are usually not available to assess the reliability of estimates and the direction and magnitude of bias. Although it is always difficult to use recall to collect precise numerical data such as income, incidences of diarrhea, or farm prices, it can be used to obtain estimates of major changes in the welfare conditions of the household. For example, families can usually recall which children attended a school outside the community before the village school opened, how children traveled to school, and travel time and cost. Families can also often provide reliable information on access to health facilities, where they previously obtained water, how much water they used, and how much it cost. On the other hand, families might be reluctant to admit that their children had not been attending school or that they had been using traditional medicine. They might also wish to underestimate how much they had spent on water if they are trying to convince planners that they are too poor to pay the water charges proposed in a new project.

Two common sources of bias in recall of expenditure data have been identified. First, the underestimation of small and routine expenditures increases as the recall period increases. Second, there is a "telescoping" of recall concerning major expenditures (such as the purchase of a cow, bicycle, home, car, or item of furniture), so that the time frame of expenditures may be misreported. Although most of the research on recall bias in income and expenditures comes from studies such as the Consumer Expenditure Surveys[1] conducted quarterly by the Bureau of Labor Statistics of the U.S. Department of Labor, the general results are potentially relevant to developing countries. The Living Standards Measurement Survey[2] (LSMS) program has conducted studies on the use of recall for estimating consumption in developing countries (Deaton & Grosh, 2000), which are discussed below.

The most systematic assessments of the accuracy of recall data in developing countries probably come from demographic studies on the reliability of reported information on contraceptive usage and fertility. A number of large-scale comparative studies such as the World Fertility Survey[3] provide national surveys using comparable data-collection methods for different points in time. For example, similar surveys were conducted in the Republic of Korea in 1971, 1974, and 1976, each of which obtained detailed information on current contraceptive usage and fertility as well as obtaining detailed historical information based on recall for a number of specific points in the past. This permitted a comparison of recall in 1976 for contraceptive usage and fertility in 1974 and 1971, with exactly the same information collected from surveys in those two earlier years. It was found that recall produced a systematic underreporting, but that the underestimation could be significantly reduced through the careful design and administration of the surveys (Pebley, Noreen, & Choe, 1986). Similar findings are available from demographic analysis in other countries. The conclusion from these studies is that recall can be a useful estimating tool with predictable and, to some extent, controllable errors. Unfortunately, it is possible to estimate the errors only where large-scale comparative survey data are available, and there are few if any other fields for developing countries with a similar wealth of comparative data to that available from demographic studies.

A major challenge in using recall is that estimates are very sensitive to changes in the research design methods, particularly the method used for data collection, the period over which estimates are obtained, and how the questions are formulated.[4]

[1] The Consumer Expenditure Surveys combine a quarterly survey of expenditures administered to a sample of U.S. households and the completion by a smaller sample of households of a diary in which all expenditures are recorded. The diary is used to check the reliability and sources of bias in the estimates, based on recall obtained from the quarterly surveys. For more information, go to www.bls.gov/cex/home.htm.

[2] The LSMS (Living Standards Measurement Survey) program was launched in the 1980s by the World Bank to develop standard survey methodologies and questionnaires for comparative analysis of poverty and welfare in developing countries (Grosh & Glewwe, 2000).

[3] The World Fertility Survey is based on demographic and fertility surveys conducted in 41 developing countries in the 1970s and 1980s by many different agencies. For more information, go to https://www.k4health.org/toolkits/info-publications/world-fertility-survey-current-status-and-findings.

[4] The references in this and the following paragraph are taken from a chapter by John Gibson (2006), "Statistical Tools and Estimation Methods for Poverty Measures Based on Cross-Sectional Household Surveys," in *Handbook on Poverty Statistics* (United Nations Department of Economic and Social Affairs). Available online at https://unstats.un.org/unsd/methods/poverty/chapters.htm.

The following examples illustrate the design sensitivity. For example, a study in Latvia found that, on average, household expenditures were 46% higher when respondents were asked to keep a record of expenditures in a diary compared with when they were asked to recall expenditures (K. Scott & Okrasa, 1998). In El Salvador, estimated expenditures were 31% higher when respondents were asked to provide detailed expenditures for 75 food categories and 25 nonfood categories compared with when they were asked to provide less-detailed information covering 18 food items and 6 nonfood items (Jolliffe, 2001). A study in Ghana found that average daily expenditures on a group of frequently purchased items fell by an average of 2.9% for every additional day over which respondents were asked to recall expenditures. The recall error leveled off at about 20% after two weeks (C. Scott & Amenuvegbe, 1991). One of the best-known studies on the sensitivity of expenditures to the recall period comes from India. Between 1989 and 1998, the National Sample Survey in India experimented with different recall periods for measuring expenditures. It was found that when the 30-day recall period for food items was replaced with a 7-day period, the total estimated food expenditures increased by around 30%. When at the same time the 30-day recall period for infrequent expenditures was replaced with a one-year recall, the estimated total expenditure increased by about 17% (Deaton, 2005).

Interestingly, a number of studies suggest that recall can provide better estimates of behavioral changes in areas such as primary prevention programs for child abuse, vocational guidance, and programs for delinquents than conventional pre- and posttest comparisons based on self-assessment (Pratt, McGuigan, & Katzev, 2002). This is because before entering a program, subjects often overestimate their behavioral skills or knowledge through a lack of understanding of the nature of the tasks being studied and of the required skills. After completing the program, they may have a better understanding of these behaviors and may be able to provide a better assessment of their previous level of competency or knowledge and how much these have changed. Self-assessment of poverty, empowerment, or community organizational capacity in developing countries might all be areas in which the *response shift* concept could potentially be applied for reconstruction of baseline data (Schwarz & Oyserman, 2001).

Another potentially useful approach concerns the use of calendars and time diaries to help respondents reconstruct past events or continuously changing activities. These methods are often referred to as life course research (Belli, Stafford, & Alwin, 2009). Both of these methods encourage respondents to incorporate temporal changes as clues in reporting events such as their parental status, childhood experiences, schooling, marriages, residences, relationships, wealth, work, stressors, health conditions, levels of happiness, what they have taught others, and how they have spent their time during the past day. According to Belli et al. (2009), these methods have "shown the ability to provide data of remarkably high quality in fields as diverse as life cycle consumption, training, labor supply, psychological development and adaptation to the environment, age stratification and life course, evolutionary theories of aging, and demographic models."[5] Alwin (2009) shows how statistical models such as the Latent Markov Model can be used to assess the reliability of event data (for example, whether a person was or was not married or was or was not working at a particular point in time) and continuous variables.

BOX A5.2-1

Examples of Retrospective Surveys Used to Reconstruct Baseline Data

Case 1. In Bangalore, India, in 1999, a sample of households was asked to respond to a "citizen report card" in which they assessed the changes in the quality of delivery of public services (water, sanitation, public hospitals, public transport, electricity, phones, etc.) since the project started in 1993. Families reported that although the quality of services was still low, on average, there had been an improvement in most services with respect to helpfulness of staff and proportion of problems resolved. The use of recall was an economical substitute for a baseline study.

Source: Paul (2002) and Operations Evaluation Department (2005).

Case 2. At the time that an evaluation of the impact of social funds in Eritrea was commissioned, the program had already been under way for several years. Baseline conditions for access to health services were estimated

(Continued)

[5] For references to each of these areas of research, see Belli et al. (2009, p. 2).

(Continued)

by asking families how frequently they had used health services before the village clinic was built, how long it took to reach these facilities, the costs of travel, and the consequences of not having had better access. The information provided by the households was compared with information from health clinic records and key informants (nurses, community leaders, etc.) so as to strengthen the estimates through triangulation. While existing documents (secondary data) were useful, it was often found that the records were not organized in the way required to assess changes and impacts. For example, the village clinics kept records on each patient visit but did not keep files on each patient or each family, so it was difficult to determine how many different people used the clinic each month/year and also the proportion of village families who used the clinic. Similar methods were used to reconstruct baseline data on village water supply and rural roads and transport for the evaluation of the water supply and road construction components.

Source: Based on unpublished local consultant impact evaluation report.

Case 3. The Operations Evaluation Department (OED) of the World Bank conducted an ex-post evaluation of the social and economic impacts of a resettlement program in Maharashtra State, India. Baseline data had been collected by project administrators on all families eligible to receive financial compensation or new land, but information was not collected on the approximately 45% of families who had been forced to move but who were not entitled to compensation. A tracer study was conducted by OED in which families forced to relocate without compensation were identified through neighbors and relatives. A significant proportion of families were traced in this way, and these were found on average to be no worse off as a result of resettlement, but it was not possible to assess how representative they were of all families relocated without compensation.

Source: World Bank (1993).

APPENDIX 5.3 CHALLENGES COLLECTING BASELINE DATA ON A COMPARISON GROUP

Comparison groups are communities, organizations, or groups selected to match the project communities as closely as possible on social, economic, physical, historical, and other characteristics relevant to the study. There are several additional sets of challenges for collecting baseline data on these groups. First, people who are not expecting to receive any benefit from the project have less incentive to cooperate and to provide information. Second, many of the people in the comparison group may be difficult to identify and to locate. This is particularly true when this population may include landless or other marginal groups. Third, there are a number of challenges to select a comparison group that has similar characteristics to the project population. The process of selecting well-matched comparison groups is often a challenging task. As discussed in Chapter 5, one of the major challenges in selecting the comparison group sample is that unless random selection is used, project beneficiaries are usually not a representative sample of the total target population. The most common beneficiary selection procedures are either self-selection or *administrative selection* by the implementing agency. When self-selection procedures are used, individuals or communities who apply are likely to include a higher-than-average proportion of subjects likely to succeed (e.g., people with no business experience are less likely to apply for a microloan, and communities with no organizational experience are less likely to apply for a community infrastructure project that requires the community to participate in implementation or maintenance). When *administrative selection* procedures are used, with few exceptions, beneficiaries or beneficiary communities are selected *purposefully* to target, for example, the poorest areas or those with the greatest development potential rather than selected *randomly*, increasing the challenge of identifying matched local groups.

Given the fact that only a small minority of projects use random selection procedures, the evaluator will almost always be faced with the likelihood that there will be systematic differences between the characteristics of the project and comparison-group populations. This means that some of the post-project differences found between the two groups may be due to existing differences (business loan recipients already had experience in running a business, parents who enroll children in after-school activities may be better educated or more motivated, and communities that participate in infrastructure projects may have greater organizational capacity than those that do not). So the practical challenge for the evaluator is how to match the project and comparison groups as closely as possible to try to eliminate these differences, or at least to understand what the differences are and to take them into account in the analysis.

There are two main RealWorld Evaluation (RWE) scenarios concerning the selection of a comparison group. The first is where good secondary survey data are available (see earlier discussion) or where resources are available to conduct a sample survey with a sufficiently large sample (see Chapter 15 for a discussion of sample size). In these cases, it is possible to use *statistical matching* (see Chapters 11 and 12) of the project and comparison groups to adjust for the effect of multiple variables (income, education, household size, educational level of parents, plot size, distance from the nearest town, etc.). However, as we will discuss, it is never possible to match the two samples on all possible factors that might affect outcomes, so there is always the question of how much outcomes might be affected by the variables, known as *omitted variables* or *unobservables*, that are not studied.

Be careful not to make the assumption that a comparison group is matched with the project groups in every respect except participation in the project. Rarely if ever in society are all factors equal between any two communities or groups. Many contextual and other factors must be considered in a holistic analysis of results, from which the evaluator should then try to determine the relative influence of the project's interventions compared with different internal and external factors in the project or the comparison group.

The second scenario is one in which well-matched comparison groups are not available, either from secondary data or sample surveys, and *judgmental matching* must be used. While research journals and quantitatively oriented textbooks tend to focus on cases in which statistical matching could be used (many journals do not publish studies where less rigorous procedures were used), unfortunately, in the real world, most evaluations probably use judgmental matching. Box 5.8 in Chapter 5 gives examples of relatively strong and relatively weak comparison groups (see also Box A5.3-2 later in this chapter).

Judgmental matching involves the pooling of all available sources of information to select a group of individuals, communities, or organizations (schools, health clinics, etc.) that match the project group as closely as possible. Often this will

involve combining the opinion of experts and key informants, review of maps and secondary data on different communities, and, where possible, diagnostic studies or at least visits to possible communities. In Chapter 16, we discuss the use of *concept mapping* as a technique through which statistical procedures are used to synthesize the opinions of large numbers of experts or stakeholders to select the comparison group.

Statistical matching techniques (see also Chapter 12). Evaluations frequently compare the project population with comparison areas selected to match the project population as closely as possible. When, as is usually the case, subjects were not randomly assigned to the two groups, this is called a nonequivalent control group or comparison group. In some cases, the comparison group may seem to match the project group quite closely on most of the socioeconomic characteristics of the households or individuals, but in other cases, there may be important differences between the two groups (see Chapter 5, Box 5.8). When relatively large and reasonably random samples have been interviewed in both groups, it is usually possible to strengthen the analysis by statistically matching subjects from two areas on a number of relevant characteristics such as education, income, and family size. The evaluations of the Ecuador cut-flower export industry and the Bangladesh microcredit programs described in Box A5.3-1 are examples of this approach.

If differences in the dependent variables (the number of hours men and women spend on household tasks, men's and women's savings and expenditures on household consumption goods, etc.) between the project group and the comparison group are still statistically significant after controlling for these household characteristics, this provides preliminary indications that the differences may be due, at least in part, to the project interventions.

Although this type of multivariate analysis is a powerful analytical tool, one important weakness is that, without baseline studies of both groups, the evaluation design does not provide any information on the initial conditions or attributes of the two groups prior to the project intervention. For example, the higher savings rates of women in the communities receiving microcredit in Bangladesh might be due to their having previously received training in financial management or to the fact that they already had small-business experience. These comparison-group designs can be strengthened by incorporating some of the methods discussed above for reconstructing baseline data. Using these methods, it is possible to assess the similarities and differences between the two groups at the time the project began. If the two groups are found to be relatively similar on key baseline indicators (socioeconomic characteristics, access to the kinds of services or benefits provided by the project), then this increases the likelihood that statistical differences found in the ex-post comparison are due at least in part to the project. If, on the other hand, there were important initial differences between the two groups in the reconstructed baselines, then it is harder to assume that the posttest differences are necessarily due to the project intervention. How effective the statistical controls are will depend on the adequacy of the control model and the reliability of the measurement of the control variables (Shadish, Cook, & Campbell, 2002, pp. 138, 249).

BOX A5.3-1

Working With Comparison Groups in Ex-Post Evaluation Designs

Case 1. An evaluation was conducted in Guayaquil, Ecuador, to assess the impact of the cut-flower export industry (which employs a high proportion of women and pays women well above average wages) on women's income and employment and on the division of domestic tasks between husband and wife. Families living in another valley about 100 miles away and without access to the cut-flower industry were selected as a comparison group. This was a nonequivalent control group because families were not randomly assigned to the project and comparison groups. The project and comparison groups were interviewed after the cut-flower industry had been operating for some time and, consequently, no baseline data were available. Multivariate analysis was used to determine whether there were differences in the dependent variables (women's employment and earnings and the number of hours spent by husband and wife on domestic chores) in the project and comparison areas after controlling for household attributes such as educational level of both spouses, family size, and so on. Significant differences were found between the two groups on each of these dependent variables, and it was concluded that there was evidence that access to higher-paid employment (in the cut-flower industry) did affect the dependent variables (the distribution of

domestic chores between men and women and the hours women and men devoted to paid and nonpaid work). Although multivariate analysis matched the project and comparison groups more closely, it was not able to examine differences in the initial conditions of the two groups before the project began. For example, it is possible that the cut-flower industry decided to locate in this particular valley because it was known that a high proportion of women already worked outside the home and that husbands were prepared to assume more household chores, thus allowing their wives to work longer hours. The analytical model used in the study was not able to examine this alternative explanation.

Source: Newman (2001).

Case 2. An ex-post evaluation was conducted of the impact of microcredit on women's savings, household consumption and investment, and fertility behavior in Bangladesh. The evaluation used household survey data from communities that did not have access to credit programs as a nonequivalent control group (comparison group). Multivariate analysis was used to control for household attributes, and it was found that women's access to microcredit programs was significantly associated with most of the dependent variables. As in the case of the Ecuador study, this design did not control for existing differences between the project and comparison groups with respect to important explanatory variables such as women's participation in small-business training programs or prior experience with microcredit.

Source: Khandker (1998).

Judgmental Matching

The following are some of the strategies used for constructing comparison groups when statistical matching is not possible:

- It is sometimes possible to construct an *internal* comparison group within the project area. Households or individuals who did not participate in the project or who did not receive a particular service or benefit can be treated as the comparison for the project in general or for a particular service.[6]

- When projects are implemented in phases, it is also possible to use households selected for the second or subsequent phases as the comparison group for the analysis of the impacts of the previous phase. This is sometimes called a pipeline comparison group. For example, the economic status of a new cohort of women about to receive their microfinance loans might serve as a comparison group to compare with the current economic status of women who received loans during the past year. The phased approach was used in the evaluation of the Tondo Foreshore Squatter Upgrading Project in Manila (see following section).

- The evaluation may be able to take advantage of *natural experiments*. For example, the start of the project in one area may be delayed due to administrative or other problems, so this area can be used as a comparison for the areas where it did start on time. Sometimes resources do not permit all areas to receive all services (for example, there may not be sufficient textbooks or specially trained teachers, so some of the treatments may not be implemented in all areas; in malaria treatment programs, it is sometimes the case that supplies of tablets or bednets run out so that some families only receive the orientation talks, others receive tablets but not bednets, and some receive the complete treatment). In other cases, another agency might provide different services in some areas (for example, the project might provide school meals and, by chance, another agency might provide school transport to some of the schools but not to others). If the evaluation has the flexibility to adapt the design, and if information is available sufficiently quickly to know about the changes, it may be possible to conduct *natural experiments* to compare the intended evaluation model with these different situations. Care must be taken in the analysis of natural experiments, as it will often be the case that the communities or schools that do not receive the full treatment may be the poorest or most remote, or the areas where other donors provide additional services might be the better-off or more accessible areas. So the evaluation must always try to understand and correct for these differences.

[6] For example, subjects may be categorized according to their distance from a road or water source, whether any family member attended literacy classes, the amount of food aid they received, and so on. This is sometimes called *intensity analysis*.

BOX A5.3-2

Potentially Strong and Weak Comparison Groups

The following three cases are examples of relatively strong comparison groups:

- In a community water-supply project in Bolivia, the number of communities that applied to obtain water far exceeded the resources available to construct water systems in that particular year. Successful communities were selected through a lottery so that the process could be seen to be transparent and unbiased.
- In the Tondo Foreshore Slum Upgrading Project in Manila, the project was designed to cover a population of more than 100,000 households in several phases over a period of 10 years. The areas to be included in Phase 2 were selected as a comparison group for Phase 1.
- In a low-cost housing project in El Salvador, all the project participants came from one of three distinct types of low-income settlements, and participants represented a relatively small proportion of the population in each of these areas. Although participants were self-selected, so that it was difficult to control for the effect of motivation, it was possible to randomly select a comparison sample from these three low-income settlements. Statistical analysis found the characteristics of the project and comparison groups were similar but not identical.

The following two cases describe situations in which it was more difficult to select a strong comparison group:

- An evaluation was conducted in Nairobi to evaluate the impacts of slum-upgrading programs that had been operating for a decade or longer. The programs had covered all the major slums that housed well over 75% of the low-income population. The slums not covered by these programs were very small, housing only a few hundred families (compared with some project areas with more than 50,000 households). All the potential comparison areas had special characteristics (such as a unique ethnic group) that distinguished them in potentially important ways from the project areas.
- The project sites to be included in an agricultural extension program in Ethiopia were selected to include the poorest and most remote rural communities and also to include only areas in which no other agencies were working. The selection process also meant that many of the project areas had unique ethnic characteristics. In addition to the difficulties of finding areas with similar characteristics, most other areas had at least one outside agency involved, making it very difficult to find suitable comparison areas.

APPENDIX 5.4 SPECIAL ISSUES AND CHALLENGES WORKING WITH COMPARISON GROUPS

Approach	Sources	Comments/Issues
Identifying and reconstructing comparison groups	Government statistics, earlier surveys, and 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 phase of project as "pipeline" comparison • Internal comparison groups when different participants receive different combinations of services • Appropriateness of potential comparison groups • Statistical matching of samples (e.g., propensity scores)
Special issues in reconstructing data on comparison groups; 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)	• Econometric analysis posttest project and control groups (this design cannot control for historical differences between the two groups—see Chapter 11)	Methodological issues include • Self-selection of participants (difficult 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 (in both cases, it is difficult to find a comparison group with similar characteristics)
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)	• Participant observation • Focus groups • Unstructured interviews • Observation • PRA techniques • Case studies • Key informants • Observation (participant and nonparticipant) • Informants from the groups • Self-reporting • Tracer studies and snowball samples • Key informants • Existing documents (secondary data) • Symbols of group identification (clothing, tattoos, graffiti)	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

APPENDICES FOR CHAPTER 7 STRENGTHENING THE EVALUATION DESIGN AND THE VALIDITY OF THE CONCLUSIONS

7.1 Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Quantitative (Experimental and Quasi-Experimental) Impact Evaluation Designs

7.2 Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Qualitative Impact Evaluation Designs

7.3 Integrated Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Impact Evaluation Designs (Standard Version)

7.4 Example of a Completed Threats-to-Validity Worksheet

7.5 Integrated Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Impact Evaluation Designs (Advanced Version)

7.6 Approaches for Assessing Validity of Mixed-Method Evaluations

7.7 Points During the RWE Cycle at Which Corrective Measures Can Be Taken

7.8 Factors Determining the Adequacy of an Evaluation Design and the Validity of the Findings

7.9 Examples of Other Checklists Used to Assess Evaluation Quality and Validity

Chapter 7 discusses ways to strengthen the evaluation design and to assess the validity of findings and conclusions. The quantitative, qualitative, and mixed-method evaluation literature have different ways of thinking about validity and use different terminology and indicators. In this chapter we draw on the mixed-method literature to develop a framework that can be applied to both quantitative and qualitive evaluation designs. Separate checklists are developed for assessing validity of both quantitative and qualitative designs, and then an integrated checklist is presented that combines elements of both the quantitative and qualitative approaches and which can be used with designs that combine elements of both approaches (i.e., mixed-method designs). Nine appendices are included. Appendices 7.1, 7.2, and 7.3 present worksheets and checklists for assessing the validity of the findings and recommendations of quantitative, qualitative, and mixed-method designs, respectively. Appendix 7.4 then gives an example of a completed worksheet. While Appendix 7.3 presents a basic worksheet for evaluating mixed-method designs (essentially combining indicators directly from the quantitative and qualitative worksheets given in Appendix 7.1 and 7.2), Appendix 7.5 presents a more advanced worksheet for mixed-method designs that incorporates theoretical and methodological concepts from the mixed-method field, and Appendix 7.6 synthesizes the different approaches presented for assessing the validity of mixed-method evaluation. Appendix 7.7 then identifies the different points in the evaluation cycle when corrective measures can be taken to address threats to validity, and Appendix 7.8 summarizes factors determining the validity of an evaluation. Finally, Appendix 7.9 identifies a number of checklists identified by other authors and agencies for assessing evaluation validity.

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 7.1 WORKSHEET FOR ASSESSING THREATS TO THE VALIDITY OF THE FINDINGS AND RECOMMENDATIONS OF QUANTITATIVE (EXPERIMENTAL AND QUASI-EXPERIMENTAL) IMPACT EVALUATION DESIGNS[1]

Part 1: Summary of the Findings of the Assessment of the Evaluation

1. Name of the project/program

2. Who conducted the evaluation? (indicate organizational affiliation)

3. Who conducted this validity assessment? (indicate organizational affiliation)

4. At what stage of the project/program did the evaluation begin?

5. At what stage of the evaluation was this assessment conducted?

6. Reason for conducting the threats-to-validity assessment

7. Summary of findings of the assessment with respect to the evaluation methodology (note whether the assessment included the use of the checklists in Part 3).

8. (If methodological problems were identified) What are the implications of these problems for the utilization of the evaluation findings and recommendations?

9. Recommended follow-up actions

1. *Source:* The checklists in Part 3 are adapted from Shadish, Cook, & Campbell (2002, Tables 2.2, 2.4, 3.1, and 3.2). Additional material, including the proposed rating systems, has been included in the checklists, and the present authors are solely responsible for the adaptation and interpretation of the data.

Part 2: Summary Assessment for Each Component

A. Threats to Internal Design Validity: Reasons why inferences about a causal relationship between two variables may be incorrect

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:[2]
- The number of methodological problems that could affect the utilization of the evaluation:

B. Threats to Statistical Conclusion Validity: Reasons why inferences about a statistical association between two variables (e.g., project treatment and outcome) may be incorrect

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

2. Note on ratings: 1 = the design and implementation of the methodology for this component are sound and there are no problems or issues; 5 = the design and/or implementation of the methodology for this component are weak, and many issues could affect the validity of the findings and recommendations.

C. Threats to Construct Validity: Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

D. Threats to External Validity: 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 (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

Part 3: Checklists Used to Assess the Four Components Describing Potential Threats to the Adequacy and Validity of a Quantitative Impact Evaluation

NOTE: Part 3, which involves a more in-depth, technical, and expensive assessment, is normally used only for large, high-priority evaluations or where the controversial nature of the topic requires a high level of credibility of the assessment. For these reasons, Part 3 would normally be conducted by an external consultant.

Dimension 1: Internal Validity

- Checklist 1: Threats to Internal Design Validity
- Checklist 2: Threats to Statistical Conclusion Validity
- Checklist 3: Threats to Construct Validity

Dimension 2: External Validity

- Checklist 4: Threats to External Validity

Checklist 1. Threats to Internal Design Validity: Reasons why inferences about a causal relationship between two variables may be incorrect	Rating (see footnote on use of ratings)	
	A	B
1. **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 (e.g., urban development programs) do not have a precise starting date but get going over periods of months or even years.		
2. **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).		
3. **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.		
4. **Maturation.** Maturation produces many natural changes in physical development, behavior, knowledge, and exposure to new experiences. It is often difficult to separate changes due to maturation from those due to the project.		
5. **Regression toward the mean.** If subjects are selected because of their extreme scores (e.g., weight, physical development), there is a natural tendency to move closer to the mean over time—thus diminishing or distorting the effects of the program.		
6. **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 dropout over time may have changed the characteristics of the project population (e.g., the poorest or least educated might drop out).		
7. **Testing.** Being interviewed or tested may affect behavior or responses. For example, being asked about expenditures may encourage people to cut down on socially disapproved expenditures (cigarettes and alcohol) and spend more on acceptable items.		
8. **Instrumentation.** As researchers gain more experience, they may change how they interpret rating scales, observation checklists, and so on.		
9. **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.		
10. **Information is not collected from the right people, or some categories of informants are not interviewed.**** Sometimes information is collected from and about only 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.		

(Continued)

(Continued)

Checklist 1. Threats to Internal Design Validity: Reasons why inferences about a causal relationship between two variables may be incorrect	Rating (see footnote on use of ratings)	
	A	B
11. **Use of less rigorous designs due to budget and time constraints.**** Many evaluations are conducted under budget, time, and data constraints, which require reductions in sample size, time for pilot studies, amount of information that can be collected, and so on. These constraints increase vulnerability to threats to internal validity.		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Notes on the four checklists:

1. **= additional categories included by the present authors.

2. How to use the checklist: Column A = the existence or seriousness of each threat to validity. It is possible to simply check issues that exist, or a rating scale can be used. A typical rating scale: 1 = the methodology is sound, and there are no issues or problems; 5 = there are major methodological problems and issues. Column B = the importance of this threat for the purposes of this particular evaluation. The same two options exist. The first is to simply check each item rated as 4 or 5 in Column A that has important implications for the purposes of this evaluation. The second option is to rate the importance of this threat for the purposes of the present evaluation. For example: 1 = the threat does not have important implications for this evaluation and 5 = the threat has serious implications for the purposes of this evaluation.

Summary scores for each column: The summary score for Column A can be calculated either as the number of items that have been checked as having methodological problems, or when a rating scale is used, the mean rating can be calculated (the sum of all scores divided by the number of indicators rated). For Column B, this will normally be the number of items in Column A that were rated as having problems that were considered to have important policy or operational implications for the use of the evaluation. If mean scores are calculated, it is important to be aware of the dangers of treating ordinal variables as if they were interval (calculating means, etc.).

Source: Adapted from Shadish, Cook, & Campbell (2002). Checklists 1, 2, 3, and 4 are adapted from Tables 2.2, 2.4, 3.1, and 3.2, respectively. The proposed rating systems and the categories indicated by ** were included by the present authors.

Checklist 2. Threats to Statistical Conclusion Validity: Reasons why inferences about statistical association (covariation) between two variables may be incorrect	A	B
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. **Some assumptions of the statistical tests have been violated.** Many statistical tests require that observations be independent of each other. If this assumption is violated (e.g., studying children in the same classroom, or patients in the same clinic who may be more similar to each other than the population in general), this can increase the risk of Type I error (false positive), wrongly concluding the project had an effect.		
3. **"Fishing" for statistically significant results.** A certain percentage of statistical tests will show "significant" results by chance (1 in 20 at the .05 significance level). Generating large numbers of statistical tables will always find some of these spurious results.		
4. **Unreliability of measures of change of outcome indicators.** Unreliable measures of, for example, rates of change in income, literacy, and infant mortality always reduce the likelihood of finding a significant effect.		
5. **Restriction of range or extrapolation from a truncated or incomplete database.**** If only similar groups are compared, the power of the test is reduced and the likelihood of finding a significant effect is also reduced. If the sample only covers part of the population (e.g., 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.		
6. **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.		
7. **External events influence outcomes (extraneous variance in the experimental setting).** External events or pressures (power failure, community violence, election campaigns) may distract subjects and affect behavior and program outcomes.		
8. **Diversity of the population (heterogeneity of units).** If subjects have widely different characteristics, this may increase the variance of results and make it more difficult to detect significant effects.		
9. **Inaccurate effect size estimation due to outliers.** A few outliers (extreme values) can significantly reduce effect size and make it less likely that significant differences will be found.		
10. **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 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.		
11. **Information is not collected from the right people, or some categories of informants are not interviewed.**** Sometimes information is collected from and about only 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.		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Checklist 3. Threats to Construct Validity: Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect	A	B
1. **Inadequate explanation of constructs.** Constructs (the 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 behavior, 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 underrepresents 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 (e.g., 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 (confounding constructs with levels of constructs).** Often a treatment is administered at only 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. **Program participants and comparison groups respond differently to some questions (treatment-sensitive factorial structure).** 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.		
7. **Participants assess themselves and their situation differently than the comparison group (reactive self-report changes).** 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.		
8. **Reactivity to the experimental situation.** Project participants' interpretation of the project situation may affect their behavior. If they believe the program is run by a religious organization, they may respond differently than if they believe it is run by a political group.		
9. **Experimenter expectancies.** Experimenters also have expectations (e.g., about how men and women or different socioeconomic groups will react to the program), and this may affect how they react to different groups.		
10. **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.		

Checklist 3. Threats to Construct Validity: Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect	A	B
11. **Compensatory effects and rivalry.** Programs create a dynamic that can affect outcomes in different ways. There may be pressure to provide benefits to nonparticipants, 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.		
12. **Using constructs developed in other countries without pretesting in the local context.**** Many evaluations import theories and constructs from other countries and may not adequately capture the local project situation. For example, many evaluations of the impacts of microcredit on women's empowerment in countries such as Bangladesh use international definitions of empowerment that may not be appropriate for Bangladeshi women.		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Checklist 4. Threats to External Validity: Reasons why inferences about how study results would hold over variations in persons, settings, treatments, and outcomes may be incorrect	A	B
1. **Sample does not cover the whole population of interest.** Subjects interviewed 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 difficult to generalize from the study findings to the whole population.		
2. **Different settings affect program outcomes (interaction of the causal relationship over treatment variations).** 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 (interaction of the causal relationship with outcomes).** Different outcome measures can produce different conclusions on project effectiveness. Microcredit 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 (interactions of the causal relationships with 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 from one particular setting to different settings.		
5. **Programs operate differently in different settings (context-dependent mediation).** 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 nongovernmental organizations.		
6. **The attitude of policymakers and politicians to the program.**** Identical programs will operate differently and have different outcomes in situations where they have the active support of policymakers 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 internal terms of trade for key exports and imports. Attempts to generalize or project findings from pilot programs must take these cycles into consideration.		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

APPENDIX 7.2 WORKSHEET FOR ASSESSING THREATS TO THE VALIDITY OF THE FINDINGS AND RECOMMENDATIONS OF QUALITATIVE IMPACT EVALUATION DESIGNS

Part 1: Summary of the Findings of the Assessment of the Evaluation

1. Name of the project/program

2. Who conducted this validity assessment? (indicate organizational affiliation)

3. At what stage of the project/program did the evaluation begin?

4. At what stage of the evaluation was this assessment conducted?

5. Reason for conducting the threats to validity assessment

6. Summary of findings of the assessment with respect to the evaluation methodology (note whether the assessment included the use of the checklists in Part 3)

7. (If methodological problems were identified) What are the implications of these problems for the utilization of the evaluation findings and recommendations?

8. Recommended follow-up actions

Part 2: Summary Assessment for Each Component

A. **Confirmability:** Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:[1]
- The number of methodology problems that affect the utilization of the evaluation:

B. **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?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodology problems that affect the utilization of the evaluation:

1. Note on ratings: 1 = the design and implementation of the methodology for this component is sound and there are no problems or issues; 5 = the design and/or implementation of the methodology for this component are weak, and many issues could affect the validity of the findings and recommendations.

C. **Credibility:** Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodology problems that affect the utilization of the evaluation:

D. **Transferability:** Do the conclusions fit other contexts, and how widely can they be generalized?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodology problems that affect the utilization of the evaluation:

Part 3: Checklists Used to Assess the Five Components Describing Potential Threats to the Adequacy and Validity of a Qualitative Impact Evaluation

NOTE: Part 3, which involves a more in-depth, technical, and expensive assessment, is normally only used for large, high-priority evaluations or where the controversial nature of the topic requires a high level of credibility of the assessment. For these reasons, Part 3 would normally be conducted by an external consultant.

Dimension 1: Internal Validity

- Checklist 1: Confirmability
- Checklist 2: Dependability
- Checklist 3: Credibility

Dimension 2: External Validity

- Checklist 4: Transferability

Dimension 3: Utilization

- Checklist 5: Utilization

Checklist 1. Confirmability:[2] Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?		
	Rating (see footnote on use of ratings)	
	A	B
1. Are the study's methods and procedures adequately described? Are study data retained and available for reanalysis?		
2. Are data presented to support the conclusions?		
3. Has the researcher been as explicit and self-aware as possible about personal assumptions, values, and biases?		
4. Were the methods used to control for bias adequate?		
5. Were competing hypotheses or rival conclusions considered?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Notes on the five checklists:

1. ** = additional categories included by the present authors.

2. How to use the checklist: Column A = the existence or seriousness of each threat to validity. There are two options: (1) Simply check all indicators where methodological problems exist, or (2) use a rating scale to indicate the severity of the issue. A typical rating scale: 1 = the methodology is sound and there are no issues or problems; 5 = there are major methodological problems and issues. Column B = the importance of each methodological threat identified in Column A for the purposes of the present evaluation. In this case, all items that have important implications for the present evaluation are checked.

Summary scores for each column: The summary score for Column A can be calculated either as the number of items that have been checked as having methodological problems, or when a rating scale is used, the mean rating is calculated (the sum of all scores divided by the number of indicators rated). If mean scores are calculated, it is important to be aware of the dangers of treating ordinal variables as if they were interval (calculating means, etc.). For Column B, this will normally be the number of items in Column A that were rated as having problems that were considered to have important policy or operational implications for the use of the evaluation.

2. *Source:* Adapted from Miles & Huberman (1994, Chapter 10, Section C). See also Guba & Lincoln (1989). The rating scales were developed by the present authors. Items indicated by ** were added by the present authors.

Checklist 2. 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?		
	A	B
1. Are findings trustworthy, consistent, and replicable across data sources and over time?		
2. Were data collected across the full range of appropriate settings, times, respondents, and so on?		
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 QUAL studies), is this recognized and addressed?		
6. Were peer or colleague reviews used?		
7. Were the rules used for confirmation of propositions, hypotheses, and so on made explicit?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Checklist 3. Credibility: Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?		
	A	B
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, or 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?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
Gender comments on this component		

Checklist 4. Transferability: Do the conclusions fit other contexts, and how widely can they be generalized?		
	A	B
1. Are the characteristics of the sample of persons, settings, processes, and so on described in enough detail to permit comparisons with other samples?		
2. Does the sample design theoretically permit generalization to other populations?		
3. Does the researcher define the scope and boundaries of reasonable generalization from the study?		
4. Do the findings include enough "thick description" for readers to assess the potential transferability?		
5. Do a range of readers report the findings to be consistent with their own experience?		
6. Do the findings confirm or are they congruent with existing theory? Is the transferable theory made explicit?		
7. Are the processes and findings generic enough to be applicable in other settings?		
8. Have narrative sequences been preserved? Has a general cross-case theory using the sequences been developed?		
9. Does the report suggest settings where the findings could fruitfully be tested further?		
10. Have the findings been replicated in other studies to assess their robustness? If not, could replication efforts be mounted easily?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Checklist 5. Utilization: How useful were the findings to clients, researchers, and the communities studied?		
	A	B
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?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

APPENDIX 7.3 INTEGRATED WORKSHEET FOR ASSESSING THREATS TO THE VALIDITY OF THE FINDINGS AND RECOMMENDATIONS OF MIXED-METHOD IMPACT EVALUATION DESIGNS (STANDARD VERSION)

Part 1: Summary of the Findings of the Assessment of the Evaluation

1. Name of the project/program
2. Who conducted this validity assessment? (indicate organizational affiliation)
3. At what stage of the project/program did the evaluation begin?
4. At what stage of the evaluation was this assessment conducted?
5. Reason for conducting the threats to validity assessment
6. Summary of findings of the assessment with respect to the evaluation methodology (note whether the assessment included the use of the checklists in Part 3)
7. (If methodological problems were identified) What are the implications of these problems for the utilization of the evaluation findings and recommendations?
8. Recommended follow-up actions

Part 2: Summary Assessment for Each Component

A. **Objectivity (Confirmability):** Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

B. **Internal Design Validity (Reliability/Dependability/Credibility/Authenticity):** Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying? 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 (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

(Continued)

(Continued)

C. **Statistical Conclusion Validity:** Reasons why inferences about statistical association (covariation) between two variables may be incorrect

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

D. **Construct Validity:** Do the constructs used to define processes, outcomes and impacts, and contextual variables adequately capture the essential elements of what is being measured? Are the constructs sufficiently comprehensive to capture the multidimensionality of many of the constructs?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

E. **External Validity (Transferability and Fittingness):** Do the conclusions fit other contexts, and how widely can they be generalized? Do they provide credible evidence on how the program would perform in other settings?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

F. **Utilization:** How useful were the findings to clients, researchers, and the communities studied?

Summary assessment and recommendations (identify most serious operational problems)

- The quality of the methodology of this component: Rating:
- The number of methodological problems that could affect the utilization of the evaluation:

Note: How to use the checklist:

Column A = the existence or seriousness of each threat to validity. It is possible to simply check issues that exist or a rating scale can be used. A typical rating scale: 1 = the methodology is sound and there are no issues or problems; 5 = there are major methodological problems and issues. Column B = the importance of this threat for the purposes of this particular evaluation. The same two options exist. The first is to simply check each of the items rated as 4 or 5 in Column A that have important implications for the purposes of this evaluation. The second option is to rate the importance of this threat for the purposes of the present evaluation. For example: 1 = the threat does not have important implications for this evaluation and 5 = the threat has serious implications for the purposes of this evaluation.

Summary scores for each column: The summary score for Column A can be calculated either as the number of items that have been checked as having methodological problems, or when a rating scale is used, the mean rating can be calculated (the sum of all scores divided by the number of indicators rated). For Column B, this will normally be the number of items in Column A that were rated as having problems that were considered to have important policy or operational implications for the use of the evaluation. If mean scores are calculated, it is important to be aware of the dangers of treating ordinal variables as if they were interval (calculating means, etc.).

Part 3: Checklists Used to Assess the Six Components Describing Potential Threats to the Adequacy and Validity of a Quantitative Impact Evaluation

NOTE: Part 3, which involves a more in-depth, technical, and expensive assessment, is normally only used for large, high-priority evaluations or where the controversial nature of the topic requires a high level of credibility of the assessment. For these reasons, Part 3 would normally be conducted by an external consultant.

Checklist 1: Objectivity (Confirmability)

Checklist 2: Internal Design Validity (Reliability/Dependability/Credibility/Authenticity)

Checklist 3: Threats to Statistical Conclusion Validity

Checklist 4: Construct Validity

Checklist 5: External Validity

Checklist 6: Utilization

Checklist 1. Objectivity (Confirmability): Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?	Ratings (see footnote on use of ratings)	
	A	B
1. Are the study's methods and procedures adequately described? Are study data retained and available for reanalysis?		
2. Are data presented to support the conclusions?		
3. Has the researcher been as explicit and self-aware as possible about personal assumptions, values, and biases?		
4. Were the methods used to control for bias adequate?		
5. Were competing hypotheses or rival conclusions considered?		
Summary score		
Number of methodological issues potentially affecting the use of the evaluation		
General comments on the component		

Notes and Sources:

See note on how to apply the ratings at the end of Part 2.

Sources for the six checklists: Items in normal text (i.e., those used most commonly in qualitative evaluations) are adapted from Miles & Huberman (1994, Chapter 10, Section C) and from Teddlie & Tashakkori (2009), especially Chapter 12. See also Guba & Lincoln (1989). Items in italics (i.e., those most commonly used in quantitative evaluations) are adapted from Shadish, Cook, & Campbell (2002, Tables 2.2, 2.4, 3.1, and 3.2). The rating scales were developed by the present authors.

**Indicates items added by the present authors.

Checklist 2. Internal Design Validity (Reliability/Dependability/Credibility/Authenticity): Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?		
	A	B
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, or 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. Are findings trustworthy, consistent, and replicable across data sources and over time?		
8. Were data collected across the full range of appropriate settings, times, respondents, and so on?		
9. Did all fieldworkers have comparable data-collection protocols?		
10. Were coding and quality checks made, and did they show adequate agreement?		
11. Do the accounts of different observers converge? If they do not (which is often the case in QUAL studies), is this recognized and addressed?		
12. Were peer or colleague reviews used?		
13. Were the rules used for confirmation of propositions, hypotheses, and so on made explicit?		
14. ***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 (e.g., urban development programs) do not have a precise starting date but get going over periods of months or even years.*		
15. ***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 best-organized communities).*		
16. ***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.*		
17. ***Maturation.*** *Maturation produces many natural changes in physical development, behavior, knowledge, and exposure to new experiences. It is often difficult to separate changes due to maturation from those due to the project.*		

Checklist 2. Internal Design Validity (Reliability/Dependability/Credibility/Authenticity): Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?		
	A	B
18. ***Regression toward the mean.*** *If subjects are selected because of their extreme scores (e.g., weight, physical development), there is a natural tendency to move closer to the mean over time—thus diminishing or distorting the effects of the program.*		
19. ***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 (e.g., the poorest or least educated might drop out).*		
20. ***Testing.*** *Being interviewed or tested may affect behavior or responses. For example, being asked about expenditures may encourage people to cut down on socially disapproved expenditures (cigarettes and alcohol) and spend more on acceptable items.*		
21. ***Instrumentation.*** *As researchers gain more experience, they may change how they interpret rating scales, observation checklists, and so on.*		
22. ***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.*		
22. ***Information is not collected from the right people, or some categories of informants are not interviewed.***** *Sometimes information is collected from and about only 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.*		
23. ***Use of less rigorous designs due to budget and time constraints.***** *Many evaluations are conducted under budget, time, and data constraints, which require reductions in sample size, time for pilot studies, amount of information that can be collected, and so on. These constraints increase vulnerability to threats to internal validity.*		
Summary score		
Number of methodological issues potentially affecting the use of the evaluation		
General comments on the component		

Note: Items in italics in this and the two following tables are those normally used in the assessment of quantitative evaluations, while items in normal text are commonly used in the assessment of qualitative evaluations. Most of the items in italics could also be used for some qualitative evaluations.

Checklist 3. Threats to Statistical Conclusion Validity: Reasons why inferences about statistical association (covariation) between two variables may be incorrect	A	B
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. This is particularly important when effect sizes are small.*		
2. ***Some assumptions of the statistical tests have been violated.*** *Many statistical tests require that observations be independent of each other. If this assumption is violated (e.g., studying children in the same classroom, or patients in the same clinic who may be more similar to each other than the population in general), this can increase the risk of Type I error (false positive), wrongly concluding the project had an effect.*		
3. ***"Fishing" for statistically significant results.*** *A certain percentage of statistical tests will show "significant" results by chance (1 in 20 at the .05 significance level). Generating large numbers of statistical tables will always find some of these spurious results.*		
4. ***Unreliability of measures of change of outcome indicators.*** *Unreliable measures of, for example, rates of change in income, literacy, and infant mortality always reduce the likelihood of finding a significant effect.*		
5. ***Restriction of range or extrapolation from a truncated or incomplete database.**** *If only similar groups are compared, the power of the test is reduced, and the likelihood of finding a significant effect is also reduced. If the sample only covers part of the population (e.g., 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.*		
6. ***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.*		
7. ***External events influence outcomes (extraneous variance in the experimental setting).*** *External events or pressures (power failure, community violence, election campaigns) may distract subjects and affect behavior and program outcomes.*		
8. ***Diversity of the population (heterogeneity of units).*** *If subjects have widely different characteristics, this may increase the variance of results and make it more difficult to detect significant effects.*		
9. ***Inaccurate effect size estimation due to outliers.*** *A few outliers (extreme values) can significantly reduce effect size and make it less likely that significant differences will be found.*		
10. ***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 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.*		
11. ***Information is not collected from the right people, or some categories of informants are not interviewed.**** *Sometimes information is collected from and about only 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.*		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Checklist 4. Construct Validity: Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect	A	B
1. **Inadequate explanation of constructs.** Constructs (the 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 behavior, hostile work environment, and 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 underrepresents the real number of incidents.		
3. **Use of a 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 (e.g., 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 (confounding constructs with levels of constructs).*** *Often a treatment is administered at only 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. ***Program participants and comparison groups respond differently to some questions (treatment-sensitive factorial structure).*** *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.*		
7. ***Participants assess themselves and their situation differently than comparison group (reactive self-report changes).*** *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.*		
8. ***Reactivity to the experimental situation.*** *Project participants' interpretation of the project situation may affect their behavior. If they believe the program is run by a religious organization, they may respond differently than if they believe it is run by a political group.*		
9. ***Experimenter expectancies.*** *Experimenters also have expectations (e.g., about how men and women or different socioeconomic groups will react to the program), and this may affect how they react to different groups.*		
10. ***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.*		
11. ***Compensatory effects and rivalry.*** *Programs create a dynamic that can affect outcomes in different ways. There may be pressure to provide benefits to nonparticipants, 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.*		

(Continued)

(Continued)

Checklist 4. Construct Validity: Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect	A	B
12. ***Using constructs developed in other countries without pretesting in the local context.***** *Many evaluations import theories and constructs from other countries and may not adequately capture the local project situation. For example, many evaluations of the impacts of microcredit on women's empowerment in countries such as Bangladesh use international definitions of empowerment that may not be appropriate for Bangladeshi women.*		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

Checklist 5. External Validity: Do the conclusions fit other contexts and how widely can they be generalized?		
	A	B
1. *Sample does not cover the whole population of interest.*		
2. *Different settings affect program outcomes.*		
3. *Different outcome measures give different assessments of project effectiveness.*		
4. *Program outcomes vary in different settings.*		
5. *Programs operate differently in different settings.*		
6. *Does the attitude (positive or negative) of policymakers and politicians to the program affect implementation or outcomes?*		
7. *Do seasonal and other cycles affect implementation or outcomes?*		
8. *Is there an adequate description of sample characteristics?*		
9. *Does the sample design 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. Do 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. *Have the findings been replicated in other studies to assess their robustness? If not, could replication efforts be mounted easily?*		
Summary score		
Number of methodological issues potentially affecting the use of the evaluation		
General comments on the component		

Checklist 6. Utilization: How useful were the findings to clients, researchers, and the communities studied?		
	A	B
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?		
Summary score		
Number of methodological issues potentially affecting the use of the evaluation		
General comments on the component		

APPENDIX 7.4 EXAMPLE OF A COMPLETED THREATS-TO-VALIDITY WORKSHEET

Using the Worksheet to Assess an Already Completed Evaluation of a Low-Cost Housing Project

This example illustrates the use of the integrated worksheet to assess the validity of a fictitious but fairly typical housing project in Central America. The three-year low-cost housing project began in June 2007 and ended in December 2010. A baseline study was conducted with project beneficiaries at the start of the project, and the survey was repeated about six months before the project closing date. The baseline survey covered a sample of 500 households selected randomly from the five project locations. A panel study design was used in which as many as possible of the original project households were re-interviewed for the posttest survey three years later. Approximately 85% were re-interviewed, and a random sample of 75 project households was selected to replace those who could not be re-interviewed. It was originally intended to include a comparison group, but this was cut out both because of budget constraints and because the Ministry of Housing did not wish to raise expectations that families interviewed as part of the comparison group would become eligible to obtain houses in a second phase of the project. When the posttest survey was being commissioned, the Ministry of Housing recognized the need to obtain comparison data to support their claim that the project had produced significant social and economic impacts, and consultants were requested to identify and interview a group of households with similar characteristics to the project participants. A matched comparison group was selected and interviewed. Project impacts were estimated by combining a posttest comparison of the project and comparison groups and a pretest–posttest estimate of the changes in the project group. The analysis found (a) the project group scored significantly higher than the comparison group on a number of economic and social indicators, and (b) the project group scores on the same indicators had improved over the life of the project. It was concluded that the project had produced a significant impact on both the economic and social conditions of participants, and it was recommended that the project should be replicated on a larger scale as part of the national poverty-reduction strategy.

Integrated Worksheet for Assessing Threats to the Validity of the Findings and Recommendations of Mixed-Method Evaluation Design (Standard Version)

Part 1: Summary of the Findings of the Assessment of the Evaluation

1. **Name of the project/program:** *Central America Low-Cost Housing Project*
2. **Who conducted the evaluation?** *A national consulting firm contracted by the Ministry of Housing*
3. Who conducted this validity assessment? (indicate organizational affiliation) *An international consultant hired by the funding agency*
4. At what stage of the project/program did the evaluation begin? *At the start of the project*
5. **At what stage of the evaluation was this assessment conducted?** *Several months after the project closed and the final evaluation report had been submitted*
6. **Reason for conducting the threats-to-validity assessment:** *Requested by the international funding agency as a standard procedure for all projects lasting more than 24 months*
7. Summary of findings of the assessment with respect to the evaluation methodology (note whether the assessment included the use of the checklists in Part 3).

[Note: The assessment did include the use of the checklists.] *The decision of the Ministry not to include a baseline survey of a comparison group seriously affected the validity of the evaluation findings. While there were statistically significant differences between the project group and the comparison group interviewed at the end of the project, the lack of baseline data on the comparison group makes it very difficult to assess the conditions of the comparison group at the start of the project and how they compared with the project group. If the comparison group scored lower than the project group on the economic and social indicators, then the post-project differences may have been at least partly due to the existing differences between the groups at the start of the project. The lack of a baseline comparison group also affected the interpretation of the observed improvements in the project group over the life of the project. It may be that the improvements were at least partly due to general improvements in the economy (for example, general wage increases and increased employment opportunities). However, it is not possible to control for these external factors without having data on the corresponding changes in the comparison group.*

Due to these evaluation design issues, it is necessary to qualify the conclusion that housing is an effective way to improve the economic and social conditions of low-income households, and caution is needed with respect to the recommendation that the project should be expanded as part of the national poverty-reduction strategy.

While the evaluation consultants were not responsible for the decision not to include a baseline comparison group survey, they could be criticized for not having pointed out to the Ministry (which did not have any previous experience with impact evaluations) the problems that this decision would cause with respect to the interpretation of the evaluation findings. The consultants could also have proposed strategies in the posttest survey for estimating the conditions of the comparison group at the start of the project and for identifying and estimating the impacts of external factors during the three-year period of project implementation. For example, consultants did not consult any of the household income and employment surveys that the government conducts every year and which could have provided information on changing earnings and employment opportunities.

8. *(If methodological problems were identified) What are the implications of these problems for the utilization of the evaluation findings and recommendations?*

The weaknesses in the evaluation design seriously affect the findings and recommendations. The finding that housing increased income and improved social indicators of the project families is not fully supported by the evaluation design. As indicated in the previous section, it is not possible to determine to what extent the improved indicators are due to external factors such as improved economic conditions or increased availability of schools and health clinics. It is also not possible to judge whether the differences in the economic and social conditions of the project and comparison groups are due to the project or whether they might be due to differences between the two groups at the time the project began.

Consequently, the recommendations that housing has been shown to be an effective way to increase income and improve social indicators and that the project should be replicated on a larger scale as part of a national poverty-reduction strategy should be carefully assessed.

9. Recommended follow-up actions:

If resources and time permit, it would be useful to commission a rapid follow-up study to review secondary data from the household income and employment surveys and other sources, to estimate how much of the increased income and access to services of the project population was due to a general trend affecting all low-income households in the cities where the project operates.

If similar evaluations are to be commissioned in the future, it is strongly recommended that baseline and posttest data should be collected on both the project population and a comparison group.

Part 2: Summary Assessment for Each Component

1. **Objectivity (Confirmability):** Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?

Summary assessment and recommendations (identify most serious operational problems):

The conclusions are based on the available evidence, but the analysis does not take into consideration the methodological problems resulting from the absence of baseline data on a comparison group. There are some indications (based on interviews with the consultant and the clients) that given the positive reaction of the Ministry of Housing and the international funding agency to the favorable findings, the consultant may have been less rigorous in pointing out the limitations of the findings than he or she should have been.

- The quality of the methodology of this component: *Average rating: 2.8*
- The number of methodological problems that could affect the utilization of the evaluation: *2*

2. **Internal design validity (reliability/dependability/credibility/authenticity):** Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying? 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 (identify most serious operational problems):

The findings were favorably received and considered credible by the client because they presented the project in a favorable light and concluded that it had achieved its objectives. The findings would have seemed less credible to professional evaluators and researchers, but they were not consulted.

- The quality of the methodology of this component: *Average rating: 2.7*
- The number of methodological problems that could affect the utilization of the evaluation: *5*

(Continued)

(Continued)

3. **Statistical conclusion validity:** Reasons why inferences about statistical association (covariation) between two variables may be incorrect

Summary assessment and recommendations (identify most serious operational problems):

The absence of baseline comparison group data weakens the statistical estimation of project impacts. A number of econometric procedures could have been used to strengthen the analysis, but these were not used.

- The quality of the methodology of this component: *Average rating: 1.5*
- The number of methodological problems that could affect the utilization of the evaluation: *1*

4. **Construct validity:** Does the adequacy of the constructs used to define processes, outcomes and impacts, and contextual variables capture the essential elements of what is being measured? Are the constructs sufficiently comprehensive to capture the multidimensionality of many of the constructs?

Summary assessment and recommendations (identify most serious operational problems):

The constructs used to estimate changes in income are weakened by an inadequate definition of income from the informal sector and from other sources such as rent and remittances, so household income may have been underestimated.

- The quality of the methodology of this component: *Average rating: 1.9*
- The number of methodological problems that could affect the utilization of the evaluation: *2*

5. **External validity (transferability and fittingness):** Do the conclusions fit other contexts, and how widely can they be generalized? Do they provide credible evidence on how the program would perform in other settings?

Summary assessment and recommendations (identify most serious operational problems):

The evaluation recommends that the project should be replicated on a larger scale and estimates that it would produce significant impacts on income and other social indicators. However, no analysis is made of any of the special characteristics of the household populations, the project locations, or the support from government agencies and politicians that might have affected project outcomes and that might not exist in other locations where a larger project might operate. Consequently, caution is required when stating that the project would have similar outcomes in other locations.

- The quality of the methodology of this component: *Average rating: 3.2*
- The number of methodological problems that could affect the utilization of the evaluation: *3*

6. **Utilization:** How useful were the findings to clients, researchers, and the communities studied?

Summary assessment and recommendations (identify most serious operational problems):

The findings were considered very useful by the Ministry of Housing and the donor and were used to support the proposal to fund a larger second phase. The project community was not involved in the planning or review of the evaluation, and they were not informed about the findings or recommendations.

- The quality of the methodology of this component: *Average rating: 3.2*
- The number of methodological problems that could affect the utilization of the evaluation: *3*

Part 3: Checklists Used to Assess the Six Components Describing Potential Threats to the Adequacy and Validity of a Quantitative Impact Evaluation

Checklist 1: Objectivity (Confirmability)

Checklist 2: Internal Design Validity (Reliability/Dependability/Credibility/Authenticity)

Checklist 3: Threats to Statistical Conclusion Validity

Checklist 4: Construct Validity

Checklist 5: External Validity

Checklist 6: Utilization

Checklist 1. Objectivity (Confirmability)		
Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?	**Ratings (see footnote on use of ratings)**	
	A	**B**
1. Are the study's methods and procedures adequately described? Are study data retained and available for reanalysis?	*1*	
2. Are data presented to support the conclusions?	*2*	
3. Has the researcher been as explicit and self-aware as possible about personal assumptions, values, and biases?	*2*	
4. Were the methods used to control for bias adequate?	*4*	√
5. Were competing hypotheses or rival conclusions considered?	*5*	√
Summary score (average rating)	***2.8***	
Number of methodological issues potentially affecting the use of the evaluation		***2***

General comments on the component:

The conclusions are based on the available evidence, but the analysis does not take into consideration the methodological problems resulting from the absence of baseline data on a comparison group. There are some indications (based on interviews with the consultant and the clients) that given the positive reaction of the Ministry of Housing and the international funding agency to the favorable findings, the consultant may have been less rigorous in pointing out the limitations of the findings than he or she should have been.

Notes and Sources:

How to use the checklist: Column A = the existence or seriousness of each threat to validity. It is possible to simply check issues that exist, or a rating scale can be used. A typical rating scale: 1 = the methodology is sound and there are no issues or problems; 5 = there are major methodological problems and issues. Column B = the importance of this threat for the purposes of this particular evaluation. The same two options exist. The first is to simply check each of the items rated as 4 or 5 in Column A that have important implications for the purposes of this evaluation. The second option is to rate the importance of this threat for the purposes of the present evaluation. For example: 1 = the threat does not have important implications for this evaluation and 5 = the threat has serious implications for the purposes of this evaluation.

Summary scores for each column: The summary score for Column A can be calculated either as the number of items that have been checked as having methodological problems, or when a rating scale is used, the mean rating can be calculated (the sum of all scores divided by the number of indicators rated). For Column B, this will normally be the number of items in Column A that were rated as having problems that were considered to have important policy or operational implications for the use of the evaluation. If mean scores are calculated, it is important to be aware of the dangers of treating ordinal variables as if they were interval (calculating means, etc.).

Sources for the six checklists: Sources used most commonly in qualitative evaluations are adapted from Miles & Huberman (1994, Chapter 10, Section C) and from Teddlie & Tashakkori (2009), especially Chapter 12. See also Guba & Lincoln (1989). Sources most commonly used in quantitative evaluations are adapted from Shadish, Cook, & Campbell (2002), Tables 2.2, 2.4, 3.1, and 3.2. The rating scales were developed by the present authors.

**Indicates items added by the present authors.

Checklist 2. Internal Design Validity (Credibility, Authenticity)		
Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?	**A**	**B**
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?	4	
2. Does the account ring true, make sense, or seem convincing? Does it reflect the local context?	3	
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	
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?	3	
5. Are areas of uncertainty identified? Was negative evidence sought, found? How was it used? Have rival explanations been actively considered?	4	√
6. Were conclusions considered accurate by the researchers responsible for data collection?	2	
7. Are findings trustworthy, consistent, and replicable across data sources and over time?	4	√
8. Were data collected across the full range of appropriate settings, times, respondents, and so on?	5	√
9. Did all fieldworkers have comparable data-collection protocols?	1	
10. Were coding and quality checks made, and did they show adequate agreement?	2	
11. Do the accounts of different observers converge? If they do not (which is often the case in QUAL studies), is this recognized and addressed?	2	
12. Were peer or colleague reviews used?	4	√
13. Were the rules used for confirmation of propositions, hypotheses, and so on made explicit?	3	
14. **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.	2	
15. **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 best-organized communities).	3	
16. **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.	5	√

Checklist 2. Internal Validity (Credibility, Authenticity)		
Are the findings credible to the people studied and to readers, and do we have an authentic portrait of what we are studying?	**A**	**B**
17. **Maturation.** Maturation produces many natural changes in physical development, behavior, knowledge, and exposure to new experiences. It is often difficult to separate changes due to maturation from those due to the project.	3	
18. **Regression toward the mean.** If subjects are selected because of their extreme scores (e.g., weight, physical development), there is a natural tendency to move closer to the mean over time—thus diminishing or distorting the effects of the program.	n/a	
19. **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 (e.g., the poorest or least educated might drop out).	3	
20. **Testing.** Being interviewed or tested may affect behavior or responses. For example, being asked about expenditures may encourage people to cut down on socially disapproved expenditures (cigarettes and alcohol) and spend more on acceptable items.	1	
21. **Instrumentation.** As researchers gain more experience, they may change how they interpret rating scales, observation checklists, and so on.	1	
22. **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.	n/a	
23. **Information is not collected from the right people, or some categories of informants are not interviewed.**** Sometimes information is collected from and about only 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.	1	
24. **Use of less rigorous designs due to budget and time constraints.**** Many evaluations are conducted under budget, time, and data constraints, which require reductions in sample size, time for pilot studies, amount of information that can be collected, and so on. These constraints increase vulnerability to threats to internal validity.	n/a	
Summary score (average rating)	**2.7**	
Number of methodological issues potentially affecting the use of the evaluation		**5**

General comments on the component:

The researchers followed appropriate procedures for sample selection and for quality control of data. The weaknesses concerned not having collected baseline data on the comparison group and not having collected data to control for economic and other events during the three years that might have affected outcomes.

Note: Items in italics are those normally used in the assessment of quantitative evaluations, while items in normal text are commonly used in the assessment of qualitative evaluations.

Checklist 3. Threats to Statistical Conclusion Validity		
Reasons why inferences about statistical association (covariation) between two variables may be incorrect	**A**	**B**
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. This is particularly important when effect sizes are small.	1	
2. **Some assumptions of the statistical tests have been violated:** Many statistical tests require that observations be independent of each other. If this assumption is violated (e.g., studying children in the same classroom, or patients in the same clinic who may be more similar to each other than the population in general), this can increase the risk of Type I error (false positive) wrongly concluding the project had an effect.	3	
3. **"Fishing" for statistically significant results:** A certain percentage of statistical tests will show "significant" results by chance (1 in 20 at the .05 significance level). Generating large numbers of statistical tables will always find some of these spurious results.	1	
4. **Unreliability of measures of change of outcome indicators:** Unreliable measures of, for example, rates of change in income, literacy, and infant mortality always reduce the likelihood of finding a significant effect.	1	
5. **Restriction of range or extrapolation from a truncated or incomplete database.**** If only similar groups are compared, the power of the test is reduced and the likelihood of finding a significant effect is also reduced. If the sample covers only part of the population (e.g., 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.	1	
6. **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.	1	
7. **External events influence outcomes (extraneous variance in the experimental setting):** External events or pressures (power failure, community violence, election campaigns) may distract subjects and affect behavior and program outcomes.	5	√
8. **Diversity of the population (heterogeneity of units):** If subjects have widely different characteristics, this may increase the variance of results and make it more difficult to detect significant effects.	1	
9. **Inaccurate effect size estimation due to outliers:** A few outliers (extreme values) can significantly reduce effect size and make it less likely that significant differences will be found.	1	
10. **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 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.	3	
11. **Information is not collected from the right people, or some categories of informants are not interviewed.**** Sometimes information is collected from and about only 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.	1	
Summary score (average rating)	**1.5**	
Number of methodological issues affecting the use of the evaluation		**1**
General comments on this component:		
The sample design and the statistical tests were correctly administered. The problem was that only very biased statistical tests (comparison of single means) were used, and some of the more powerful econometric tests that could have partially controlled for the absence of baseline data were not used.		

Checklist 4. Construct Validity		
Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect	**A**	**B**
1. **Inadequate explanation of constructs:** Constructs (the 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 behavior, hostile work environment, and sex discrimination).	1	
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 underrepresents the real number of incidents.	4	√
3. **Use of a 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, and domestic violence) will usually produce bias.	3	
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 (e.g., observing communities in formal meetings will produce different results than observing social events or communal work projects).	3	
5. **Only one level of the treatment is studied (confounding constructs with levels of constructs):** 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.	1	
6. **Program participants and comparison groups respond differently to some questions (treatment-sensitive factorial structure):** 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.	2	
7. **Participants assess themselves and their situation differently than the comparison group (reactive self-report changes):** 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.	1	
8. **Reactivity to the experimental situation:** Project participants' interpretation of the project situation may affect their behavior. If they believe the program is run by a religious organization, they may respond differently than if they believe it is run by a political group.	1	
9. **Experimenter expectancies:** Experimenters also have expectations (e.g., about how men and women or different socioeconomic groups will react to the program), and this may affect how they react to different groups.	4	√
10. **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.	1	
11. **Compensatory effects and rivalry:** Programs create a dynamic that can affect outcomes in different ways. There may be pressure to provide benefits to nonparticipants, 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.	1	

(Continued)

(Continued)

Checklist 4. Construct Validity		
Reasons why inferences about the constructs used to define implementation processes, outputs, outcomes, and impacts may be incorrect	**A**	**B**
12. Using constructs developed in other countries without pretesting in the local context.** Many evaluations import theories and constructs from other countries and may not adequately capture the local project situation. For example, many evaluations of the impacts of microcredit on women's empowerment in countries such as Bangladesh use international definitions of empowerment that may not be appropriate for Bangladeshi women.	1	
Summary score (average rating)	1.9	
Number of methodological issues affecting the use of the evaluation		2

General comments on this component:

The definition and measurement of income did not adequately capture informal sector earnings or nonearned income such as rent and remittances. Consequently, income may have been underestimated.

Experimenter expectancies concerning a positive project outcome may have made researchers less critical of the weak evaluation design and the potential bias of the findings.

Checklist 5. External Validity		
Do the conclusions fit other contexts, and how widely can they be generalized?	**A**	**B**
1. Sample does not cover the whole population of interest.	*3*	
2. Different settings affect program outcomes.	*3*	
3. Different outcome measures give different assessments of project effectiveness.	*1*	
4. Program outcomes vary in different settings.	*3*	
5. Programs operate differently in different settings.	*3*	
6. Does the attitude (positive or negative) of policymakers and politicians to the program affect implementation or outcomes?	*3*	
7. Do seasonal and other cycles affect implementation or outcomes?	*3*	
8. Is there an adequate description of sample characteristics?	*3*	
9. Does the sample design permit generalization to other populations?	*5*	√
10. Does the researcher define the scope and boundaries of reasonable generalization from the study?	4	√
11. Do the findings include enough "thick description" for readers to assess the potential transferability?	4	√
12. Do a range of readers report the findings to be consistent with their own experience?	n/a	
13. Do the findings confirm or are they congruent with existing theory? Is the transferable theory made explicit?	n/a	
14. Are the processes and findings generic enough to be applicable in other settings?	3	
15. Have narrative sequences been preserved? Has a general cross-case theory using the sequences been developed?	n/a	
16. Have the findings been replicated in other studies to assess their robustness? If not, could replication efforts be mounted easily?	4	
Summary score (average rating)	**3.2**	
Number of methodological issues potentially affecting the use of the evaluation		**3**

General comments on the component:

Many of the issues mentioned in the checklist have not been addressed. The main weakness is that the factors affecting potential replicability have not been assessed.

Checklist 6. Utilization		
How useful were the findings to clients, researchers, and the communities studied?	**A**	**B**
1. Are the findings intellectually and physically accessible to potential users?	3	
2. Were any predictions made in the study and, if so, how accurate were they?	n/a	
3. Do the findings provide guidance for future action?	4	√
4. Do the findings have a catalyzing effect leading to specific actions?	4	√
5. Do the actions taken actually help solve local problems?	3	
6. Have users of the findings experienced any sense of empowerment or increased control over their lives? Have they developed new capacities?	4	
7. Are value-based or ethical concerns raised explicitly in the report? If not, do some exist that the researcher is not attending to?	1	
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.)	3	
9. Is there evidence that the evaluation had a significant influence on future project design?	2	
10. Is there evidence that the evaluation influenced policy decisions?	4	√
Summary score	**3.2**	
Number of methodological issues potentially affecting the use of the evaluation		**3**

General comments on the component:

The main problem is that the evaluation recommended that the project should be replicated based on questionable analysis. Policymakers have been influenced by these recommendations and used them to support a proposal to finance a second and larger project.

APPENDIX 7.5 INTEGRATED WORKSHEET FOR ASSESSING THREATS TO THE VALIDITY OF FINDINGS AND RECOMMENDATIONS OF A MIXED-METHOD IMPACT EVALUATION DESIGNS (ADVANCED VERSION)[1]

Part 1: Summary of the Findings of the Assessment of the Evaluation

1. Name of project/program
2. Who conducted the evaluation? (indicate organizational affiliation)
3. Who conducted this validity assessment? (indicate organizational affiliation)
4. At what stage of the project/program did the evaluation begin?
5. At what stage of the evaluation was this assessment conducted?
6. Reason for conducting the threats to validity assessment
7. Summary of findings of the assessment (note whether the assessment included the use of the checklists in Part 3)
8. (If methodological problems were identified) What are the implications of these problems for the utilization of the evaluation findings and recommendations?
9. Recommended follow-up actions (if any)

1. *Source:* Adapted from Teddlie & Tashakkori (2009, Chapter 12, Tables 12.5 and 12.6). Additional material included from Appendices 7.1 and 7.2 of the present publication (additional sources cited in these two appendices).

Part 2: Summary Assessment for Each Dimension and Category (see Part 3 for more detailed assessments)

Dimension 1: Internal Threats to Validity
Level 1: Overall design quality: How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented and analyzed, and how well were both integrated into a mixed-method approach? (see Checklist 1)
Summary assessment and recommendations
• The quality of the methodology of this component: Summary Rating:[2] • The number of methodology problems that affect the utilization of the evaluation:
Level 2: Quality assessment for each stage of the evaluation (see Checklist 2)
Summary assessment and recommendations

Data quality						
Data analysis						
Inference						
Integration						
Overall rating						

2. Note on ratings: If the Part 3 checklists are not used, then judgmental ratings will be used on a 5-point scale where 1 = this aspect of the evaluation methodology is sound and there are no issues affecting the findings and recommendations and 5 = this aspect of the evaluation methodology has serious problems that affect the findings and recommendations. The rating system is described in more detail in the note to Checklist 1. If the Part 3 checklists are used, the summary rating and number of methodological problems are included from the respective checklists.

Dimension 2: External Threats to Validity (Checklist 3)

How well will the findings and recommendations apply to different groups and in different settings? Have these differences been adequately addressed in the reports?

Summary assessment and recommendations

- The quality of the methodology of this component: Rating:
- The number of methodology problems that affect the utilization of the evaluation:

Dimension 3: Utilization Validity (Checklist 4)

How useful were the findings to clients, researchers, and the communities studied? Is there evidence that the findings and recommendations influenced policymakers or the design of future projects?

Summary assessment and recommendations

- The quality of the methodology of this component: Rating:
- The number of methodology problems that affect the utilization of the evaluation:

Dimension 4: Interpretative Rigor (see Checklist 5)

Are the findings credible on the basis of the results obtained in the evaluation?

Summary assessment and recommendations

- The quality of the methodology of this component: Rating:
- The number of methodology problems that affect the utilization of the evaluation:

Part 3: Checklists Used to Assess Each Dimension and Component

Checklist	Dimension		Level	
1	1	Internal threats to validity	1	Overall design quality
2	1	Internal threats to validity	2	Quality assessment for each stage of the evaluation
3	2	External validity		
4	3	Utilization validity		
5	4	Interpretative rigor		

Checklist 1. Dimension 1: Internal Threats to Validity Level 1: Overall Design Quality			
How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into a mixed-method approach?			
		Ratings[3]	
		Threat	**Importance**
		A	**B**
1.	Design suitability: Was the design appropriate for the purpose of the study?		
a.	Are the methods of the study appropriate for answering the research questions? Does the design match the research questions?		
b.	Does the mixed-method design match the stated purpose for conducting an integrated study?		
c.	Do all the strands of the mixed-method study address the same research questions?		
	Summary score for design suitability		
2.	Internal design validity: Does the design adequately address threats to internal design validity? (see Part 4: Guidelines for Checklist 1)		
a.	Temporal precedence of interventions and effects		
b.	Project selection bias		
c.	History		
d.	Attrition		
e.	Maturation		
f.	Regression toward the mean		
g.	Testing		
h.	Instrumentation		
i.	Potential biases or distortion during the process of recall		
j.	Information not collected from the right people		
	Summary score for internal design validity		
3.	Construct validity: The adequacy and comprehensiveness of constructs used to define processes, outcomes, impacts, and contextual variables (see Part 4 guidance for explanation)		
a.	Inadequate explanation of constructs		
b.	Indicators do not adequately measure constructs		
c.	Use of a single indicator or method to measure a complex construct		
d.	Use of a single method to measure a construct (mono-method bias)		
e.	Only one level of the treatment is studied		
f.	The implicit program theory model is not well documented		
g.	Program participants and comparison groups respond differently to some questions		
h.	Participants assess themselves and their situation differently from comparison group		
i.	Reactivity to the experimental situation		

(Continued)

3. See note at the end of the checklist.

Checklist 1. (Continued)			
How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into a mixed-method approach?			
		Ratings[3]	
		Threat	**Importance**
		A	**B**
j.	Experimenter expectancies		
k.	Novelty and disruption effects		
l.	Compensatory effects and rivalry		
m.	Using indicators and constructs developed in other countries without pretesting in the local context		
n.	The process of "quantizing" or "qualitizing" changes the nature or meaning of a variable in a way that can be misleading		
o.	Does multilevel, mixed-method analysis accurately reflect how the project operates and interacts with its environment?		
	Summary score for construct validity		
4.	Design fidelity: Are the QUANT, QUAL, and mixed-method procedures (sampling, data collection, etc.) implemented with the necessary quality and rigor? (refer to Checklist 2)		
a.	QUANT data collection		
b.	QUAL data collection		
c.	QUANT data analysis		
d.	QUAL data analysis		
	Summary score for design fidelity		
5.	Within-design consistency: consistency among all elements of the design		
a.	Do the components of the design fit together in a seamless manner?		
b.	Do the strands of the MM study follow each other (or are they linked) in a logical and seamless manner?		
	Summary score for within-design consistency		
6.	Analytic adequacy: Was the analytic strategy appropriate and adequate?		
a.	Are the data analysis procedures/strategies appropriate and adequate to provide possible answers to research questions?		
b.	Are the mixed-method analytic strategies implemented effectively?		
	Summary score for analytic adequacy		

Checklist 1. (Continued)			
How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into a mixed-method approach?			
		Ratings[3]	
		Threat	**Importance**
		A	**B**
7.	Objectivity: Are the conclusions drawn from the available evidence, and is the research relatively free of researcher bias?		
a.	Are the conclusions and recommendations presented in the executive summary consistent with, and supported by, the information and findings in the main report?		
b.	Are the study's methods and procedures adequately described? Are study data retained and available for reanalysis?		
c.	Are data presented to support the conclusions? Is evidence presented to support all findings?		
d.	Has the researcher been as explicit and self-aware as possible about personal assumptions, values, and biases?		
e.	Were the methods used to control for bias adequate?		
f.	Were competing hypotheses or rival conclusions considered?		
	Summary score for objectivity		
Summary score for Checklist 1			
Number of methodological issues affecting the use of the evaluation			
General comments and conclusions for this component			

Notes on the five checklists:

1. ** = additional categories included by the present authors.

2. How to use the checklist: Column A = the existence or seriousness of each threat to validity. There are two options: (1) Simply check all indicators where methodological problems exist, or (2) use a rating scale to indicate the severity of the issue. A typical rating scale:

1 = the methodology is very strong and there are no issues or problems; 2 = the methodology is sound and there are no major weaknesses; 3 = the methodology is generally sound but there are some areas of weakness; 4 = the methodology has a number of important weaknesses; 5 = there are major methodological problems and issues. Column B = the importance of each methodological threat identified in Column A for the purposes of the present evaluation. In this case, all items that have important implications for the present evaluation are checked.

Summary scores for each column: The summary score for Column A can be calculated either as the number of items that have been checked as having methodological problems, or when a rating scale is used, the mean rating is calculated (the sum of all scores divided by the number of indicators rated). If mean scores are calculated, it is important to be aware of the dangers of treating ordinal variables as if they were interval (calculating means, etc.). For Column B, this will normally be the number of items in Column A that were rated as having problems that were considered to have important policy or operational implications for the use of the evaluation.

Checklist 2. Dimension 1: Internal Threats to Validity Level 2: Quality assessment for each stage of the evaluation			
How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into a mixed-method approach?			
		Rating[4]	
		Threat	Importance
		A	B
1.	Data collection: The adequacy of the data-collection methods and the quality of the data		
1A.	Quantitative data collection (see Part 4 guidelines for explanations)		
a.	The sample is too small to detect program effects (low statistical power)		
b.	Unreliability of measures of change in outcome indicators		
c.	Restriction of range		
d.	Diversity of the population (heterogeneity of units)		
e.	Project and comparison group samples do not cover the same population		
f.	Information is not collected from the right people, or some categories of informants are not interviewed		
g.	Did budget, time, or data constraints affect the quality of the data? If so, were adequate measures taken to address these limitations?		
h.	Data-collection methods were not appropriate for collecting information on sensitive topics or for interviewing difficult-to-reach groups		
1B.	Qualitative data collection		
a.	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?		
b.	Are findings trustworthy, consistent, and replicable across data sources and over time?		
c.	Were data collected across the full range of appropriate settings, times, respondents, and so on?		
d.	Did all fieldworkers have comparable data-collection protocols?		
e.	Were coding and quality checks made, and did they show adequate agreement?		
f.	Do the accounts of different observers converge? If they do not (which is often the case in qualitative studies), is this recognized and addressed?		
g.	Were peer or colleague reviews used?		
h.	Did budget, time, or data constraints affect the quality of the data? If so, were adequate measures taken to address these limitations?		
i.	Does the account ring true, make sense, or seem convincing? Does it reflect the local context?		
j.	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?		
k.	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?		
l.	Are areas of uncertainty identified? Was negative evidence sought, found? How was it used? Have rival explanations been actively considered?		
m.	Were conclusions considered accurate by the researchers responsible for data collection?		

4. See note at the end of Checklist 1.

Checklist 2. Dimension 1: Internal Threats to Validity Level 2: Quality assessment for each stage of the evaluation			
How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into a mixed-method approach?			
		Rating	
		Threat	**Importance**
		A	**B**
	Summary score for data collection		
2.	Data analysis and the validity of how data are interpreted (inference)		
2A.	Quantitative data analysis and inference (see Part 4 guidelines for more details)		
a.	The sample is too small to detect program effects (low statistical power)		
b.	Some assumptions of the statistical tests have been violated		
c.	"Fishing" for statistically significant results		
d.	Restriction of range		
e.	Unreliability of treatment implementation not captured in the analysis		
f.	Diversity of the population reduces the statistical power of the analysis		
g.	Extrapolation from a truncated or incomplete database		
h.	Project and comparison group samples do not cover the same population		
i.	The sample is too small to detect statistically significant effects		
j.	Sample size for group- and community-level variables is too small to permit statistical significance testing		
2B.	Qualitative data analysis and inference		
a.	Were the rules used for the confirmation of propositions and hypotheses made explicit?		
b.	Were coding and quality checks made, and did they show adequate agreement?		
	Summary score for data analysis		
3.	Integration of quantitative and qualitative approaches at all stages of the evaluation		
a.	Hypotheses development		
b.	Developing the conceptual framework and the program theory model		
c.	Same design		
d.	Data collection		
e.	Data analysis		
	Summary score for integration of quantitative and qualitative approaches		
Summary score for Checklist 2			
Number of methodological issues affecting the use of the evaluation			
General comments on this component			

Checklist 3. Dimension 2: External Validity		
Reasons why inferences about how study results would hold over variations in persons, settings, treatments, and outcomes may be incorrect (see Part 4 guidelines for explanations)		
	Rating[5]	
	Threat	**Importance**
	A	**B**
1. Sample does not cover the whole population of interest		
2. Different settings affect program outcomes		
3. Different outcome measures give different assessments of project effectiveness		
4. Program outcomes vary in different settings		
5. Programs operate differently in different settings		
6. The attitude of policymakers and politicians to the program		
7. Seasonal and other cycles		
8. Adequate description of sample characteristics		
9. Does the sample design 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. Do 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?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

5. See note on ratings at the end of Checklist 1.

Checklist 4. Dimension 3: Utilization Validity		
How useful were the findings to clients, researchers, and the communities studied?		
	Rating[6]	
	Threat	**Importance**
	A	**B**
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?		
Summary score		
Number of methodological issues affecting the use of the evaluation		
General comments on this component		

6. See note on ratings at the end of Checklist 1.

Checklist 5. Dimension 4: Interpretative Rigor			
Are the interpretations credible on the basis of the results obtained in the evaluation?			
		Rating[7]	
		Threat	**Importance**
		A	**B**
1.	Interpretative consistency		
a.	Do the inferences closely follow the relevant findings in terms of type, scope, and intensity?		
b.	Are multiple inferences made on the basis of the same findings consistent with each other?		
2.	Theoretical consistency		
a.	Are the inferences consistent with theory and state of knowledge in the field?		
3.	Interpretive agreement		
a.	Are other scholars likely to reach the same conclusions on the basis of the same results?		
b.	Do the inferences match participants' constructions?		
4.	Interpretive directness		
a.	Is each inference distinctively more credible/plausible than other possible conclusions that might be made on the basis of the same results?		
5.	Integrative efficacy (mixed and multiple methods)		
a.	Do the meta-inferences adequately incorporate the inferences that are made in each strand of the study?		
b.	If there are credible inconsistencies between the inferences made within/across strands, are the theoretical explanations for these inconsistencies explored and possible explanations offered?		
6.	Interpretive correspondence		
a.	Do the inferences correspond to the stated purposes/questions of the study? Do the inferences made in each strand address the purposes of the study in that strand?		
b.	Do the meta-inferences meet the stated need for using a mixed-method design?		
Summary score			
Number of methodological issues affecting the use of the evaluation			
General comments, conclusions, and recommendations from this component			

7. See note on ratings at the end of Checklist 1.

Part 4: Guidelines for Using the Checklists

Note: Explanations are only provided for technical questions that may be more difficult to interpret.

Guidelines for Checklist 1: Overall Design Quality: How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into the mixed-method approach?		
Question No.	**Question**	**Explanation**
1. Design suitability		No explanation required
2. Internal design validity		
a.	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 (e.g., urban development programs) do not have a precise starting date but get going over periods of months or even years.
b.	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).
c.	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.
d.	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 (e.g., the poorest or least educated might drop out).
e.	Maturation	Maturing produces many natural changes in physical development, behavior, knowledge, and exposure to new experiences. It is often difficult to separate changes due to maturation from those due to the project.
f.	Regression toward the mean	If subjects are selected because of their extreme scores (e.g., weight, physical development), there is a natural tendency to move closer to the mean over time, thus diminishing or distorting the apparent effects of the program.
g.	Testing	Being interviewed or tested may affect behavior or responses. For example, being asked about expenditures may encourage people to cut down on socially disapproved expenditures (cigarettes and alcohol) and spend more on acceptable items.
h.	Instrumentation	As researchers gain more experience, they may change how they interpret rating scales, observation checklists, and so on.
i.	Potential bias 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.

(Continued)

(Continued)

Guidelines for Checklist 1: Overall Design Quality: How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into the mixed-method approach?		
Question No.	**Question**	**Explanation**
j.	Information not collected from the right people	Sometimes information is collected from and about only 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. Some categories of informants are not interviewed.
3. Construct validity		
a.	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 behavior, hostile work environment, and sex discrimination.
b.	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 underrepresents the real number of incidents.
c.	Use of a 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, and domestic violence) will usually produce bias.
d.	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 (e.g., observing communities in formal meetings will produce different results than observing social events or communal work projects).
e.	Only one level of the treatment is studied	Often a treatment is only administered at one, usually low, level of intensity (e.g., 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.
f.	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.
g.	Program participants and comparison groups 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.
h.	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.
i.	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.

Guidelines for Checklist 1: Overall Design Quality: How suitable was the mixed-method design for the purposes of the evaluation? How well were the QUANT and QUAL procedures and design components implemented, analyzed, and integrated into the mixed-method approach?		
Question No.	**Question**	**Explanation**
j.	Experimenter expectancies	Experimenters have expectations (e.g., about how men and women or different socioeconomic groups will react to the program), and this may affect how they react to different groups.
k.	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.
l.	Using indicators and constructs developed in other countries without pretesting in the local context	Many evaluations import theories and constructs from other countries, and these may not adequately capture the local project situation. For example, many evaluations of the impacts of microcredit on women's empowerment in countries such as Bangladesh have used international definitions of empowerment that may not be appropriate for Bangladeshi women.
m.	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.
n.	Does multilevel, mixed-method analysis accurately reflect how the project operates and interacts with its environment?	The reviewer may consult with key informants, beneficiaries, and stakeholders, requesting them to review and comment on the descriptions included in the evaluation reports.

Guidelines for Checklist 2: Dimension 1: Internal Threats to Validity Level 2: Quality Assessment for Each Stage of the Evaluation		
Question No.	**Question**	**Explanation**
1B. Qualitative data collection		
a.	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?
j.	Did triangulation among complementary methods and data sources produce generally converging conclusions?	Where interpretations do not converge, are the differences in interpretation and conclusion noted and discussed?
k.	Are the data well linked to the categories of prior or emerging theory?	Are the findings internally coherent, and are the concepts systematically related?
2A. Quantitative data analysis		
a.	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.
b.	Some assumptions of the statistical tests have been violated.	Computer software analysis packages make it simple to run statistical tests without understanding the assumptions on which they are based. Sometimes this can invalidate or weaken the findings and how they are interpreted.
c.	"Fishing" for statistically significant results	If large numbers of regressions or other statistical tests are run, a certain number will come out "positive" due to the laws of probability. To avoid these spurious findings, it is important to base the analysis design on a set of hypotheses derived from the program theory.
d.	Restriction of range	Many evaluations only cover part of a population (e.g., only the poorest sectors are studied), and this can weaken the strength of some kinds of analysis.
e.	Unreliability of treatment implementation not captured in the analysis	If the treatment is not administered in an identical way to all subjects, the probability of finding a significant effect is reduced.
f.	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.
g.	Extrapolation from a truncated or incomplete database	If the sample only covers part of the population (e.g., 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.
h.	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 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.

Guidelines for Checklist 2: Dimension 1: Internal Threats to Validity Level 2: Quality Assessment for Each Stage of the Evaluation		
Question No.	**Question**	**Explanation**
i.	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.
j.	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 to conduct statistical significance testing. This is frequently the case when data are collected at the group level to save time or money.

Guidelines for Checklist 3: External Validity

	Question	Explanation
1.	Sample does not cover the whole population of interest.	*Subjects may only 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. Microcredit 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 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 nongovernmental organizations.*
6.	The attitude of policymakers and politicians to the program	*Identical programs will operate differently and have different outcomes in situations where they have the active support of policymakers 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, alternatively, 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, and so on described in enough detail to permit comparisons with other samples?	*This may require presenting information on the social and economic characteristics of the study population and of comparison groups.*
9.	Does the sample design theoretically permit generalization to other populations?	*Does the program theory explain the factors determining success or failure of the project, and are these presented in a way that makes it possible to assess the potential for replication in other settings?*
10.	Does the researcher define the scope and boundaries of reasonable generalization from the study?	*Does the analysis assess the extent to which these characteristics are likely to be present in the areas where the program might be replicated?*
13.	Do the findings confirm or are they congruent with existing theory? Is the transferable theory made explicit?	*Is there a review of existing theory, and does this discuss how the program theory relates to current theory? Is the program theory consistent with current theory, and if not, is there a credible case made for the differences?*

	Question	Explanation
14.	Are the processes and findings generic enough to be applicable in other settings?	*Does the study present sufficient evidence to show whether processes are unique to this setting or can be generalized? Is the evidence convincing?*
15.	Have narrative sequences been preserved? Has a general cross-case theory using the sequences been developed?	*Is a significant amount of the original narrative preserved or are only the researchers' notes and interpretations available? Is sufficient narrative available to test whether the findings and interpretations are consistent across cases?*

Guidelines for Checklists 4 and 5: External Validity: Explanations were not considered necessary for these two checklists.

APPENDIX 7.6 APPROACHES FOR ASSESSING VALIDITY OF MIXED-METHOD EVALUATIONS

1. The Standard Mixed-Method Worksheet

Appendix 7.3 presents a worksheet for what we call the "standard version" of the mixed-method validity assessment. This combines the main elements of the checklists for QUANT (Appendix 7.1) and QUAL (Appendix 7.2) evaluations. Each of the six checklists uses italics to indicate the items that are normally used in QUANT validity analysis and normal text for the items that are normally used in QUAL validity analysis. The worksheet has the same format and the same system of rating as for the two previous worksheets. This worksheet can be considered the default worksheet that can be used for most validity assessments, as in practice, most evaluations combine some elements of both QUANT and QUAL approaches.

Appendix 7.4 presents a completed version of Worksheet 3 that has been applied to a hypothetical but fairly typical project evaluation. This illustrates how the detailed technical assessments in Part 3 are summarized through the ratings and then presented in a condensed form in Part 2, which is intended for evaluation managers who are often not research specialists but who require a paragraph to explain the main findings for each of the six checklists. The findings are further summarized in Part 1 to provide for senior management and partner funding agencies a brief assessment of the strengths and weaknesses of the evaluation methodology and what this means for the interpretation of the findings and recommendations. In this example, due to the lack of baseline data on the comparison group, it is concluded that the evaluation's findings concerning the positive project impacts are not fully supported by the interpretation of the data.

2. A More Advanced Approach to the Assessment of Mixed-Method Evaluations (Appendix 7.5)

Over the past few years, mixed-method (MM) evaluation has emerged as a distinct approach to evaluation that is much more than simply a combination of different methods of data collection. Appendix 7.5 presents a worksheet that tries to capture and assess the complexities of an integrated MM evaluation approach. MM evaluations combine QUANT and QUAL conceptual frameworks, methods of hypothesis generation, data collection, analysis, and interpretation. This means that a two-stage assessment of validity is required. In Stage 1, each element of the evaluation design and implementation must be assessed using the appropriate QUANT or QUAL validity criterion. Then, in Stage 2, the overall evaluation design must be assessed to determine how well the QUANT and QUAL components have been integrated into the MM design. The validity analysis is further complicated by the fact that evaluation designs form a continuum from mainly QUANT designs to mainly QUAL designs (see Chapter 14), so there is an element of judgment in deciding which components of an MM evaluation design should be assessed in terms of QUANT threats to validity, which in terms of QUAL dimensions of adequacy and trustworthiness, and which may require the application of both QUANT and QUAL criteria.

Appendix 7.5 presents a worksheet with a set of five components for assessing the validity of MM designs. The MM validity analysis is based on the three dimensions (internal, external, and utilization validity), but a number of additional indicators have been added to reflect the unique nature of the mixed-method approach (see Chapter 7, Figure 7.3):

a. During the process of data analysis, MM will sometimes *quantitize* QUAL variables by transforming them into QUANT variables and *qualitize* QUANT variables by transforming them into QUAL variables (see Teddlie & Tashakkori, 2009, Chapter 11). This is defined as a *conversion mixed-method design.* Consequently, for these designs, in addition to assessing QUANT and QUAL data collection, it is also necessary to assess the adequacy of the process of transformation of the variables.

b. One of the strengths of MM approaches is that they permit multilevel analysis, for example, to examine the interactions between the project-implementation process and individuals, households, communities, and organizations and to model how these interactions affect project implementation and outcomes. This requires an assessment of the whole modeling process as well as the adequacy of the data collection and analysis at each of the levels (see Teddlie & Tashakkori, 2009, Chapter 11).

c. Most MM designs combine the use of separate QUANT and QUAL techniques for data collection, analysis, and interpretation, and these must be examined separately using the appropriate QUANT and QUAL assessment criteria (see Appendix 7.5, Checklist 2).

d. The overall evaluation design must then be reviewed to assess how well the QUANT and QUAL components have been integrated to produce a true MM design as opposed to the separate use of QUANT and QUAL techniques that are used separately and not integrated into a true MM design (see Appendix 7.5, Checklist 1).

e. Finally, a new element must be introduced to assess the overall interpretative rigor with which the MM framework has been applied. This is required because MM designs are required to integrate both QUANT and QUAL approaches to the development of conceptual frameworks and the definition and testing of hypotheses. It is quite possible for an evaluation to include methodologically sound interpretations of both QUANT and QUAL data but fail to ensure their integration into a fully integrated MM interpretation of all aspects of the evaluation (see Appendix 7.5, Checklist 5).

APPENDIX 7.7 POINTS DURING THE RWE CYCLE AT WHICH CORRECTIVE MEASURES CAN BE TAKEN

The RWE Integrated Checklist (Appendix 7.3) and other similar checklists can help identify threats to validity and adequacy of designs, as well as strategies for addressing the threats once they have been identified. Threats can be addressed at three points in the evaluation: during design, during implementation, and during report preparation. Some of the questions to be addressed when assessing the adequacy and validity of evaluation designs include the following:

- Are methods appropriate for information needs? For example, is the evaluation question to be addressed in data collection one of meaning, understanding, or process? Will the methods generate data that will help the client and other stakeholders address their information needs?
- Is there a clear, positive relationship of the methodology to the evaluation focus or questions?
- Are the evaluation team's expertise and capacity sufficient for the approach and methods?
- Are strategies to ensure validity, trustworthiness, and confirmability adequate and appropriate?
- Using this design, are the data to be collected likely to be relevant, comprehensive, and representative of the program?
- Are adequate and appropriate technical and ethical safeguards in place?
- Are there procedures to ensure cultural sensitivity, competence, and capacity?

Increasing the comprehensiveness of the methods would improve many evaluations—for example, increasing the period over which data is collected and number of observations or increasing the number of stakeholders and stakeholder groups interviewed or surveyed. Adding experts in the application of these methods (e.g., sociologists, statisticians) or in the type of program to be evaluated (e.g., economists to study a microcredit program, civil engineers to study a clean-water program) can also strengthen evaluation designs. However, in many RWE contexts, budget and time constraints limit the ability of the evaluation team to make these additions.

Strengthening the Evaluation Design

The Integrated Checklist (see Appendix 7.3) can be used to review a proposed evaluation design and make corrections before the evaluation begins. If analysis by an evaluator, evaluation team, or technical advisory panel determines that an evaluation design is weak, steps may be taken to improve the design, protecting the quality of the evaluation before data collection begins. The specific improvements needed vary from design to design, of course, but some general advice may be helpful.

Strengthening Data-Collection Methods

As the evaluation focus, questions, and information needs are considered with reference to the planned methods, it may become clear that the methods may not elicit adequate or strong enough data to provide sufficient evidence of program quality. The substitution of more appropriate methods or addition of complementary methods may be needed.

For example, in an evaluation of state assessment programs in education, an examination of methods may suggest that the evaluation will generate data describing the types of achievement testing used by the state, the types of items on the tests, whether the tests are standards based or norm referenced, how they are scored, how scores are reported, and the proportions of students across time who are passing and failing the tests. Such information can be obtained from

state documents and online sources, perhaps supplemented by interviews with state education agency personnel. The combined use of several or all of these methods would provide extremely useful data for judging the quality of the state assessment system.

But these data would not be sufficient to determine how much measurement error is included in the tests, whether the tests are well aligned with state content standards and state curriculum guides such that teaching is likely to provide appropriate preparation for test takers, and whether failure rates impose inappropriate negative consequences on students and educators. For information of this type, more detailed study of the test content, content standards, and standards-based curricula are needed. Interviews with a representative sample of teachers, school administrators, students, and parents are needed to provide the perspectives of critical stakeholders regarding whether the state testing program is strengthening or corrupting educational practices and outcomes—and these interviews need to be followed up with surveys to determine how widespread the experiences of those interviewed are in the affected population. Observations in classrooms are needed to determine whether standards-based curricula are integral to learning opportunities and, therefore, aligned to standards-based tests. The design needs to expand from review of documents and state-level interviews to classroom-level observation, interviews of a variety of stakeholder groups, and general surveys.

In such a case, a wider range of methods and data sources would often provide triangulation, strengthening the validity protections in the evaluation design. More carefully focused methods—in this case, getting into the details of test content, measurement soundness, and the appropriateness of score-based consequences—allow for the refinement and validity of findings. Such findings would be more informative, better supported by the data, and more useful for state assessment directors who may be hoping to improve their systems on the basis of evaluation data and findings.

Strengthening Capacity of the Evaluation Team

So that the benefits of different types of methodological approaches and techniques can help protect the soundness of the evaluation and the validity of its interpretations, specialized types of methodological expertise may be needed on the evaluation team. QUANT team members may be needed to ensure appropriate sample sizes for data to be used in representing populations or determining statistical significance, comparable experimental groupings, and selection and use of statistical tests and procedures. QUAL team members may be needed to ensure a fine-grained focus on individually and socially constructed meaning, the human implications of program effects, and the identification and magnitude of unintended consequences.

The complementarity of these approaches—that is, using mixed methods—can be extraordinarily useful in producing high-quality evaluation. It takes considerable open-mindedness and managerial expertise, however, to find working consensus among disparate points of view. A positional stance that privileges macro-level descriptors of the program, the population overall, or aggregated results and costs may be dismissive of micro-level effects on individuals, their experiences, and their perspectives of program quality. From an opposing positional stance, macro-level indicators and trends may be accorded little meaning if intended beneficiaries disagree with numbers that suggest they should be satisfied with program outcomes they may actually find unsatisfactory. Working consensus, rather than melding or integration, may be useful for preserving the benefits of different methodological perspectives and expertise.

Often, an evaluation benefits from a variety of methodological specialties. Professionals in the field of the program who are familiar with relevant theory, professional literature, and best practices are as critical to the strength of data analysis as evaluators trained and experienced with inquiry procedures, including different strategies of data analysis. The situation is analogous to that in critical thinking: General critical-thinking strategies can be taught, but one must be able to think critically within a content area, where general strategies may be greatly differentiated and augmented. Analysis attentive to the literature in the field of the program (e.g., the field of epidemiology in programs intended to combat HIV/AIDS or the field of early childhood education in programs intended to reduce school drop-out rates through early intervention) is as important as methodological strength.

For example, survey data in an evaluation of a state grant program to support the inclusion of children with disabilities suggested that all the 150 or so grantees had included, as intended, children with disabilities in prekindergarten,

kindergarten, and first-grade classrooms of typically developing peers. Telephone interview data, however, suggested that some of the grant programs provided more inclusion than others. Observations and interviews at daylong site visits and a few weeklong case studies indicated that, at some sites, no inclusion was practiced. While all members of the evaluation team were able to recognize when programs segregated special-education from regular-education classrooms, initially, only the special-education members of the evaluation team were able to distinguish whether classrooms with mixed populations met the definition of *inclusion* or segregated regular-education from special-education students within the same physical spaces. Only the trained special-education eyes perceived whether instruction was differentiated and developmentally appropriate.

Determining the point of entry of such content specialists in the evaluation of a program and the scope of the role they might play is a consideration in strengthening evaluation design and conclusions. In an ideal world, of course, such resources would be abundant for an evaluation. However, in the real world, balancing the costs of differentiated expertise (feasibility) with the value to the evaluation (quality, validity, credibility) is almost always necessary.

Strengthening the Implementation of the Evaluation

The RealWorld Evaluator is always encouraged to consider the use of mixed-method approaches. One of the strengths of these approaches is that QUAL methods can be used in parallel with QUANT data-collection methods in a compensatory manner. For example, in the evaluation of a community water-supply project, surveys may not reveal that residents have to pay bribes to local leaders to obtain water, information more likely to be revealed through ethnographic observation.

Strengthening Data-Analysis Procedures

Too often, data analysis progresses no further than trends or correlations, sometimes in combination with calculations of statistical significance. Consider a trend toward increased numbers of children from an indigenous population in the highlands of Papua New Guinea in English-speaking primary schools. The trend itself is not a finding; it must be interpreted. Should the interpretation be that education is improving for indigenous children living in the highlands of Papua New Guinea? If analysis ventures no further than enrollment figures, using easy-to-analyze simple QUANT procedures, important underlying explanations may be missed.

On the other hand, analysis complicated by both QUANT and QUAL data, including ethnographic observations and systematic interviews, requires more complex analysis. Content analysis may, for example, make it much less clear whether education in English-language primary schools is beneficial for children in Papua New Guinea. While English-language primary schools may improve children's chances to succeed in higher education, those chances may come at the risk of extinguishing the indigenous language and culture. Records kept by state officials are unlikely to include such data and important stakeholder considerations, while sensitive inquiry—in this case, such as that by Malone (1997)—can offer more comprehensive and informative analysis.

Strengthening the Evaluation When Preparing to Report

Many evaluators submit draft reports to clients for review, comment, and correction prior to finalization. This practice is sometimes referred to as *negotiation drafts*. Feedback from clients may identify issues or gaps needing attention. Some time and budget should be reserved to permit additional fieldwork or analysis, if called for. These may include follow-up clarification interviews with key informants, focus group interviews with newly identified groups of stakeholders, or site visits to newly identified natural comparison groups.

APPENDIX 7.8 FACTORS DETERMINING THE ADEQUACY OF AN EVALUATION DESIGN AND THE VALIDITY OF THE FINDINGS

The adequacy of an evaluation design and the validity of the findings and conclusions resulting from this design is contingent on a number of factors:

- How well suited the evaluation focus, approach, and methods are for obtaining the types of information needed, for example:
  - Information regarding managerial decision making. Does the evaluation focus on program procedures, personnel, and product quality? Does the evaluation design take into consideration organizational infrastructure, resources, training, safety, and access as well as outcomes?
  - Information regarding stakeholder perspectives of program adequacy. Does the evaluation design include procedures for understanding the experiences and perceptions of intended beneficiaries? Are the methods to be used sensitive to the gender, cultural, and linguistic characteristics of stakeholders?
- Availability of data and data sources, for example:
  - Whether appropriate data exist or can be generated. Do records provide accurate information about how, when, and to whom program services and benefits are delivered? Are financial records available and accurate, and have they been audited?
  - Are all stakeholders accessible to evaluators? Can stakeholders be identified and located? Are they willing and able to provide data? Are there language, cultural, political, or other barriers to their provision of information?
- Will the data support valid interpretations about the program, for example:
  - The achievement of program goals. Which program objectives were accomplished, and how well? Did some intended beneficiaries fare better than others? Which factors, if any, proved critical to program success? What, if anything, hindered or undermined goal attainment?
  - Cost-effectiveness of the program. How accurate and comprehensive are the program's financial records? What was the cost of program delivery per beneficiary (or other unit of analysis)? Did benefits justify the cost of the program? How do the costs and benefits of this program compare with similar programs?
  - The extent of delivery of program benefits. Did all intended beneficiaries receive benefits as planned? Did all appropriate stakeholders enjoy sufficient and equal access to program benefits or services? Should benefits or services have been accessible to a larger group than those defined by the program? Was there any significant "leakage" of benefits to groups not entitled to receive them? Were the benefits readily available or hard to obtain?
  - The adequacy of resources affecting goal attainment. Were funding, personnel, and other resources sufficient for satisfactory program implementation? Where funding was authorized, were needed resources actually available within the program's context?
  - Unintended consequences. Were there unexpected benefits resulting because of the program? How important were they? Were there negative side effects? To what extent did they undermine or counteract benefits?
- The professional expertise and knowledge of the evaluation team in terms of both evaluation methodology and the specific sector or field of the program, for example:
  - Expertise in terms of evaluation methodology. If content analysis of QUAL data is part of the design, does the team include specialists in QUAL methodology? If a cost-effectiveness analysis is part of the design, does the team include financial analysts?
  - Expertise in terms of the specific field of the program. If evaluating a well-child program, does the team include medical and public health specialists? If evaluating an adult literacy program, does the team include specialists in reading curriculum and pedagogy for adult education?
  - Capacity of evaluation resources for the scope of the program. Are there enough interviewers to collect the interview data? Are they sufficiently trained, or are there sufficient resources (expertise and budget) to provide training? Is there sufficient evaluation capacity in terms of database development and capacity, and statistical analysis of QUANT data? Are there adequate logistical resources for communication, transportation, and other needs?

APPENDIX 7.9 EXAMPLES OF OTHER CHECKLISTS USED TO ASSESS EVALUATION QUALITY AND VALIDITY

In addition to the checklists developed by Shadish et al. (2002), Guba and Lincoln (1989), and Miles and Huberman (1994) referred to earlier in the appendices, a number of other checklists have been developed for assessing the quality of evaluations or the validity of their conclusions (see also Chapter 9 for a discussion of evaluation guidelines and standards). The following are some widely used examples:

A. *The Western Michigan University Evaluation Checklist Project*[8] provides refereed checklists for designing, budgeting, contracting, staffing, managing, and assessing evaluations of programs, personnel, students, and other evaluands; collecting, analyzing, and reporting evaluation information; and determining merit, worth, and significance. Each checklist is a distillation of valuable lessons learned from practice. The site's stated purpose is to improve the quality and consistency of evaluations and enhance evaluation capacity through the promotion and use of high-quality checklists targeted to specific evaluation tasks and approaches. The checklists are classified into the following groups: evaluation management, evaluation models, evaluation values and criteria, meta-evaluation and evaluation capacity development, and institutionalization. Many but not all of the checklists focus on the education sector.

   The site includes a number of widely cited checklists, among which are Michael Scriven's *Key Evaluation Checklist* (2007); Daniel Stufflebeam's *Program Evaluations Metaevaluation Checklist* (1999); Daniel Stufflebeam's *CIPP Evaluation Model Checklist* (2007); and Michael Patton's *Qualitative Evaluation Checklist* (2003) and *Utilization Focused Evaluation Checklist* (2002c).

B. The American Evaluation Association's *Guiding Principles for Evaluators* (2004), which provides guidance to evaluators on five areas: systematic enquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare.

C. The OECD/DAC *Quality Standards for Development Evaluation* (2010b). This provides standards for (1) the rationale, purpose, and objectives of an evaluation; (2) evaluation topic; (3) context; (4) evaluation methodology; (5) information sources; (6) independence; (7) evaluation ethics; (8) quality assurance; (9) relevance of the evaluation results; and (10) completeness.

While these are extremely valuable resources, they focus mainly on ensuring quality in the design and implementation of the evaluation and ensuring that evaluators follow appropriate professional standards. There is little direct discussion of threats to validity and how they can be addressed.

8. For information on the Western Michigan University Evaluation Center Checklist Project, see http://www.wmich.edu/evalctr/checklists/checklist_topics.

APPENDICES FOR CHAPTER 10 THEORY-BASED EVALUATION AND THEORY OF CHANGE

10.1 Results-Based Reporting and Logical Frameworks

10.2 The Two Components of a Program Theory Framework: Program Impact Models and Implementation Models

Chapter 10 discusses the applications of program theory in evaluation. Theory-based evaluation and particularly theory of change underpins many evaluations by defining the logic through which program inputs and activities are intended to produce outputs, outcomes, and finally impacts. A well-articulated theory of change will also identify the hypotheses intended to explain the processes through which the intended changes are to be achieved. The chapter also explains the linkages between theories of change, logical frameworks, and results frameworks.

Two appendices are included for this chapter: Results-based management and logical frameworks (Appendix 10.1), and the two components of a program theory framework: program impact and program implementation models (Appendix 10.2).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 10.1 RESULTS-BASED REPORTING AND LOGICAL FRAMEWORKS

As discussed in Chapter 10, Section 1, a theory of change may either provide the framework for a stand-alone evaluation or be part of a program management system that often includes a results framework (where intended program outputs, outcomes, and impacts are translated into monitorable indicators) and a logical framework, where the program design is represented graphically (see Chapter 10, Figure 10.1).

1. Results-Based Management (RBM) Systems

RBM systems translate program theory into sets of measurable indicators so that progress can be tracked and factors determining achievement or non-achievement of outputs and impacts can be assessed. Over the past decade, most international development agencies and many governments and NGOs have introduced a results-based management (RBM) approach that defines and measures results rather than just monitors outputs. This permits development agencies to better assess whether the resources they provide and the initiatives they support have contributed to achieving their development objectives. Many RBM models include multiple cause–effect chains in a pyramid-type graphic, rather than a single linear chain that fits neatly into the classical logframe matrix. Morra-Imas and Rist (2009) present a 10-step approach for the design and implementation of a results-based M&E system.[9]

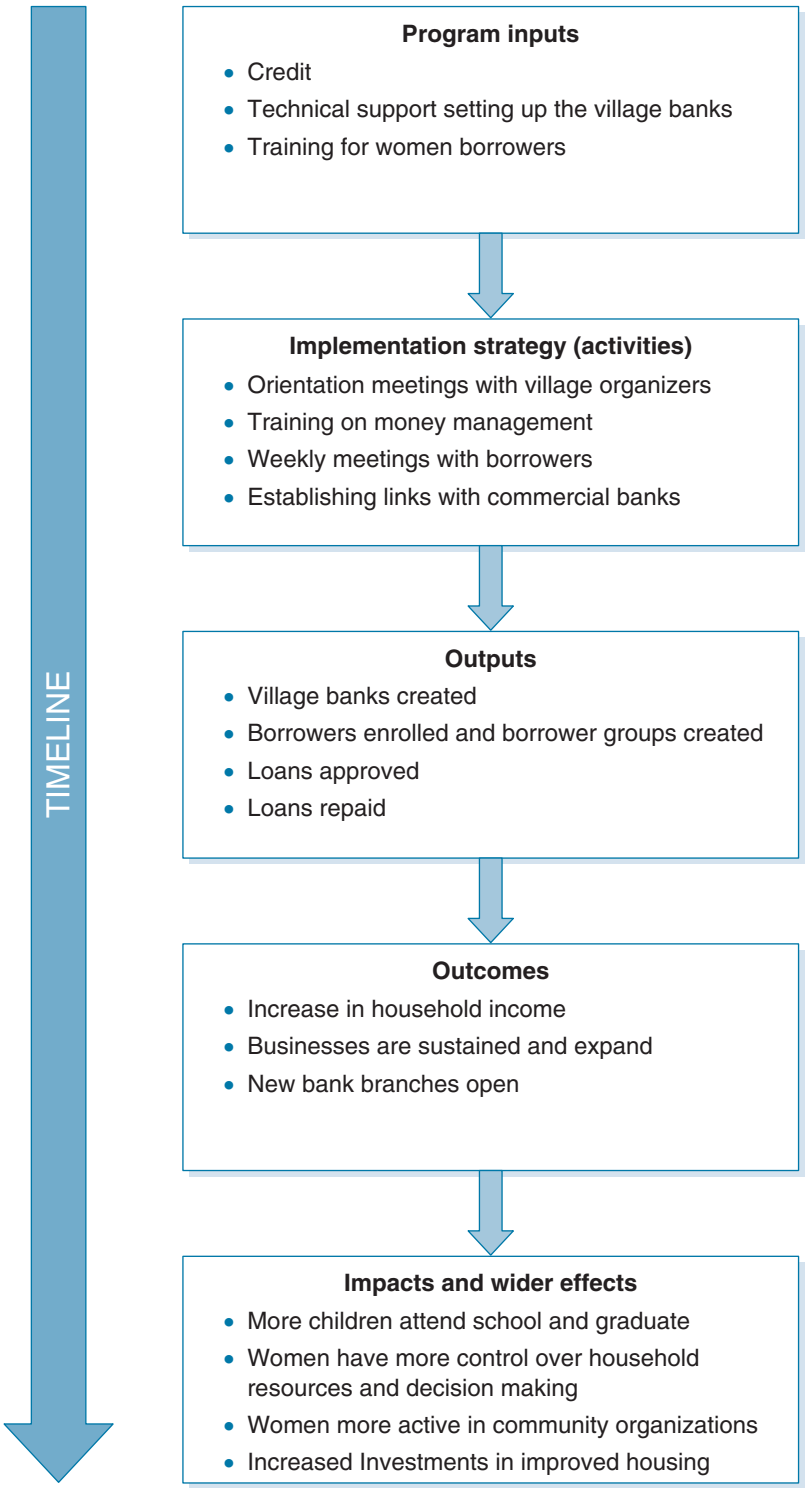
2. Logical Framework (Logframe)

A logframe presents a graphical representation of the program theory. It identifies the linkages between inputs, activities, outputs, and the intended outcomes, impacts, and broader program goals. Ideally the logframe should also identify the critical assumptions on which the choice of inputs, the selection of implementation processes, and the expected linkages between the different stages of the program cycle are based. Figure A10.1-1 presents a typical logframe for a hypothetical program to promote women's economic empowerment through microcredit. The figure identifies five stages of the program:

- *Program inputs.* These are usually a combination of financial, infrastructure, equipment, materials and supplies, and technical assistance. In the present example, inputs include credit, technical support, and training for women borrowers.
- *Implementation strategy (activities).* These explain the activities that will be implemented to transform inputs into outputs. In the present case, activities include orientation meetings, training on money management, weekly meetings with borrowers, and establishing links with commercial banks.
- *Outputs.* These are the intended products that the program is committed to producing. Many of these will be tangible, quantifiable products such as buildings or people provided with services, but they may also include less easily measurable outputs such as women's increased participation in meetings. In the present case, outputs include the number of village bank branches opened, the number of borrowers, and the number of loans approved and repaid.
- *Outcomes.* While outputs are largely under the control of the implementing agency, outcomes are the short- and medium-term effects resulting from the availability of the outputs. The agency has much less control over these as they depend on the attitudes and behavior of project beneficiaries, often influenced by outside circumstances over which the project may have only limited control. In the present example, outcomes include increased household income, businesses are sustained and expand, and new bank branches are opened.
- *Impacts and wider effects.* These are the medium- and longer-term changes that result from the program interventions. There is no clear-cut distinction between outcomes and impacts, and to some extent the difference is a matter of judgment. Impacts are considered broader and more significant and occur over a longer period of time, but they must be interpreted within the local social, economic, political, and cultural context. In some microcredit programs in very conservative rural areas, the ability of a woman to travel to the local market on her own and without her husband's permission may be considered a major impact. However, in another area this may only be considered an outcome or perhaps even an output.

[9] For examples on how the World Bank uses results frameworks in their project reports, see http://documents.worldbank.org/curated/en/docsearch/document-type/791001.

FIGURE A10.1-1 ● A Logic Model for a Hypothetical Program to Promote Women's Economic Empowerment Through Microcredit



One of the important and very useful elements of the logframe is that it identifies some of the critical assumptions about the linkages between the different stages of the model. Table A10.1-1 illustrates critical assumptions that might be included at each stage of a logframe of the women's economic empowerment program. For example, the use of credit as the major input is based on two assumptions: first, that lack of access to credit is one of the main constraints on women's ability to start a small business; and second, that if women have access to credit, this will significantly increase the number and sustainability of small businesses they start. Both of these assumptions can be tested, and their correctness will be an important determinant of the project's success. Similar assumptions can be identified and tested for each stage of the model.

Importantly, this logic model also includes a projected timeline over which the different stages of the model are expected to be achieved. As we will see later, the timeline describing when (usually in years) the different stages of the project are scheduled to start and to produce their respective outputs, outcomes, and impacts can be critical in the design of the impact evaluation. When the timeline is not included, there is a risk that outcome or impact evaluations will be scheduled when it is too early in the project cycle for them to have been achieved and to be measurable.

TABLE A10.1-1 ● Testing Critical Assumptions in a Logic Model of a Project to Strengthen Women's Economic Empowerment Through Microcredit

Stage of Project	Critical Assumptions to Be Tested
Design	• Poor women have the skills needed to operate viable income-generating projects but lack only capital. • Women are able to decide what business to start/expand. • Women will be able to control how the loan is used, and the money will not be appropriated by the husband.
Inputs	• Access to credit, in a form that the woman can control, is critical to enhance women's access to economic opportunities.
Implementation process	• The creation of solidarity groups through which loans are approved and technical support provided is essential to enable women to control their use of their loans and to manage their small businesses. • Solidarity groups must select their own members without any outside pressures.
Outputs	• Women will use loans to invest in small businesses (not just to pay off debts or pay for consumption or ceremonial activities). • Women will be able to control the use of the loan (despite cultural traditions that economic resources are controlled by male household members).
Outcomes	• If women produce goods, they will be able to market them. • Their businesses will be profitable. • Women will control, or share in the control of, the profits.
Impact	• Profits will increase household consumption, women's savings, and quality of life of members of their households.
Sustainability	• The women's solidarity groups will be able to continue providing loans after the project's external credit and support has ended. • Their businesses will continue to operate and to grow.

APPENDIX 10.2 THE TWO COMPONENTS OF A PROGRAM THEORY FRAMEWORK: PROGRAM IMPACT MODELS AND IMPLEMENTATION MODELS

A program theory includes both descriptive assumptions about the causal processes, explaining the social problems a program is trying to address, and prescriptive assumptions about the components and activities that program designers and other stakeholders consider necessary to a program's success. While many authors propose a single logic model, others use two separate models to describe the program impact model (based on the theory of change) and the program implementation model.

Figure A10.2-1 illustrates an impact (change) model for the women's small-business development model. The middle level of the figure includes the same stages as the logic model (except that outcomes and impacts are combined for reasons of space). The implementation strategy box illustrates how questions, in this case about the implementation strategy, can be included as well as the activities. However, the main difference is that the model also includes five sets of contextual factors and a set of mediators (factors affecting project implementation and outcomes that can be modified by the project). While both the contextual factors (economic, political, policy, institutional and organizational, environmental, and socioeconomic characteristics of the affected populations) and the mediators can affect project implementation and outcomes, the difference is that the project has little or no control over the contextual factors, but it can influence the mediators. For example, there are ways to make men more willing to allow their wives to work outside the home, and to help wives to gain more confidence to visit a bank to apply for a loan.

Descriptive assumptions are generally based, at least in part, on an analysis of available research and evidence from other programs and often include a needs assessment or rapid diagnostic study of the social, economic, cultural, security, and other characteristics of the subjects or communities that the program is intended to affect. *Prescriptive assumptions*, on the other hand, are based on judgments and values about which intervention strategy should be selected. These assumptions may be based on a review of earlier projects or consultation with specialists, or they may be largely based on personal values concerning what is the "right" way to address the problems.

The descriptive assumptions are translated into a *program impact model* (Donaldson, 2003) articulating the assumptions about the causal processes underlying the decisions to use certain program strategies. Some program theories also develop a *program implementation model* describing how the program will be organized to achieve the intended outcomes and impacts. Depending on the focus of the evaluation, it may be that only one of these two models will suffice, often the case for RWE operating with budget and time constraints. If the purpose of the evaluation is to assess the achievement of goals and impacts, then the impact model will probably be sufficient. If, on the other hand, the purpose is to assess effective implementation and the production of outputs in order to help improve performance (formative evaluation), then the implementation model may be preferred. In some cases, and resources permitting, it may be useful to use both models.

Program Impact Model

Different terminology has been used to describe the components of an *impact model* (Bamberger, Mackay, & Ooi, 2004; Chen, 2005; Creswell, Clark, Gutmann, & Hanson, 2003; Donaldson, 2003; Rossi, Lipsey, & Freeman, 2004), sometimes indicated by terms such as *change model*.[10] Figure A10.2-1 illustrates a program impact model for a women's microcredit project. This model integrates the following components:

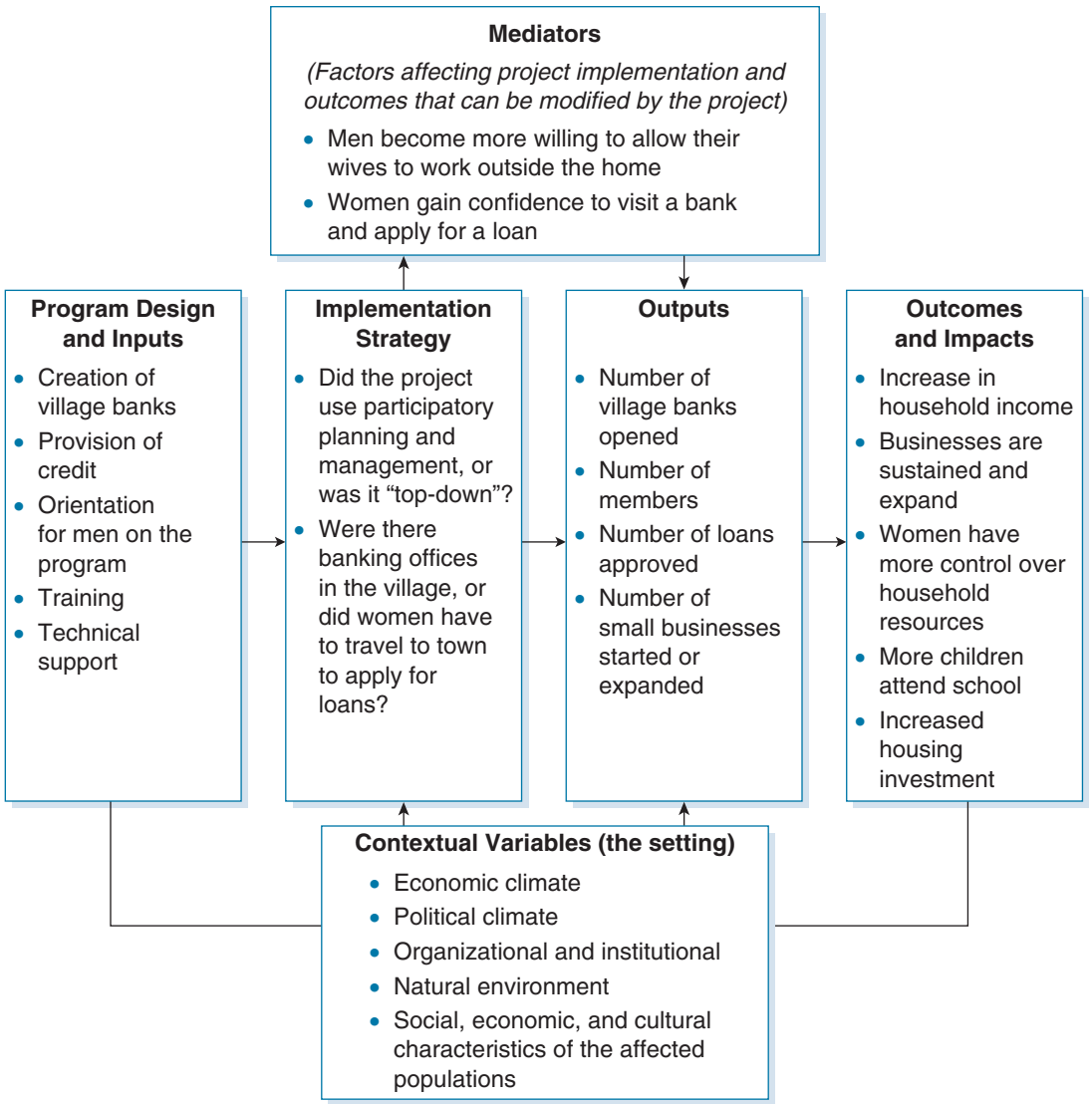
Contextual Variables (the Setting).[11] A contextual analysis is conducted to identify and understand the social, economic, psychological, security, and environmental setting within which the project operates and the problems, needs, priorities, and constraints of the intended project beneficiaries and the different stakeholders. Projects implemented in the same way in different communities, schools, or regions can have very different outcomes because of differences in these contextual variables. This analysis may cover the following:

[10] Chen (2005) uses the term *change model*.

[11] This component is not included in the Chen (2005) and Donaldson (2003) models. However, conceptual factors are similar to what Donaldson (2003) defines as *moderators*.

- *The economic climate.* Are economic conditions getting better, remaining constant, or getting worse? This will influence the decisions of families as to whether or not they want to participate in any project that either requires payments or promotes present or future income-generating activities.
- *The political climate.* Is the local political climate likely to support or undermine the project?
- *Organizational and institutional factors.* To what extent do local organizations (government, NGOs, and private sector) support or hinder the project?
- *Natural environmental factors.* In what ways do environmental factors influence the project?
- *The characteristics of the communities affected by the project.* How do social, cultural, economic, and other characteristics influence how different groups respond to the project or are affected by it? How might the needs, problems, constraints, and expectations of the different groups affect the project?

FIGURE A10.2-1 ⬢ An Impact Model for a Women's Small-Business Development Program



The contextual analysis can be derived from a review of the literature, the opinions of stakeholders, a rapid assessment study, or a needs assessment during which the target group is consulted. The needs assessment might use a market research approach in which a target group is surveyed and asked to select among a set of options relating to the planned program strategies, or it might use a participatory assessment method in which families identify their concerns, needs, and proposed solutions (see, for example, Dayal, van Wijk, & Mukherjee, 2000; Rietberger-McCracken & Narayan, 1997; Theis & Grady, 1991). In a microcredit example (see Figure 10.4 in Chapter 10), a contextual analysis identified a number of constraints on women's access to economic opportunities, including labor laws, labor union protection of male workers, and women's difficulties in obtaining bank loans.

Program Design and Inputs. *Inputs* refer to resources allocated to program efforts to achieve objectives and goals (e.g., money, staff, materials, trainers, vehicles, teaching materials).[12] This includes what some authors call *interventions*. Although some interventions are intended to directly achieve program goals (for example, emergency food or medical supplies), in most cases, programs are designed to influence mediators (see below) through changing knowledge, attitudes, or practices. In the case of a women's small-business development project, interventions might include the organization of women's savings and loans groups and orientation sessions to reduce the opposition of men to their wives starting their own businesses, in addition to the provision of credit.

Implementation Strategy. The implementation strategy refers to the way in which inputs are used to produce the planned outputs and achieve the desired effects. Many agencies refer to these as *activities*. Two programs using the same inputs might achieve different results depending on their implementation strategies and the quality of how their activities are carried out. For example, one project may use participatory planning and management in which stakeholders were actively involved in program design, implementation, and monitoring, while another uses top-down planning with design and management by the client or funding agency. Similarly, one microcredit project might open an office in a community, while another might require women to travel to the nearest town to apply for loans.

Mediators.[13] Mediators are intervening variables potentially affecting project performance that the project may be able to influence (Donaldson, 2003). Most programs are influenced by many factors that exert influence on program outcomes. There are, of course, many external, contextual variables that are outside the direct control of an intervention program. What we are referring to here are those factors that are essential to motivating people to participate in a program, or potential barriers to such participation. For example, in the case of a health program, mediators might include how a person's course of action is influenced by his or her perceived susceptibility to illness, perceived seriousness of the problem's consequences, perceived benefits of a specific action, and perceived barriers to taking action (Strecher & Rosenstock, 1997, cited in Chen, 2005, p. 21). For a microcredit program, mediators could be the willingness of husbands to allow their wives to work outside the home following an orientation session on the program, or poor rural women gaining sufficient confidence, after joining a village banking group, to visit the local bank to apply for a loan.

Outputs. Outputs are the immediate results that a project seeks to achieve, and over which it has direct control. In the case of a small-business development project, these might include the creation of a certain number of village banks with a certain number of members and the authorization and repayment of a certain number of loans.

Outcomes and Impacts. These are the short-, medium-, and long-term changes in behaviors or conditions or systems that the project hopes to affect, the achievement of which it may influence but does not have direct control. Again referring to the microcredit example, outcomes and impacts might include the number of businesses started or expanded, the additional income earned, the increase in women's control over family resources and role in family decision making, and the improvements in family living conditions, such as increased school attendance, better health, and increased investment in housing. A project's implementations will hopefully make significant contributions to such impacts, but there are many other factors that affect such changes, positively or negatively.

---

[12] Some authors, for example, Chen (2005), use the term *interventions* to cover both the resources (what we call *inputs*) and the way the resources are used (what we call the *implementation strategy* or *activities*). However, we consider it helpful to distinguish between the two because the same inputs can have very different effects depending on the implementation strategy and how well the activities are carried out.

[13] Chen (2005) combines moderators and mediators into a single category, *determinants*, but the present authors consider it better to separate the two because they operate in different ways (see Donaldson, 2003).

APPENDICES FOR CHAPTER 11
EVALUATION DESIGNS

11.1 A More Detailed Look at the RealWorld Evaluation Design Frameworks

11.2 The RWE Approach to the Classification of Factors Affecting the Choice of Evaluation Design

11.3 The Strengths and Weaknesses of the Seven RWE Evaluation Design Frameworks

11.4 Challenges Facing the Use of Experimental and Other Statistical Designs in RealWorld Evaluation Contexts

11.5 Examples of Randomized Control Trials

Chapter 11 reviews all of the quantitative, qualitative, and mixed-method evaluation designs that can be used to assess program outcomes and impacts under different scenarios: depending on the stage of the program when the evaluation starts and ends, whether an experimental approach is possible, and whether there will be a comparison group. This includes potential applications of big data and data analytics. The chapter concludes by presenting a number of strategies that can be used to strengthen RealWorld Evaluation designs that must be conducted under different kinds of constraints.

There are five appendices for this chapter: a detailed look at all of the most common evaluation design frameworks (Appendix 11.1); the RWE approach to the classification of factors affecting the choice of evaluation design (Appendix 11.2); the strengths and weaknesses of the seven RWE evaluation design frameworks (Appendix 11.3); challenges facing the use of experimental and other statistical designs in RealWorld Evaluation contexts (Appendix 11.4); and examples of randomized control trials (Appendix 11.5).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 11.1 A MORE DETAILED LOOK AT THE REALWORLD EVALUATION DESIGN FRAMEWORKS

This appendix describes in more detail all the designs described in Chapter 11. The designs are presented in the same order as in Table A11.1-1 (which reproduces Table 11.3 from Chapter 11).[14] Examples are given illustrating how each of these designs has been applied in the field. All the designs can be strengthened if used in combination with the methods described in Chapter 11. Table A11.1-3 (reproduced from Table 11.5) summarizes some of the strategies that can be used to strengthen all of the designs discussed in this appendix. First the longitudinal (Design 1) and then the randomized control trial (RCT; Design 2.1) designs are discussed in more detail to explain some of the basic procedures used in all the subsequent designs.

To make it easier for the reader to navigate all the evaluation designs addressed in this rather long appendix, we provide below a Table of Contents just for Appendix 11.1.

Table of Contents for Appendix 11.1

1.	**Longitudinal Design With Pre-, Midterm, Post-, and Ex-Post Observations on the Project and Comparison Groups (Design 1)**	
1.1	When to Use	
	Table A11.1-1	List of Experimental, Quasi-Experimental, and Nonexperimental Evaluation Design Options, With a Focus on How the Counterfactual Is Determined
1.2	Description of the Longitudinal Design	
	Table A11.1-2	Design 1: Longitudinal Design
	Box A11.1-1	An Example of Design 1.1 (Longitudinal Design With Randomized Assignment): The Evaluation of GAIN: A Welfare-to-Work Program in California
	Box A11.1-2	A Longitudinal Evaluation Combining Design 2.1 (Randomized Design) for One Component and Design 2.2 (Quasi-experimental Design) for the Other Two Components: The Bolivian Social Investment Fund
1.3	Tools and Techniques for Strengthening All of the Basic Impact Evaluation Designs Described in This Appendix	
	Table A11.1-3	Tools and Techniques to Strengthen Any of the Basic Impact Evaluation Designs (Reproduction of Chapter 11's Table 11.5)
1.4	Threats to Validity and Adequacy of the Conclusions	
	Table A11.1-4	Some Threats to Validity That Must Be Checked for Designs 1 and 2
2.	**Strong Statistical Designs**	
2.1	Randomized Control Trial (Design 2.1)	
	When to Use	
	Description of the Design	
	Table A11.1-5	Experimental Design 2.1: Randomized Control Trial (RCT)
	Box A11.1-3	Example of Design 2.1: Randomized Control Trial Testing a Vaginal Microbicide Gel to Reduce the Likelihood of HIV Infection

(Continued)

[14] The numbering of the designs discussed in this appendix follows the numbering used in Chapter 11's Table 11.3, in that the lead number reflects the number of the seven basic designs of Table 11.2, and the second number represents a variant of that basic design.

(Continued)

2.2	Pretest–Posttest Project and Comparison Group Designs With Statistical Matching (Designs 2.2–2.5)	
	When to Use	
	Description of the Design	
	Table A11.1-6	Design 2.2: Pretest–Posttest (Quasi-Experimental) Design With Statistical Matching of the Two Samples
	Box A11.1-4	Example of Design 2.2A: Pretest–Posttest Comparison With Statistical Matching—Option A: Evaluation Commissioned at the Start of the Project
	Refinements to the Design	
	Box A11.1-5	Propensity Score Matching (PSM)
	Box A11.1-6	Example of Design 2.2B: Pretest–Posttest Design Commissioned Posttest and Using Statistical Matching: Evaluating the Effects of a Scholarship Program in Cambodia in Increasing Girls' Secondary School Enrollment
2.3	Regression Discontinuity (RD) (Design 2.3)	
	When to Use	
	Description of the Design	
	Figure A11.1-1	Example of Regression Discontinuity (Design 2.3): Assessing Impact of School Construction and Complementary Interventions on School Enrollment Rates Where Assignment Variable Is Average Household Income in School District
	Box A11.1-7	Example of Design 2.3: Regression Discontinuity Combined With a Mixed-Method Design: Evaluating the Effects of a Social Safety Net Based on a Poverty Index in Jamaica
	Table A11.1-7	Design 2.3: Regression Discontinuity (RD)
3.	**Weaker Statistical Designs**	
3.1	Pipeline Design (Design 2.4)	
	When to Use	
	The Evaluation Design	
	Box A11.1-8	Example of Design 2.4: Pipeline Design: A Quasi-Experimental Pipeline Design to Evaluate a Social Protection Program in Argentina
	Table A11.1-8	Design 2.4: Pipeline Design
3.2	Pretest–Posttest Project and Comparison Group Design With Judgmental Matching of the Two Samples (Design 2.5)	
	When to Use	
	The Evaluation Design	
	Table A11.1-9	Design 2.5: Quasi-Experimental Design With Judgmental Matching of the Two Samples
	Box A11.1-9	Case Study for Design 2.5A: Using Judgmental Matching to Evaluate the Impacts of Improved Housing on Household Income in El Salvador

3.3	Truncated Pretest–Posttest Project and Comparison Group Design (Design 3.1)	
	When to Use	
	Description of the Design	
	Table A11.1-10	Design 3: Truncated Pretest–Posttest Project and Comparison Group Design Starting at the Time of the Midterm Review
	Box A11.1-10	Case Study for Design 3 (Truncated Design) Combined With Design 2.5B (Judgmental Matching Pretest–Posttest): Assessing the Social and Economic Impacts of a Feeder Roads Project in Eritrea
3.4	Elimination of the Baseline Comparison Group (Pretest–Posttest Project Group Combined With Posttest-Only Analysis of Project and Comparison Group) (Design 4.1)	
	When to Use	
	Description of the Design	
	Incorporating Mixed-Method Approaches Into the Design	
	Table A11.1-11	Design 4: Elimination of the Baseline Comparison Group (Pretest–Posttest Project Group Combined With Posttest Analysis of Project and Comparison Groups)
	Box A11.1-11	Case Study for Design 4.1 (With Elements of Design 1): Comparing the Effects of Resettlement on Project Beneficiaries and Nonbeneficiaries in the Second Maharashtra Irrigation Project, India
3.5	Posttest-Only Comparison Group Design (Design 5)	
	When to Use	
	Description of the Design	
	Incorporating Mixed-Method Approaches Into the Design	
	Box A11.1-12	Case Study for Design 5: Assessing the Impacts of Microcredit on the Social and Economic Conditions of Women and Families in Bangladesh
	Table A11.1-12	Design 5: No Baseline Data (Posttest-Only Project and Comparison Groups)
	Box A11.1-13	Example of Design 5B: Posttest Comparison Combining Statistical Matching With a Mixed-Method Design—Evaluating Nicaragua's School-Based Reform: A Retrospective, Mixed-Method Design With Statistical Matching
4.	**Nonexperimental designs (NEDs) (Designs 6 and 7)**	
4.1	Pretest–Posttest No-Comparison-Group Designs (Design 6)	
	When to Use	
	Description of the Design	
	Table A11.1-13	Design 6: Basic Pretest–Posttest Project Group Design With No Comparison Group

(Continued)

(Continued)

	Box A11.1-14	Case Study for Design 6.4: Using a Before-and-After Survey of Resettled Households to Evaluate the Impact of the Khao Laem Hydroelectric Project in Thailand
	The Single-Case Design (SCD)	
	Box A11.1-15	Design 6.1: Single-Case Design: Treating a Schoolgirl With Asperger's Disorder
	Longitudinal NED Design (Design 6.2)	
	Box A11.1-16	Design 6.2: Longitudinal Design Without Comparison Group: The 12–18 Project—Making Lives Modern in Australia
	Interrupted Time Series (Design 6.3)	
	Box A11.1-17	Design 6.3: Interrupted Time Series: Estimating the Effects of Raising the Drinking Age
	Case Study Designs	
	Box A11.1-18	Example of a More Rigorous Nonexperimental Design: Evaluating the Effectiveness of the Natural Resources Leadership Program
4.2	Data Collected Only on the Posttest Project Group (Design 7)	
	When to Use	
	Description of the Design	
	Strengthening the Design	
	Incorporating Mixed-Method Approaches Into the Design	
	Table A11.1-14	Design 7: Posttest Analysis of Project Group Without a Baseline or Comparison Group
	Box A11.1-19	Case Study for Design 7: Assessing the Impacts of the Construction of Village Schools in Eritrea

TABLE A11.1-1 ● List of Experimental, Quasi-Experimental, and Nonexperimental Evaluation Design Options, With a Focus on How the Counterfactual Is Determined

	Start of Project	Intervention	Midterm	End of Project	Post-Project	Stage of Project When Evaluation Commissioned
Design	T_1		T_2	T_3	T_4	
Experimental (Randomized) Designs						
1.1[a] Longitudinal design starting with randomized selection of intervention and control group	P_1 C_1	X	P_2 C_2	P_3 C_3	P_4 C_4	Start

	Start of Project	Intervention	Midterm	End of Project	Post-Project	Stage of Project When Evaluation Commissioned
1.2 Randomized control trial with only pretest and posttest	P_1 C_1	X		P_3 C_3		Start
Quasi-Experimental Designs						
2.1 Longitudinal design (without randomized selection)	P_1 C_1	X	P_2 C_2	P_3 C_3	P_4 C_4	Start
2.2 Option A. Pretest–posttest comparison group design with statistical matching of samples. Evaluation commissioned at start of project.	P_1 C_1	X		P_2 C_2		Start
2.2. Option B. Pretest–posttest comparison group design with statistical matching of samples. Evaluation commissioned at end of project.[b]	(P_1) (C_1)	X		P_2 C_2		End
2.3 Regression discontinuity	P_1 C_1	X		P_2 C_2		Start
2.4. Pipeline comparison group design. Can be used when projects are implemented in phases. Individuals, households, or communities entering in Phase 2 (P2) and later phases can be used as the comparison group for those entering in Phase 1 (P1) and subsequent phases.	$P1_1$	X	$P1_2$ $C1_1 = P2_1$	$P1_3$ $C1_2 = P2_2$ $C2_1 = P3_1$	(. . . etc.)	Start
2.5 Option A. Pretest–posttest comparison group design with judgmental matching. Evaluation commissioned at start of project.	P_1 C_1	X		P_2 C_2		Start
2.5 Option B. Pretest–posttest comparison group design with judgmental matching. Evaluation commissioned at end of project. Recall or secondary data used to reconstruct initial status of both project and comparison groups.		(P_1) (C_1) X		P_2 C_2		End

(Continued)

(Continued)

	Start of Project	Intervention		Midterm	End of Project	Post-Project	Stage When Evaluation Commissioned
3.1 Pretest–posttest comparison group design in which initial data collection (delayed baseline) is not conducted until project has been under way for some time.			X	P_1 C_1	P_2 C_2		During implementation
4.1 Option A. Posttest comparison group design combined with collection of baseline data on project. Evaluation commissioned at start of project.		P_1	X		P_2 C_1		Start
4.1 Option B. Posttest comparison group design combined with collection of baseline data on project. Evaluation commissioned at end of project. Recall or secondary data used to reconstruct initial status of project group.		(P_1)	X		P_2 C_1		End
5.1 Posttest-only comparison group design.			X		P_1 C_1		End
Nonexperimental Designs							
6.1 Pretest–posttest single case project group design with no external comparison group. *Note:* Though it looks similar to the pipeline design, it is based on single cases, not group observations. This design is complicated to represent because the methodology requires that the treatment is applied sequentially in three separate cases, so more cells are required to represent all three phases.	C1 [($P1_1$)]		X_1	C1 [$P1_2$] C2 [$P2_1$] X_2	C2 [$P2_2$] C3 [$P3_1$] X_3	C3 [$P3_2$]	Start
6.2 Longitudinal design with no comparison group	P_1		X	P_2	P_3	P_4	Start

	Start of Project	Intervention		Midterm	End of Project	Post-Project	Stage When Evaluation Commissioned
6.3 Interrupted time series: This is a special case of Design 6.2, where more frequent observation points are available.		P_1 P_2 P_3 P_4 P_5	X	P_6 P_7 P_8	P_9 P_{10} P_{11}	P_{12} P_{13} P_{14}	Before start
6.4 Option A. Pretest–posttest project group design without comparison group. Evaluation commissioned at start of project.		P_1	X		P_2		Start
6.4 Option B. Pretest–posttest project group design without comparison group. Evaluation commissioned at end of project.		(P_1)	X		P_2		End
7.1 Posttest analysis of project group without a baseline or comparison group.			X		P_1		End

a. Initial number refers to design frameworks in Table A11.1-2; second digit refers to a variant of the basic design.

b. The parentheses—for example, (P_1) and (C_1)—indicate designs in which the evaluation was not commissioned until late in the project cycle, and baseline conditions of participants and comparison groups need to be estimated either through the use of secondary data from surveys conducted by other agencies or through the baseline reconstruction techniques discussed in Chapter 5.

1. LONGITUDINAL DESIGN WITH PRE-, MIDTERM, POST-, AND EX-POST OBSERVATIONS ON THE PROJECT AND COMPARISON GROUPS (DESIGN 1)

1.1. When to Use

This is the strongest RealWorld Evaluation (RWE) design framework. Although not all variants of this design involve randomly assigning units into experimental and control groups, random assignment is possible. Box A11.1-1 illustrates a longitudinal design that did use random assignment, and Box A11.1-2 gives an example where the project and comparison groups were statistically compared. Longitudinal designs require more observations than the typical pretest–posttest comparison group designs classified as Design 2. A longitudinal design framework requires collecting data on both the project and comparison groups during at least four different points in time. It also requires the evaluation to continue throughout and even after the life of a project. The first round of data is collected at the start of the project (baseline), and the final round of data is not collected until sometime after the project has ended (ex-post). The advantages of this design are that it provides the most comprehensive and methodologically robust assessment of project impacts and sustainability as well as getting "inside the black box" to describe the process of project implementation, as well as trends over time.

Although more expensive than other RWE designs, this design is recommended for projects that have an important research function, to thoroughly test an experimental strategy or approach, and where the evaluation will be used to guide

future decisions on the continuation, modification, or expansion of a major project. The design can be used only when the evaluation begins at the start of the project so that relevant baseline data can be collected. It also requires the evaluation to continue over a relatively stable environment in which the project is expected to continue to operate more or less as planned and where it will be possible to revisit the comparison groups over a number of years. It may be difficult, for example, to apply the design to a large-scale, low-income urban development project because it is likely that the housing of many of the comparison groups will be demolished or dramatically restructured over the life of the project (even though the comparison groups are expected to remain largely intact during the period of project implementation).

1.2. Description of the Longitudinal Design

As for all designs, the design should be based on a program theory model (see Chapter 10). Design 1 compares the project and a comparison group at the start of the project (T_1), at midterm or even more frequently during the life of the project (T_2), at project completion (T_3), and some time (preferably several years) after the project has ended (T_4) to assess sustainability (see Table A11.1-2). Among the advantages of having multiple observations made during the life of project implementation are being able to describe and assess the implementation process and to identify trends in outcome indicators over time.

A comparison or nonequivalent control group I is selected at the start of the project to approximate as closely as possible the project beneficiary group (P). The observations from the comparison group represent the counterfactual, or what the conditions of the project group would have been like if the project's interventions had not taken place. In other words, if the average household income of the comparison group increases between T_1 and T_3, it needs to be known whether there would have been a similar increase in the project group if the project intervention had not taken place.

Our use of the term *comparison* group, rather than *control* group, reminds us that it is rarely possible to assign subjects randomly to the project and control groups, so differences may exist between the characteristics of the two groups that could distort the interpretation of the findings. It also acknowledges that in the real world, communities not participating in our project cannot be "controlled" in terms of external factors or alternative developmental activities undertaken in those communities by other agencies.

Both groups are interviewed at Time 1 (T_1) before the project begins, and information is obtained on a set of indicators ($I_1, I_2, \ldots I_n$) measuring the variables that the project is intended to affect (e.g., household income, daily travel time, number of children attending school).

Information is also collected on the social and economic characteristics of the individuals or families, called intervening variables, that might affect project outcomes. Data collection is repeated at Time 3 (T_3) at the completion of project implementation (e.g., occupation of new houses, turning on the water supply, completion of the rural roads) and again a relevant amount of time after the project has been completed, to assess the sustainability of any impacts (T_4). Ideally, the analysis should also include the contextual factors and how they may have changed over the life of the project.

If the observations P_1, P_3, C_1, C_3, and so on refer to the mean scores for each group (e.g., average income, average educational test scores, or average anthropometric scores) before and after, as well as with and without, project implementation, then a statistical test such as the t-test for the difference of means is used to determine whether the observed difference is statistically significant. If, on the other hand, the values refer to proportions (e.g., the proportion of children attending school or the proportion correctly answering a test question), then the appropriate statistical test would be a measure such as the Z-score for difference of proportions (Moore & McCabe, 1999, Chapter 8; see also Chapter 12 of the present book).

The advantages of this design are that it examines the processes and outcomes of project implementation over time, and it addresses the important issues of sustainability, which are ignored in many evaluations. This is the evaluation design that ideally would be most compatible with the program theory model presented in Chapter 2 and discussed in more detail in Chapter 10. It can serve both formative and summative evaluation purposes. Design 1 is a very robust design that can address many of the "threats to validity" discussed in Chapter 7. However, it also requires sizable resources in terms of budget and personnel because the survey questionnaire, or similar data-collection instrument, is administered at four or more points in time during (and after) the life of a project, within both the beneficiary population and a comparison community.

TABLE A11.1-2 ● Design 1: Longitudinal Design

	Time	T_1 (baseline)	Project Intervention (over time)	T_2 (midterm)	T_3 (end of project)	T_4 (after project)
	Sample selection procedure					
Project group Comparison group	While randomized selection is occasionally used, in most cases, participants will either be self-selected or be selected by the project agency. Quasi-experimental matching procedures will be used for the comparison group: either using statistical matching (the strongest option) or judgmental matching.	P_1 C_1	X	$P_{2(n)}$ $C_{2(n)}$	P_3 C_3	P_4 C_4

T_1 = pre-project observation.

T_2 = observation(s) during project implementation period.

T_3 = project completion (when project funding ends).

T_4 = ex-post follow-up evaluation sometime after project completion to determine sustainability.

P_1 and C_1 = baseline observations on the project and comparison groups before start of the project.

$P_{2(n)}$ and $C_{2(n)}$ = periodic (longitudinal) observations on both groups during project implementation to observe the processes of change; (n) indicates the number of observation points during implementation. Some evaluation designs use panel informants (a small subsample of the participants and perhaps comparison groups) to track trends over time. If only one observation is made at the midterm review, this would be simplified to P_2 and C_2.

X = implementation of the project (construction of schools, water system, etc.). It is important to remember that project implementation is a process that continues over time and is not a finite event occurring at a specific point in time.

As we will discuss in the RCT section below, it is rarely feasible to randomly allocate subjects to the project or control groups in most real-world social development programs. In these cases, multivariate analysis can be used to statistically control for differences between the project and comparison groups. Many refinements can be introduced into the basic design to assess multiple treatments or to capture impacts that evolve gradually over time (Shadish, Cook, & Campbell, 2002; Valadez & Bamberger, 1994).

The GAIN welfare-to-work program in California (see Box A11.1-1) is an example of a longitudinal design that used random assignment to determine who would receive assistance with job search, basic education, vocational training, and unpaid work experience.

The Bolivian Social Investment Fund evaluation (Box A11.1-2) is an example of a longitudinal quasi-experimental design that used statistical matching to evaluate two of the three components (health and water/sanitation) and randomized assignment for the third component (education).

BOX A11.1-1

An Example of Design 1.1 (Longitudinal Design With Randomized Assignment): The Evaluation of GAIN: A Welfare-to-Work Program in California

GAIN was created in 1985 by the California legislature as California's main welfare-to-work program for recipients of Aid to Families with Dependent Children (AFDC). The program was designed to provide welfare recipients with job search assistance, basic education, vocational training, postsecondary education, and unpaid work experience to help them prepare for and find employment. Welfare recipients assigned to the program were required to participate in these activities. Those refusing to cooperate without "good cause" could have their welfare payments reduced as a penalty. The evaluation addressed four main goals: (1) to learn about the counties' experiences in turning this ambitious legislation into an operating program and welfare recipients' participation and experiences in it; (2) to determine GAIN's effects or impacts on recipients' employment, earnings, welfare, and other outcomes and whether positive effects could be achieved in a variety of settings; (3) to assess how different program strategies influence those results; and (4) to determine the program's economic costs and benefits. Implementation issues were studied in 13 counties, and the full evaluation was concentrated in a subset of six of these counties, representing diverse areas of the state.

The evaluation made use of an array of QUANT and QUAL data, including employment and welfare administration records, program case data, staff and recipient surveys, field research, and fiscal data from a wide range of agencies. For the impact evaluation, more than 33,000 welfare recipients were randomly assigned to GAIN or a control group. The control group did not have access to GAIN services and was not subject to a participation mandate as a condition of receiving their full welfare grants. However, they remained free to enroll themselves in any services normally available in the community. Program impacts were analyzed using data from existing administrative records on employment, earning, welfare receipt, and food stamp payments (for all sample members) and the recipient survey (for a subsample). A cost–benefit analysis was used to estimate the overall financial gain or loss caused by the program for welfare recipients, government budgets and taxpayers, and society as a whole. The GAIN participant and control samples were interviewed regularly over a five-year period.

Source: Fitzpatrick, Christie, & Mark (2009).

BOX A11.1-2

A Longitudinal Evaluation Combining Design 2.1 (Randomized Design) for One Component and Design 2.2 (Quasi-experimental Design) for the Other Two Components: The Bolivian Social Investment Fund

The Bolivian Social Investment Fund (SIF) was established in 1991 as a financial institution promoting sustainable investments in the social sectors, notably health, education, and sanitation. The program directs investments to poor communities that have been largely neglected by public service networks. The SIF was the first institution of its kind and has served as a prototype for programs in Latin America, Africa, Asia, and the Middle East.

The evaluation began in 1991 at the start of the project and continued over a period of 10 years. A baseline study was conducted in 1991, a second study was conducted in 1993 at the start of the second phase, and a follow-up study was conducted in 1997. A number of other studies were conducted up to 2000. It also included separate studies of the education, health, and water projects. The evaluation uses a wide range of techniques to assess the effectiveness of the targeting systems in reaching the poor: the impact of the social service investments on desired social outcomes such as improved school enrollment rates, health conditions and water availability, and the overall cost-effectiveness of the SIF as a mechanism for delivering social services to low-income communities.

For the health and water supply components, interested communities applied to participate in the project. These communities were then statistically matched with comparison communities using socio-data from project records and other sources.

The evaluation design used random assignment for the education component with eligible communities being randomly assigned to the treatment and control groups, while for the health and water components, communities that elected to participate were statistically matched with comparison communities. Three subsamples were used: (a) a random sample of all households in rural Bolivia (plus the Chaco region), (b) a sample of households that live near the schools in the project and comparison areas (for the evaluation of the education component), and (c) a sample of households that would benefit from the water and sanitation component. The evaluation combined various data-collection techniques at the level of the community and the household, and data were collected over a 10-year period.

The comparison of findings from randomized allocation designs and statistically matched designs (in education) found that randomization produced better-matched samples and detected some positive outcomes, such as improvements in school infrastructure in the project areas, that were not detected by the judgmentally matched samples.

Source: Baker (2000).

1.3. Tools and Techniques for Strengthening All of the Basic Impact Evaluation Designs Described in This Appendix

All of the designs presented in Table A11.1-1 and described in this appendix depict only the frameworks or scenarios under which these evaluation designs operate. The methodological validity and practical utility of all of these designs can be enhanced by incorporating one or more of the refinements described in Table A11.1-3. These refinements are as follows:

- Basing the design on a program theory model that incorporates a theory of change and a theory of action (see Chapter 10)
- Incorporating process analysis to understand how the project is actually implemented and to assess how any deviations from the intended implementation plan affect efficiency, accessibility, and outcomes (see Chapter 10)
- Incorporating contextual analysis to understand the interaction between the project and the economic, political, institutional, and sociocultural context within which it operates
- Using mixed-method approaches that combine quantitative and qualitative design, data collection, and analysis methods so as to combine breadth (the ability to generalize from the sample to the total population) with depth of understanding of the lived experience of individual families, groups, or communities (see Chapter 14)
- Maximizing the use of all available types of secondary data
- Using triangulation to combine data and interpretation of findings from two or more independent sources so as to increase reliability and validity of the findings

Many of these techniques can also be combined to help reconstruct baseline data when the evaluation is not commissioned until late in the project cycle. (See Chapter 5 for a full discussion of strategies for reconstructing baseline data.)

Many of the examples of RealWorld Evaluation designs presented in this appendix illustrate how evaluators incorporate these different techniques to strengthen the basic designs. For example, the evaluation of the GAIN welfare-to-work program (Box A11.1-1) strengthened the longitudinal design with randomized assignment by incorporating a range of quantitative and qualitative methods to understand how different groups responded to the different components of the program and to understand how the program actually operated in different contexts.

All of these techniques can be used singly or in combination to strengthen all of the designs described in this appendix, but for reasons of space and to avoid repetition, the importance of incorporating these techniques will not be repeated.

TABLE A11.1-3 ⬢ Tools and Techniques to Strengthen Any of the Basic Impact Evaluation Designs

The tools and techniques described in this table can enhance the methodological rigor of all of these designs, including the designs that are considered to be statistically rigorous.

Essential Tools and Techniques	Why Required	How to Implement
1. Basing the evaluation on a theory of change and a program logic model	The purpose of an evaluation is not just to estimate *how much* change has occurred but also to explain *why* and *how* the changes were produced. Clients also wish to know to what extent the changes were due to the intervention and whether similar changes would be likely to occur if the program is replicated in other contexts. To achieve the above objectives, it is necessary to explain the underlying theory and the key assumptions on which the program is based and to identify how these can be tested in the evaluation.	The design and use of program theory is discussed in Chapter 10. That chapter also illustrates how the theory can be articulated graphically through a logic model.
2. Process analysis	Project outcomes are affected by how well a project is implemented and by what happens during implementation. Without process analysis, it is not possible to assess whether failure to achieve outcomes is due to design failure or to implementation failure.	See Chapters 10, 11, and 17
3. Contextual analysis	Projects implemented in an identical way in different locations will often have different outcomes due to different local economic, political, or organizational contexts or different socioeconomic characteristics of target communities. This can result in wrong estimations of project impact, often leading to underestimation of impacts (due to increased variance of the estimations).	See Chapters 10 and 11
4. Using mixed-method approaches to strengthen evaluation design, data collection, and analysis	Most evaluation designs can be strengthened by combining QUANT techniques that ensure the statistical representativity of the data with QUAL methods that permit in-depth analysis and assessment of the quality of implementation, outputs, and outcomes.	See Chapters 12, 13, and 14
5. Identification and use of available secondary data	Many evaluations do not identify and use all of the available secondary data. Secondary data can often reduce the costs of primary data collection and provide independent estimates of key variables.	See Chapter 5
6. Triangulation	The validity of the data and the quality and depth of the interpretation of the findings are enhanced when two or more independent estimates can be compared.	See Chapters 13 and 14
7. Reconstructing baseline data	When evaluations are not commissioned until late in the project, it will frequently be the case that no baseline data had been collected. Several of the techniques described above (mixed methods, secondary data, and program theory models) can be combined to "reconstruct" and estimate the baseline conditions of the project and comparison groups.	See Chapter 5

1.4. Threats to Validity and Adequacy of the Conclusions

Design Framework 1 is the methodologically strongest RWE scenario (data collected at multiple times during and after the life of a project, with relevant counterfactual). Thus, if it is applied appropriately, most threats to validity and adequacy (see Chapter 7) should be less serious than for most of the subsequent design frameworks, especially if the design is complemented with appropriate QUAL methods. However, under real-world conditions, it is never possible to design a perfect evaluation, so it is always important to review the Integrated Checklist (Appendix 7.3) to identify potential problems. Table A11.1-4 identifies from this checklist some of the potentially important threats to conclusion validity affecting this design to which particular attention should be paid. Most of these threats to validity also affect the subsequent designs, but again, for reasons of space and to avoid repetition, we will not address threats to validity for each individual design.

These potential threats should be examined during the design phase and at later points during the evaluation. Where possible, corrective measures (see Chapter 7) should be taken. If this is not possible, the evaluation report must clearly identify the existence of these threats and how they affect the findings and conclusions.

TABLE A11.1-4 ⬢ Some Threats to Validity That Must Be Checked for Designs 1 and 2

Threats to statistical conclusion validity	
1. Low statistical power	The sample is too small to be able to detect statistically significant effects (see Chapter 12).
2. Unreliability of measures	The indicators do not adequately measure key variables.
3. Restriction of range	The sample does not cover the whole population. For example, the lowest or highest income groups are excluded, or the sample covers only enterprises employing more than 10 people.
4. 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, but others received only the tablets.
Threats to internal validity	
1. Selection bias	This refers to differences between project and comparison groups with respect to factors affecting outcomes.
2. 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 drop-out rates, so the people who are actually exposed to the project are no longer representative of the whole population.
3. 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.
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.

(Continued)

(Continued)

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 Checklists 1 to 4 in Appendix 7.1 for the full list of threats. The numbers in the left column correspond to those given in the checklists.

2. STRONG STATISTICAL DESIGNS

2.1 Randomized Control Trial (Design 2.1)

2.1.1 When to Use

RCTs are a powerful statistical design because, when the sample is sufficiently large to estimate the significance of changes given a particular effect size, they can eliminate selection bias (initial differences between the project and control groups that might explain part of the change that was assumed due to the project intervention). Thus, this design can be useful when it is important to obtain credible estimates of the impact of a major social or economic development project or program. The designs work best when there is a single intervention that can be administered in a standard way and where the effects can be quantified (e.g., increase in income, increase in school enrollment rates, reduction in infant mortality). The designs also work better for large projects reaching large numbers of people and being implemented in many different locations. In this way, the effects of particular local contextual factors can be cancelled out and the average effect size estimated.

A number of factors, however, limit scenarios in which this design can be used. As indicated earlier (see Chapters 11 and 12), RCTs can only be used when subjects can be randomly assigned to treatment and nontreatment groups. In many cases, projects use participant self-selection, or participants are selected by the agency responsible for implementing the project. In other cases, there are political or ethical objections to the use of this model. The RCT design also only works well in a stable project environment where participant selection and service delivery can be implemented according to the project guidelines and in a stable way. The project environment must also remain relatively stable. A final limitation is that the basic RCT only estimates average project effect (it averages out contextual factors or participant attributes that can affect implementation and outcomes). So while this is useful for estimating the average effect of a large project, it makes it more difficult to use the findings to assess potential replicability in other locations with different population groups and different contextual factors.

The promotion of RCTs for the evaluation of international development programs, as well as within the United States, has sparked many heated debates among evaluators, donors, and other stakeholders. (See Chapter 19 for a summary of the limitations of RCTs.)

2.1.2 Description of the Design

Table A11.1-5 illustrates a typical RCT design. The design can be represented as follows:

Box A11.1-3 illustrates how a randomized control trial was used to assess the efficacy of a vaginal microbicide gel in reducing the likelihood that a woman would become infected with HIV after sex.

TABLE A11.1-5 ● Experimental Design 2.1: Randomized Control Trial (RCT)

	Time	T_1		T_2
	Sample Selection Procedure		**Project Intervention[a]**	
Project group	Subjects randomly assigned to project and control groups[b]	P_1	X	P_2
Control group		C_1		C_2

a. With RCT, the treatment is often administered at a specific point in time (e.g., a drug), but it can also continue to be administered over time (e.g., school meals, the use of a flip chart throughout the school year).

b. Random assignment will often be done in two or more stages, with one reason being to increase both external and internal validity (Khandker, Koolwal, & Samad, 2010, Chapter 9).

BOX A11.1-3

Example of Design 2.1: Randomized Control Trial Testing a Vaginal Microbicide Gel to Reduce the Likelihood of HIV Infection

The Centre for the AIDS Program of Research in South Africa (CAPRISA) announced breakthrough research in 2010 on testing a vaginal microbicide gel containing an antiretroviral drug known as Tenofovir. Randomized control trials found that women who used the drug before and after sex were 39% less likely to become infected with HIV compared to the placebo gel users. The risk fell by 54% for women who used the drug consistently, and the risk of acquiring herpes simplex virus fell by more than half. The research, which is largely funded by USAID under the President's Emergency Plan for AIDS Relief (PEPFAR), is considered to be the first "proof of concept" of the efficacy of a vaginal gel that women can apply themselves without having to face the problems of negotiating the use of condoms with their sex partner. Susan Brems, the senior deputy assistant administrator in the Bureau for Global Health, stated, "This is the first ever 'proof of concept' (meaning that it works) that a vaginal microbicide can reduce the risk of sexually transmitted HIV. We all know that science consists of trial and error, and previous studies of microbicide candidates had not proved promising. While further testing is necessary, we now have renewed hope for a microbicide as we move forward."

Source: Claypool (2010).

2.2. Pretest–Posttest Project and Comparison Group Designs With Statistical Matching (Designs 2.2–2.5)

2.2.1 When to Use

These designs are used under similar conditions to Design 1. However, they are used more frequently because they cover a shorter time period (ending around the same time as the project) and are also relatively less expensive than the longitudinal

design (Design 1) (although still more expensive than the less robust quasi-experimental designs [QED]) because data are collected at only two points in time rather than four or more. There are at least three variants of this design. If there is random selection of participants and a control group, it is a randomized control trial (Design 2.1, described above). In one variant (Design 2.2), the size of the sample and the quality of data make it possible to statistically match the two samples (see below), thus reducing selection bias. In another variant (Design 2.3), the two samples are matched using judgment (because the available statistical base is weaker) so that the potential selection bias is larger. In both cases, the design can only be used when it is possible to select a reasonably well-matched comparison group and to collect baseline data on both groups. This design is still fairly robust and is often the design of choice when the evaluation starts at the same time as the project and when a fairly rigorous project impact evaluation is justified (and, of course, when the resources are available). Design 2.3 (the judgmental matching of treatment and comparison groups) is a widely used and useful design, but greater care is needed to address the potential selection bias and to strengthen the design using the RWE techniques discussed earlier.

TABLE A11.1-6 ⬢ Design 2.2: Pretest–Posttest (Quasi-Experimental) Design With Statistical Matching of the Two Samples

Time		T_1 (baseline)	Project Intervention	T_3 (end of project)
	Sample selection procedure			
Project group	The project group is selected through either self-selection or administrative selection by the project agency.	P_1	X	P_2
Comparison group	Where secondary data are available from previously conducted surveys, it is possible to use techniques such as propensity score matching (PSM) (see Box A11.1-5), or instrumental variables can be used to improve the matching. Where a large sample survey is possible, techniques such as PSM can also be used. Also, large samples permit the use of multiple regression to statistically control for the effect of subject attributes such as income, education, farm size, and so on.	C_1		C_2

2.2.2 Description of the Design

Design 2.2 represents a simplified version of Design 1. It involves comparison of the project and nonequivalent comparison groups at the start of the project (T_1) and again at project completion (T_3). As always, it is recommended that the design

should be based on a program theory model and should be combined with a process evaluation to analyze the project implementation and a contextual analysis to assess the influence of the economic, political, organizational, and natural environment within which the different project sites operate and to study the influence of cultural characteristics of the affected populations on program outcomes.

Box A11.1-4 presents an example of how this design was used to evaluate a conditional cash transfer program in Colombia (*Familias en Accion*).

BOX A11.1-4

Example of Design 2.2A: Pretest–Posttest Comparison With Statistical Matching—Option A: Evaluation Commissioned at the Start of the Project

Familias en Accion (FeA) is a conditional cash transfer (CCT) program launched in Colombia in 2001 and funded by the World Bank and the Inter-American Development Bank (IDB). It promoted increased access to health and education by providing monthly grants to poor families on the condition that children were brought to the local clinic for regular health check-ups and vaccinations and that children attended school regularly. All payments were made to the mother on the assumption that the money was more likely to benefit children. The program operated in municipalities with populations of less than 100,000 and required a bank branch to which funds could be transferred. Beneficiaries were selected from the lowest stratum of the social security register (Sisben).

A pretest–posttest comparison group design was used with the comparator groups selected from municipalities ineligible to participate in the program, in most cases because there was no bank branch to handle the funds transfer. The availability of good secondary data permitted the use of propensity score matching to reduce sample selection bias. The baseline studies were conducted in 2002 with follow-ups in 2003 and 2006. A total of 57 project and 65 control municipalities were sampled with approximately 100 interviews per municipality. Political pressures due to the upcoming elections forced FeA to advance the program launch, and families in a number of municipalities had already received payments before the baseline study was conducted. The World Bank and IDB were able to convince the government to delay program launch in some areas until the baseline studies could be conducted. Consequently, the baseline was divided into two groups: those who had not received any payments prior to the baseline and those who had.

Positive results could already be seen at the time of the first follow-up study one year after project launch (2003), particularly in rural areas. The results of the second 2006 follow-up were similar to the 2003 study: There was increased primary school enrollment in rural but not urban areas, increased secondary school enrollment in rural and urban areas, some improvements in nutritional status in rural areas but not in urban areas, and an impact on diarrhea occurrence for younger rural children but not for children older than 36 months. A major concern was the lack of effects on anemia, which affects half of all poor children. Reservations were expressed in the report and in conversations with policymakers concerning the extent to which findings from the small municipalities could be extrapolated to urban areas.

Source: Bamberger & Kirk (2009).

2.2.3 Refinements to the Design

The design presented in this section is in fact the most basic design framework for pretest–posttest comparison group evaluations. Although cost and time constraints mean that this design is more likely to be used than the longitudinal Design 1 in RWE applications, a number of refinements can significantly strengthen the design (Shadish et al., 2002, Chapter 5). Several of the refinements increase the number of pretest or posttest observation points, whereas others reverse the treatment between the project and comparison groups at different points. It is also possible to use cohort analysis in which successive groups passing through the same cycle (e.g., third-grade students, medical trainees, women receiving microcredit loans, families who will receive houses or public services in different phases of a project) are compared. One

variant of this is to use subjects who will receive services or benefits in the next phase of a project as a comparison group for subjects who received the services in the first phase (see description of pipeline Design 2.4).

Box A11.1-6 illustrates how Design 2.2B was used to evaluate the effectiveness of a scholarship program in increasing girls' secondary school enrollment in Cambodia.

BOX A11.1-5

Propensity Score Matching (PSM)

PSM uses logistical regression to strengthen the match of the project and comparison group samples. The technique is most commonly used when a large sample survey is available that covers the geographical areas of interest to the evaluation and the appropriate information has been collected on the appropriate population groups. The technique is used both to reconstruct baseline data or for ex-post surveys. PSM can also be used when a sufficiently large primary sample survey is being conducted, but this application is less common. The technique involves the following steps:

a. Logistical regression (Logit) analysis is conducted with a combined sample of project participants and nonparticipants from areas that are considered appropriate as a comparison group. The first stage of the analysis, covering the project group, determines subject characteristics that are good predictors of participation.

b. For each subject in the project sample, around five "nearest neighbors" are selected from the comparison population who match the subject closely on all of these characteristics.

c. Baseline scores on the indicators to be used to measure changes in outcome variables are recorded for each participant, and the average score for the nearest neighbors of each subject are also computed.

d. The posttest scores are then calculated for each participant and the nearest neighbors, and a "change score" is calculated as the difference in the change for each participant compared to the average change for the nearest neighbors.

e. All of the individual change scores are summed to estimate the total change score.

The *Familias en Accion* evaluation discussed under Design 2.2A (Box A11.1-4) is an example of the use of PSM.

Source: Baker (2000, pp. 48–52) and Khandker et al. (2010, Chapter 4).

BOX A11.1-6

Example of Design 2.2B: Pretest–Posttest Design Commissioned Posttest and Using Statistical Matching: Evaluating the Effects of a Scholarship Program in Cambodia in Increasing Girls' Secondary School Enrollment

The Japanese Fund for Poverty Reduction (JFPR) in Cambodia awarded scholarships to poor girls who were completing sixth grade and who wished to enter secondary school. The goal of the program was to encourage girls from poor families to enroll in secondary school and to complete the full three years of lower secondary school. The program covered 15% of all secondary schools, and in each, a maximum of 45 girls were awarded scholarships.

As the evaluation was not commissioned until late in the project, a retrospective evaluation design was used. Two sources of data were used: application forms for the scholarship program (information on parental education, household composition, ownership of assets, housing materials, and distance to the nearest secondary school) and data on school enrollment and attendance collected during unannounced school visits. Regression analysis was used to statistically control for socioeconomic

characteristics of the households after the data had been collected on the project and comparison groups (i.e., it was not possible to use propensity score matching to match samples before data collection).

The evaluation found that scholarship recipients had significantly lower socioeconomic status than nonrecipients, confirming that the program had been successful in targeting poorer girls. After controlling for household characteristics, it was found that girls receiving scholarships had an almost 30% higher enrollment and attendance rate than nonrecipients and that the effects were greatest for the most disadvantaged girls.

In the evaluation of a follow-up World Bank project, the Ministry of Education had become aware of the benefit of a well-designed evaluation, and the evaluation was built into the project design. It was originally proposed to use a randomized control trial, but the Ministry was concerned that this would not permit targeting of the poorest families, so it was agreed to use a regression discontinuity design.

Source: Adapted from Bamberger & Kirk (2009).

2.3. Regression Discontinuity (RD) (Design 2.3)

2.3.1 When to Use

RD is an evaluation technique that was used widely in the 1960s and 1970s in fields such as criminal justice (e.g., to assess the impacts of different treatments on recidivism rates). Its use then appeared to decline, but there has been renewed interest in recent years. RD can be used when the population of interest can be ranked on an ordinal or interval scale (assignment variable) that is used as the eligibility criteria for selection to participate in the project. A numerical eligibility cut-off point is defined on the scale; people above this cut-off are always accepted, and people below the cut-off are always rejected. Examples of assignment variable scales are number of hours of vocational training that prisoners received while in jail, a rating on a scale of psychological disorder or stress, or a rating on a scale of likelihood of success in a women's entrepreneurial management training program.

There are three very attractive features with RD: one methodological, one political and ethical, and one logistical. The methodological advantage is that when properly designed and administered, an RD design can produce unbiased estimates of project impact. The political and ethical advantage is that experts or program management can define the eligibility criteria for project participation. For example, an agricultural development project could be targeted for the smallest and poorest farmers (defined, for example, in terms of hectares of land owned), a school feeding program could be targeted for children from the poorest households (using an appropriate indicator of wealth, expenditure, or consumption), and a women's entrepreneurial development program could be targeted either to women considered most likely to succeed (based on an assignment variable scale rated by experts or managers) or to the poorest and most needy women. In all cases, eligibility criteria are determined by clients and stakeholders, thus avoiding political concerns that managers and influential stakeholders have no control over who receives benefits and the ethical concerns that benefits are withheld from the neediest groups while other less needy people might participate based on the luck of the lottery or other randomization process.

The logistical advantage is that the comparison group is generated automatically through the selection process (all falling below the cut-off eligibility criterion). In many cases, it is a major challenge to identify a comparison group, so having the group already selected can be a significant advantage.[15]

2.3.2 Description of the Design

The design requires the definition of a target population (e.g., prisoners being released from jail, high school students in low-income areas, small farmers). An assignment variable must then be identified. Normally this will be related either to need (poverty, low school test scores, low agricultural productivity) or to likelihood of success (hours of vocational training

[15] Some programs also provide a small benefit to the comparison group so as to give them an incentive to cooperate with the study. For example, an entrepreneurial training program invited members of the comparison group to attend a workshop on how to obtain information resources on the Internet. This benefit was considered a sufficient incentive to encourage people to cooperate with the study but not sufficient to affect their entrepreneurial activities.

received, rating on a likelihood of entrepreneurial success scale). The scale must be ordinal or interval with precise and measurable scale positions, and it must be possible to rate everyone on the scale. A precise and measurable eligibility cut-off must also be defined, and it must be clear who falls above and below the cut-off.

A strict selection procedure must be applied so that everyone above the cut-off point is accepted and everyone below the cut-off is rejected. In practice, the strict application of the selection rule has proved to be a challenge. Sometimes the data are not available to apply the rating (hectares of land are not known, it is difficult to determine household wealth or income), sometimes there are pressures to relax the criteria to accept friends or people with political contacts, and in other cases the administrative procedures are not well monitored.

Once selection has been completed and the program is implemented, the evaluation involves comparing subjects just above the eligibility cut-off point with those just below it. A regression is calculated between the score on the assignment variable and the posttest outcome score. If the project had an effect, there will be a discontinuity ("jump") in the regression line at the cut-off point. This is illustrated in Figure A11.1-1 to assess the effect of constructing new schools and complementary interventions on school enrollment rates where the assignment variable is average household income in each school district and the cut-off point is average household income of $100 per month. In this case, the unit of analysis is the school district (not the family with school-age children). It can be seen that there appears to be a significant discontinuity around the $100 cut-off point. Of course, the significance of this break has to be calculated statistically.

One practical issue with the RD design is that it requires a relatively large target population because the sample for the evaluation only compares subjects falling just above or just below the cut-off point. As a rule of thumb, the 25% above the cut-off point are compared with the 25% just below it. So if a sample of 100 was required to obtain statistically significant results, it would be necessary to start with a target population of 200.

Box A11.1-7 illustrates the use of a regression discontinuity design to evaluate the effects of a social safety net program in Jamaica.

BOX A11.1-7

Example of Design 2.3: Regression Discontinuity Combined With a Mixed-Method Design: Evaluating the Effects of a Social Safety Net Based on a Poverty Index in Jamaica

In 2001, the government of Jamaica initiated the Program of Advancement through Health and Education (PATH) to increase investments in human capital and improve the targeting of welfare benefits to the poor. The program provided health and education grants to children in eligible poor households, conditional on school attendance and regular health care visits. Eligibility for the program was determined by a scoring formula with a clearly defined cut-off. A regression discontinuity (RD) analysis was conducted comparing families 2 to 15 points below and above the cut-off. Researchers justified the use of the RD design because a baseline study showed that the treatment and comparison groups had similar levels of poverty and also had similar levels of motivation as both groups had applied to participate in the program. The RD was strengthened by using a mixed-method design that gathered data using information systems, interviews, focus groups, and household surveys.

The analysis found that the PATH program increased school attendance for children ages 6 to 17 by an average of 0.5 days per month, which is significant given an already high attendance rate of 85%. Also, health care visits by children ages 0 to 6 increased by approximately 38%.

Source: Adapted from Gertler, Martinez, Premand, Rawlings, & Vermeersch (2011).

FIGURE A11.1-1 ⬢ Example of Regression Discontinuity (Design 2.3): Assessing Impact of School Construction and Complementary Interventions on School Enrollment Rates Where Assignment Variable Is Average Household Income in School District

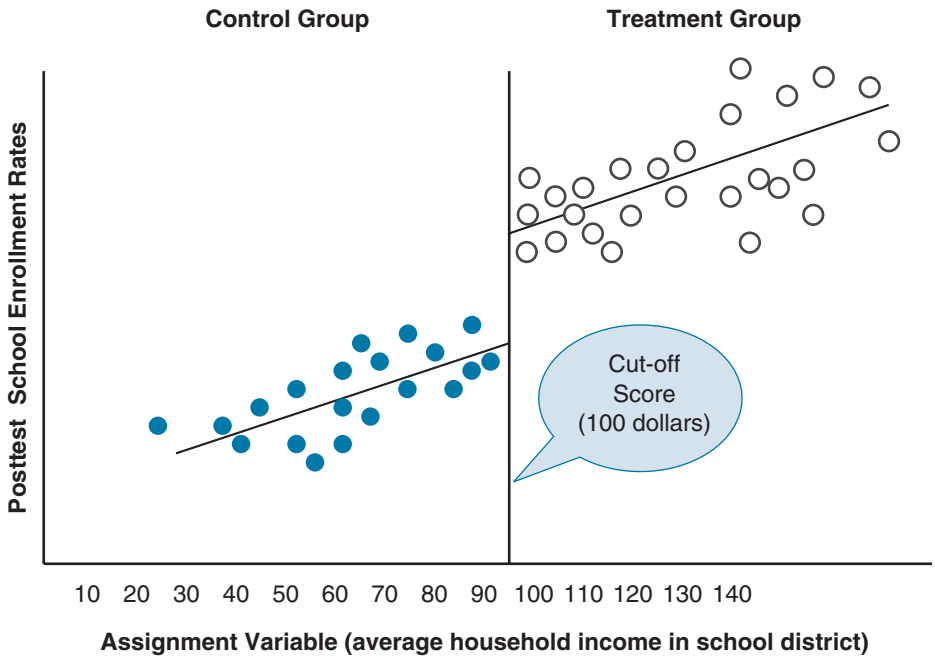


TABLE A11.1-7 ⬢ Design 2.3: Regression Discontinuity (RD)

	Sample Selection Procedure	T_1 Start of Project	T_2 Implementation	T_3 End of Project
Project group	The 25% of subjects above and closest to the eligibility criteria cut-off point	P_1	X	P_2
Comparison group	The 25% of subjects below and closest to the eligibility cut-off point	C_1	Sometimes the group is given a small service as an incentive to cooperate with the study	C_2

3. WEAKER STATISTICAL DESIGNS

3.1. Pipeline Design (Design 2.4)

3.1.1 When to Use

This design is used when a project is implemented in phases and where communities or people who will not benefit until the second or a subsequent phase can be used as the comparison group for Phase 1. The design works best where the communities or individuals in each phase are similar, where the implementation schedule for each phase is known in advance, where all phases will eventually receive similar benefits, and where communities or individuals in Phase 2 and subsequent phases do not have access to benefits during Phase 1. In the real world, many projects do not satisfy all these conditions. For example, in a slum upgrading project, families in Phase 2 may walk several kilometers to collect water from the drinking water standpipes being installed for Phase 1. Also, it is quite common that the characteristics of communities in each phase are not identical.

3.1.2 The Evaluation Design

As mentioned above, it is first necessary to determine that the project will be implemented in clearly defined phases with an implementation schedule that is known in advance. It is also necessary to check that the characteristics of individuals or communities in each phase are similar and that people or communities in Phase 2 and subsequent phases will not have access to services during Phase 1. Once the project begins, monitoring studies should be conducted periodically to ensure that these conditions are satisfied in practice.

Assuming these conditions are met, a sample of households or communities that will participate in Phase 1 is randomly selected. A decision will then be made as to how the comparison group sample will be selected. If a project is implemented in several phases, it would be possible to only select the comparison group from Phase 2 (geographically closest to Phase 1 and probably more similar than communities farther away) or to include households or communities from all subsequent phases or even to only select from the final phase. While Phase 2 families may be more similar to Phase 1, there are two potential disadvantages of using them as the comparison group. First, they can serve only as a comparison group until the time they begin to receive project benefits in Phase 2. Consequently, they can only be used to assess initial changes in Phase 1 up to the point when Phase 2 begins. Because in many cases, each phase is completed in one year or less, Phase 2 cannot be used to assess changes that take place more slowly. The second issue is that given its greater proximity, Phase 2 subjects are more likely to start having access to some project benefits before Phase 2 officially begins. The advantage of selecting the comparison group from Phase 3 or later is that they can continue to be used as a comparison over several years. Also, given their greater distance, they are less likely to gain access to the benefits of Phase 1. However, their greater distance might increase the likelihood that the group will have different characteristics from the participants in Phase 1. There is also the disadvantage of interviewing people from later phases because they may lose patience while waiting for the project's interventions to reach them. Box A11.1-8 illustrates how a pipeline design was used to evaluate a social protection program in Argentina, and Table A11.1-8 describes the logic of the design.

For projects that are not geographically based but that provide cash or other services to people who may be distributed over a wider area, the considerations in selecting the comparison group may be slightly different. For example, the "leakage" of benefits such as access to water might not be a consideration when families are receiving cash payments.

BOX A11.1-8

Example of Design 2.4: Pipeline Design: A Quasi-Experimental Pipeline Design to Evaluate a Social Protection Program in Argentina

A large-scale social protection program, Jefes y Jefas (translation: male and female household heads), was launched in Argentina in 2001 in response to the financial crisis. The program was a public safety net that provided income to families with dependents for whom their main source of earnings was lost in the crisis. There were

questions about how strictly the eligibility rules had been enforced, so it was decided to use a quasi-experimental design to assess the impacts of the program.

The evaluation design (Galasso & Ravallion, 2004) took advantage of the fact that the program was scaling up rapidly, so it was possible to construct a comparison group of families who had not yet received benefits from the program. Propensity score matching (see Boxes A11.1-4 and A11.1-5) was used to match the project and comparison groups on a number of socioeconomic household characteristics. Panel data were also collected by the central government before and after the crisis, and these were used to help remove fixed unobserved heterogeneity through double-difference analysis.

The analysis found that the program's eligibility criteria were not enforced, with about one-third of those receiving program benefits not satisfying the eligibility criteria. Also, about 80% of the adults who were eligible did not receive the program. Using double-difference analysis to control for selection bias, the study did find some positive benefits from the program—most important, a reduction in the drop of income that would have occurred without participation.

Source: Adapted from Khandker et al. (2010).

TABLE A11.1-8 ● Design 2.4: Pipeline Design

	Sample Selection Procedure	T_1 Start of Phase 1	T_2 Implementation Phase 1	T_3 Start of Phase 2	T_4 Implementation Phase 2	T_5 Start of Phase 3
Project group for Phase 1	Random sample of Phase 1 households or communities	$P1_1$	X_1	$P1_2$		
Comparison group for Phase 1 [$C1_1$ and $C1_2$], which is also the project group for Phase 2 [$P2_1$ and $P2_2$]	Random sample of households or communities from Phase 2, from a later phase, or from all phases after Phase 1	$C1_1$		$C1_2 = P2_1$	X_2	$P2_2$
Comparison group for Phase 2	Random sample of households or communities from Phase 3, from a later phase, or from all phases after Phase 2			$C2_1$		$C2_2 = P3_1$

3.2 Pretest–Posttest Project and Comparison Group Design With Judgmental Matching of the Two Samples (Design 2.5)

3.2.1 When to Use

This design (Table A11.1-9) is similar to Designs 2.1 and 2.2 except that the two samples have to be matched judgmentally because secondary survey data are not available to permit statistical matching and the sample is not large enough to permit close matching through regression analysis. So typically, Design 2.5 is used in cases where a project is being implemented in a number of specific locations (e.g., low-cost housing, slum upgrading, micro-catchment irrigation projects) and where the evaluator must use judgment to select comparison communities or areas that match as closely as possible the project locations. Typically, the selection will be based on consultations with local experts and community leaders, review of secondary data (studies, maps, etc.), and where possible rapid visits to possible comparison sites. The evaluator then uses his or her judgment to select the best match.

3.2.2 The Evaluation Design

Box A11.1-9 illustrates how a pretest–posttest comparison group design with judgmental matching was used to evaluate the social and economic impacts of a low-cost housing program in El Salvador.

TABLE A11.1-9 ● Design 2.5: Quasi-Experimental Design With Judgmental Matching of the Two Samples

Time		T_1 (baseline)	Project Intervention	T_3 (end of project)
	Sample selection procedure			
Project group	The project group is selected through either self-selection or administrative selection by the project agency.	P_1	X	P_2
Comparison group	The comparison group is selected through judgmental matching where consultations with experts, community leaders, and other relevant groups are combined with information from secondary sources (agency records, maps, any studies that have been conducted) and, where possible, rapid visits to possible comparison sites. The evaluator then uses judgment to select the best locations or groups.	C_1		C_2

BOX A11.1-9

Case Study for Design 2.5A: Using Judgmental Matching to Evaluate the Impacts of Improved Housing on Household Income in El Salvador

A four-year evaluation was conducted in El Salvador between 1976 and 1980 to assess the impacts of improved housing on poor households in San Salvador, the capital. In 1976, a randomly selected sample of households was interviewed shortly before they entered a self-help housing construction project. A comparison group was selected by combining samples of randomly selected families from the three main types of inner-city housing from which project participants were selected. The samples were repeated in 1980. The survey was combined with various QUAL methods such as participant observation during project workdays and in comparison group communities, case studies of individual families, and interviews with key informants. These assessed the quality of implementation, examined factors affecting the participation of particular groups such as female-headed households and the self-employed, and also documented the influence of certain contextual variables such as the local economy and the involvement of government housing agencies.

It was found that between T_1 and T_3, the average household income for project participants had increased by 70.0%, compared with an increase of 74.6% for the comparison group. Therefore, there was no evidence that improved housing had a positive impact on income, and in fact, the income of the comparison group rose slightly faster. This illustrates the importance of a carefully selected comparison group. If only project participants had been studied, one might have concluded that "improved housing has a significant impact on household income because the income of participants in the low-cost housing project increased by 70% in four years."

Source: Valadez & Bamberger (1994).

3.3 Truncated Pretest–Posttest Project and Comparison Group Design (Design 3.1)

3.3.1 When to Use

This design is used when an adequate budget is available but the major constraint is that the evaluation did not begin until the project was already under way for some time. Often the initial evaluation event will be part of the midterm project review, which is often commissioned between two and three years after the start of a five-year project. The size of the sample and the complexity of the design can be adapted according to whether this is a pilot project or a well-tested program, and according to the types of decisions to which the evaluation will contribute. As with any evaluation design, the pretest–posttest impact assessment should be complemented with a process evaluation.

3.3.2 Description of the Design

The truncated QED design is used when the evaluation cannot begin until around midterm, when the project has already been operating for some time but at least another 18 months remain in its operational cycle. The first observation is considered a proxy (delayed) baseline, while recognizing that there is reduced time to produce and measure outcomes. If there is time (e.g., several more years), observations could be repeated at several points between the midterm and end of the project to observe the process of project implementation. The final observation is then made at T_3, around the time of project completion.

As always, it is recommended that the design be based on a program theory model and include a contextual analysis to assess the influence of the economic, political, organizational, and natural environment within which the different project sites operate and to study the influence of cultural characteristics of the affected populations on program outcomes. Box A11.1-10 illustrates how QUAL and mixed methods were incorporated into the design of an evaluation of a feeder roads project (rural roads connecting villages to the main road network) in Eritrea. These were used for purposes of triangulation (comparing secondary sources with observational estimates of the volume and types of road transport), conducting a process analysis of project implementation, and evaluating the quality of road construction and maintenance. Observational techniques were also used to reduce the costs of obtaining data on vehicular and pedestrian traffic.

TABLE A11.1-10 ● Design 3: Truncated Pretest–Posttest Project and Comparison Group Design Starting at the Time of the Midterm Review

Time		T_1 (baseline)	Project Intervention (continues past midterm)	T_2 (midterm)	T_3 (end of project)
	Sample selection procedure				
Project group	Selection of the comparison group will be based on statistical or judgmental matching. The approach is similar to Design 2, except selection is made when the project has already been operating for some time (there was no initial baseline).		X	P_1	P_2
Comparison group				C_1	C_3

BOX A11.1-10

Case Study for Design 3 (Truncated Design) Combined With Design 2.5B (Judgmental Matching Pretest–Posttest): Assessing the Social and Economic Impacts of a Feeder Roads Project in Eritrea

This evaluation did not begin until the project had reached midterm. It combined an end-of-project survey of the project areas with a simple longitudinal study to observe the process of change. The purpose of the evaluation was to assess impacts of the road on access to schools, use of local health facilities, and volume and prices of agricultural produce sold. The evaluation was not able to identify a comparison group that would serve to assess the three different outcomes, so judgmental matching was used to identify separate comparison groups for assessing each outcome:

- A sample of schools in the same regions as the roads that did not have access to the newly constructed roads. School attendance records were used to measure changes in enrollment in the year prior to the construction of the road and in the year the road was constructed.
- A sample of local health centers in the same region as the roads but that did not have access to the newly constructed roads. Records on the average number of patients attended to each week in the year before the road was constructed and the year the road was constructed were analyzed.
- Almost all agricultural produce was bought and sold in the local cooperative markets. A sample of agricultural markets was selected that served both the villages with access to the new roads and those that did not have access. Volumes and prices of sales and purchases were compared in the year prior to road construction and the year of construction.

Observations were made at three points during the approximate nine-month period of road construction to document changes such as the number of small businesses operating along the route of the future road, the number of pedestrians, the number of trucks and buses, and the number of people visiting the health clinic. Observation was also used during road construction workdays to assess the efficiency of organization and the quality of construction and maintenance. Data reliability was strengthened through triangulation of findings from focus groups, key informant interviews, and direct observation. Baseline conditions were reconstructed through recall, and secondary data compensated for the lack of a comparison group.

Source: Unpublished national consultant report. The study was supervised by one of the present authors.

3.4. Elimination of the Baseline Comparison Group (Pretest–Posttest Project Group Combined With Posttest-Only Analysis of Project and Comparison Group) (Design 4.1)

3.4.1 When to Use

This design is used either when the evaluation begins at the start of the project (Option A) or when the evaluation starts later but previously secondary or recall data representing baseline conditions for the project group can be obtained (Option B). Baseline data are not collected on comparison groups for a number of reasons: to save money, because of technical or ethical difficulties in collecting the data, or because management does not consider this necessary. It is quite common that when the project begins, management does not feel it is necessary to include a comparison group because they feel the purpose of the evaluation is simply to compare the project group before and after the project intervention. However, as the project evolves and there is a need to justify an extension or the launch of a new project, the need for a comparison group is understood. Consequently, it is not uncommon for the evaluator to be asked to create a comparison group for the end-of-project evaluation.

3.4.2 Description of the Design

In this design (Option A), a baseline survey is conducted with the intended project beneficiaries before the project begins, but no comparison group is used at this stage. Only at the time of the final project evaluation (in T_3) is a survey conducted that includes both project and comparison groups. This design works reasonably well for assessing how a project is being implemented and whether it is able to produce the intended outputs. It is also able to compare the characteristics of the project group and the comparison group at the time the project completes its work. If, retrospectively, it can be adequately documented that the comparison group was essentially the same as the project group at the time the project started (T_1), this design may be sufficient to demonstrate project effects. Table A11.1-11 describes the logic of this design.

For example, with a rural road construction project, surveys and participatory consultations with the community may have identified a number of factors affecting the willingness of the community to participate in the project and the benefits they obtain from the road. These factors might include whether local culture permits women to participate in road construction and to travel to the market, the distance from the local market, and the agricultural surplus available to sell. A comparison group, if it is well selected, could rate other local communities on these variables and hence determine the likelihood that the project would be well received and would have an impact in other areas. The project and comparison groups could also be compared on indicators such as amount of produce sold in the local markets, average number of trips and distance traveled, and kinds of consumer goods available in community shops.

As always, the design should be based on a program theory model and should be combined with a process evaluation and a contextual analysis.

3.4.3 Incorporating Mixed-Method Approaches Into the Design

In addition to the incorporation of a process and contextual analysis, a mixed-method approach can be very helpful for reconstructing the baseline conditions of the comparison group. For the interpretation of the findings, it is important to have the best information possible on how similar the characteristics of the project and comparison groups were at the time the project started. In the case of the Maharashtra (India) irrigation project (see Box A11.1-11), QUAL methods (informal conversations with neighbors and key informants) were used to identify former residents who had not been eligible for compensation of new land and who had moved out of the area. A tracer study was then conducted with these families to compare their situation with the project families who had received compensation.

TABLE A11.1-11 ● Design 4: Elimination of the Baseline Comparison Group (Pretest–Posttest Project Group Combined With Posttest Analysis of Project and Comparison Groups)

Time		T_1 (baseline)	Project Intervention	T_3 (end of project)
	Sample selection procedure			
Project group	Purposeful selection of project participants	P_1	X	P_2
Comparison group	No baseline comparison group. Posttest comparison group selected statistically or judgmentally depending on the availability of secondary data and the sample size.			C_1

BOX A11.1-11

Case Study for Design 4.1 (With Elements of Design 1): Comparing the Effects of Resettlement on Project Beneficiaries and Nonbeneficiaries in the Second Maharashtra Irrigation Project, India

Sample surveys were conducted periodically between 1978 and 1985 in areas from which families were to be resettled as a consequence of a large-scale irrigation project. The study covered only families who were eligible to receive land or housing plots in the relocation areas. The surveys were repeated in 1990 after project families had been relocated. An ex-post comparison group survey was conducted in 1990 with a sample of families who had remained in the command area of the irrigation project. This was not an ideal comparison as many of the sample households forced to move as a result of the dam had remained in the project, so their situation did not really represent families not affected by the dam. Recognizing that no information was available on the approximately 45% of families who were forced to relocate but who were not eligible for compensation, a tracer study was conducted in 1990 to try to identify these families. The study found that the economic conditions of most families receiving compensation had improved. The situation concerning the families who had not received compensation was more mixed, but in general, forced resettlement appeared to have had less negative consequences than had been expected.

Source: Valadez & Bamberger (1994, pp. 264–266, summarizing World Bank, 1993).

3.5. Posttest-Only Comparison Group Design (Design 5)

3.5.1 When to Use

This design is used when the evaluation does not begin until near the end of project implementation or the project has recently been completed. Despite the late start, evaluations using this design are often quite well funded, and it is often possible to administer surveys to quite large samples of subjects in both the project and comparison groups. In other cases, the evaluation is based on secondary data from household, agricultural, labor force, and other surveys that had been collected for some other purposes.

3.5.2 Description of the Design

This design relies entirely on an end-of-project comparison between the project group and a nonequivalent control group (comparison group), and no baseline data were collected on either the project or comparison areas. It can be used to obtain an approximate estimate of project impacts. The design works better in isolated communities where the project is the only major outside intervention so that it is less important to isolate the effects of other interventions taking place at the same time. It can also be used to compare the characteristics of project participants with people from other similar communities. If project households have similar characteristics to other communities (other than those characteristics affected by the project's interventions), then it is more likely that the results of the pilot project can be generalized. If, on the other hand, there are significant differences between the two groups, it will be more difficult to generalize. Table A11.1-12 explains the logic of the design.

The fact that large, carefully selected samples are often used means that the analytical power of the design can be strengthened through the use of multivariate analysis to statistically control for differences between the two groups.

3.5.3 Incorporating Mixed-Method Approaches Into the Design

Often this design relies exclusively on QUANT data collection and analysis, particularly when the evaluation is based on survey data collected for some other purpose. A fundamental weakness of the design is that it does not control for historical events or for preexisting differences between the project and comparison groups. Mixed methods can significantly strengthen the design by obtaining information on the characteristics of the two groups at the time the project began.

This can be obtained through focus groups, participatory appraisal (PRA) techniques (see Chapter 5), or interviews with key informants. For example, in the evaluation of the impacts of microcredit programs in Bangladesh (Box A11.1-12), the analysis found that when women obtained loans, this had a greater impact on household expenditures, investment in housing, and children's school enrollment than when men obtained loans. It was assumed that the difference was due to the strong emphasis of the credit groups on group development and strengthening women's self-confidence. However, an alternative hypothesis, which could not be tested on the basis of the survey data, would be that the women who decided to apply for loans already had more self-confidence and perhaps entrepreneurial experience. This hypothesis could have been tested through some of the mixed-method approaches described above. This assumes, of course, that it would be possible for the researchers to go to the field to apply these techniques.

BOX A11.1-12

Case Study for Design 5: Assessing the Impacts of Microcredit on the Social and Economic Conditions of Women and Families in Bangladesh

In 1991–1992, a random sample of households was interviewed in a sample of villages where some of the leading village bank programs were operating in rural Bangladesh. A comparison group was interviewed in villages where none of these village bank programs were operating. The surveys were conducted ex-post when the village banks had already been operating for several years. No baseline information had been collected on the condition of families before the banks began to operate. It was found that borrowing from a village bank had a much greater impact on women than on men (although the latter also benefited). Per capita household expenditures increased almost twice as fast when women received loans rather than men, housing conditions improved, and personal savings increased. Interestingly, it was found that contraceptive usage declined for women borrowers, and their fertility increased. The lack of baseline data made it difficult to determine to what extent the observed differences between the project and comparison groups were due to the effects of the project or whether they were due, at least in part, to differences that already existed before the project began.

TABLE A11.1-12 ● Design 5: No Baseline Data (Posttest-Only Project and Comparison Groups)

Time		T_1 (baseline)	Project Intervention	T_3 (end of project)
	Sample selection procedure			
Project group	Purposeful selection of project participants	P_0 secondary data sometimes permit the reconstruction of baseline data (see Chapter 5).	X	P_1
Comparison group	Many evaluations use secondary data to select the sample and to use statistical matching. In some cases, secondary data are also available at the time of project launch as this is used to reconstruct baseline data. When secondary survey data are not available, judgmental matching must be used.	C_0 secondary data are sometimes available to reconstruct a baseline comparison group. See Chapter 5 for a discussion of ways to address the issue of missing variables (also called "unobservables").		C_1

Sources: S. Khandker (1998), Baker (2000, Annex 1.2), and World Bank (2001).

BOX A11.1-13

Example of Design 5B: Posttest Comparison Combining Statistical Matching With a Mixed-Method Design—Evaluating Nicaragua's School-Based Reform: A Retrospective, Mixed-Method Design With Statistical Matching

The decentralization reform in Nicaragua was a principal element of the coalition government that replaced the Sandinista regime in 1990. The reform aimed to give schools power over key managerial and pedagogical decisions and transfer financial administration directly to the schools. By the end of 1995, more than 100 secondary schools had signed a contract with the Ministry of Education to establish a directive council and become autonomous. The three questions addressed in the first phase of the evaluation were as follows:

- Whether autonomous public schools exercise greater control over their management than do traditional public schools
- Whether (and which) local stakeholders (directors, teachers, council members) affect school decisions
- How local stakeholders perceive the changes that have occurred in schools since autonomy

The evaluation was not commissioned until 1995, at the start of Phase 3 of the reform program but when the program had already been under way for five years. Consequently, a retrospective mixed-method evaluation design was used. The mixed-method design permitted the use of triangulation to first *corroborate* the quantitative findings and to ensure *convergent* validity and, second, to use qualitative techniques for *elaboration* to expand our understanding of the reform as presented by the initial results of the quantitative studies.

A matched comparison group design was used in which autonomous (reformed) schools were compared with nonautonomous (traditional, nonreformed) schools. Matching was based on the timing of the reform and the school's size and location. The lack of pre-reform baseline data meant that the matching was not as precise as when secondary data permitted the use of propensity score matching. Three quantitative data-collection instruments were used:

- A school survey using a random sample of 242 schools, including both autonomous and nonautonomous schools at both the primary and secondary levels. The survey covered, among other indicators, enrollment, level of absenteeism, grade repetition and drop-out, physical resources, training and experience of staff, and changes since reform. A special questionnaire was developed to determine whether the school made important decisions and to determine whether respondents felt influential in the decision-making process.
- A household survey covering 3,000 randomly selected students from the surveyed schools and followed to their household. This determined the socioeconomic status of the households and parents' participation in school affairs.
- Achievement tests in math and language were applied to a sample of third-grade primary and second-year secondary students to compare schools' academic performance.

Qualitative methods were applied in a subsample of 18 schools to develop typologies, assess beneficiary perspectives, examine the context in which the reform was introduced, and examine the decision-making dynamics in each school. The aim was to detect patterns and to highlight variations across schools and actors. Methods included focus groups with parents, teachers, and school council members, and key informant surveys with the school director and local officials from the Ministry of Education.

The initial findings showed that the reforms were successful in expanding the role of the schools in governance. In the quantitative surveys, about half of the respondents reported that school performance had improved in the reform schools, while the other half said there was no change.

Source: Rawlings (2000).

4. NONEXPERIMENTAL DESIGNS (NEDs) (DESIGNS 6 AND 7)

We classify nonexperimental designs (NEDs) into two categories: those that include baseline data on the project group (so that a pretest–posttest project group–only design can be used) and designs that only collect data on the project group

at one point in time, normally toward the end of the project cycle. There is another important distinction that we emphasize and that concerns the reasons for selecting a NED. Many evaluation textbooks assume that NEDs are used only as a last resort when budget and time constraints do not permit the use of a "stronger" evaluation design. However, there are many situations in which experienced evaluators select a NED, considering it the strongest design for assessing causality in a given context. So we make a distinction between situations in which a "basic" NED is used as a default option due to budget and time constraints and situations in which NEDs are considered the strongest available design.

4.1. Pretest–Posttest No-Comparison-Group Designs (Design 6)

4.1.1 When to Use

With this design, the evaluation begins at the start of the project, but for budget, technical, or political reasons, baseline data are collected only on the project group, and there is no comparison group, either at the beginning or at the end of the project. In some cases, this design is selected in consultation with the client due to budget constraints. However, there are other situations in which it was originally intended to collect baseline data on a comparison group, but this proved to not be possible due to political or technical reasons. The design works reasonably well for projects using "best practice" interventions that have been previously proven to work under very similar conditions—for example, the construction of a village school or clinic where there was previously no such facility within easy access. It can also work well when the purpose of the evaluation is to understand the project implementation process and where QUANT assessment of impacts is less important.

4.1.2 Description of the Design

This design is based on a comparison of baseline data collected on the project population at the start of the project (pretest), with similar posttest data collected toward the end of the project. No comparison group is included, and consequently, it is not possible to use a conventional counterfactual to control for alternative explanations of the observed changes. In its basic form, this design is very weak because it implicitly assumes that all of the observed changes in outcome indicators are due to the project intervention.

As always, the design should be based on a program theory model and should be combined with a process evaluation and a contextual analysis. However, in many cases, none of these options are included due to budget and time constraints.

An Option B will sometimes be used when the evaluation is commissioned at the end of the project and baseline data are "reconstructed" using some of the strategies discussed earlier.

Table A11.1-13 presents the basic design.

Box A11.1-14 illustrates how this basic design was used to evaluate the impact of a hydroelectric project in Thailand on households that were resettled.

TABLE A11.1-13 ● Design 6: Basic Pretest–Posttest Project Group Design With No Comparison Group

Time		T_1 (baseline)	Project Intervention	T_3 (end of project)
	Sample selection procedure			
Project group	Random sample of project participants at start of the project. If a panel sample is used, the same subjects will be reinterviewed at T_3 (with appropriate procedures to adjust for subjects who cannot be reinterviewed or new subjects who have entered) (see Chapters 11 and 12). A second option is to select a new random sample at T_3.	P_1	X	P_2
Comparison group	No comparison group was included in the design.			

Source: Valadez & Bamberger (1994, pp. 259–261).

BOX A11.1-14

Case Study for Design 6.4: Using a Before-and-After Survey of Resettled Households to Evaluate the Impact of the Khao Laem Hydroelectric Project in Thailand

The project called for the involuntary resettlement of 41 affected villages with a total of 1,800 families. A survey of 50% of the intended beneficiaries was conducted in 1978–1979 prior to the start of the project, and a follow-up survey was conducted in 1989–1990 with 200 resettled families. No formal comparison group was used either before or after resettlement, although the research team consulted available secondary sources. While the comparison of quantitative surveys conducted before and after resettlement showed that families were better off on the basis of a set of economic indicators, a qualitative survey found that the majority of families considered themselves to be worse off. No information was available on the families who did not move to resettlement areas or on the 30% who did not remain in the project.

4.1.3 The Single-Case Design (SCD)

This design has traditionally been used in the behavioral and health sciences to assess the effects of a treatment on an individual or small group, for example, in a school classroom or a managed health care facility. However, there has recently been renewed interest in this design due to the recommendation of a panel of experts convened by the Department of Education that the SCD, when properly administered, could be considered a methodologically rigorous design for assessing a range of educational interventions (Kratochwill et al., 2010). According to Kratochwill et al. (2010), SCDs have the following features:

- An individual "case" is the unit of intervention and the unit of data analysis. A case may be a single participant or a cluster of participants (e.g., a classroom or community).
- Within the design, the case provides its own control for purposes of comparison. For example, the case's series of outcome variables are measured prior to the intervention and compared with the measurements taken during (and after) the intervention.
- The outcome variable is measured repeatedly within and across different conditions or levels of the independent variable. These different conditions are referred to as phases (e.g., baseline, intervention phase).

Very rigorous evidence standards must be applied before the results can be considered convincing evidence of a causal relationship. The present authors are not aware of the SCD approach having been applied outside school, clinical, and managed health care settings, so it is not yet clear whether and how the approach could be used in development evaluation. However, the SCD approach is potentially attractive for the evaluation of development interventions at the level of a village or small group because an analysis of the specific group context and behavioral dynamics might produce change. It would then be possible to test whether similar results were found in similar settings to assess how far it would be possible to generalize the results. This approach is at the other end of the spectrum from RCT and other statistical designs that estimate average effect for a large population but can say nothing about the likely effect of the treatment on a particular individual or group.

Box A11.1-15 illustrates how SCD was used to evaluate the effectiveness of a behavioral therapy for a schoolgirl with Asperger's disorder. This study used the "withdrawal design" (defined as ABA) in which the baseline condition is observed and measured for some time to obtain a stable measurement, and the treatment is administered for a certain time (which can be as short as a few minutes or may last several weeks) and is then withdrawn. This treatment cycle is repeated several times (at least three times under similar conditions and ideally four or five times), and if the same pattern of change is observed (and if this is rated as large enough to be significant) in each cycle, this is considered evidence that the treatment has been successful. For research purposes, the cycle would be repeated in other similar situations to build a body of evidence on the conditions under which the treatment appears to be successful.

BOX A11.1-15

Design 6.1: Single-Case Design: Treating a Schoolgirl With Asperger's Disorder

Lakeesha was a 12-year-old African American girl beginning seventh grade and diagnosed with Asperger's disorder. Some of the symptoms included difficulty in interacting with peers, reduced eye contact, "odd" facial expressions and gestures, a lack of social or emotional reciprocity, and failure to develop appropriate peer relationships. The treatment involved the most common single-case "withdrawal design" in which baseline measurements are made, usually based on observation, followed by the administration and then withdrawal of the treatment. The indicators are recorded during treatment and again after withdrawal. This is referred to as the ABA single-case design. The ABA cycle is repeated three or, ideally, four times: ABA ABA ABA ABA. A group of experts reviews the observational data, and if significant change has occurred in each cycle, this is considered credible evidence that the treatment had an effect.

In the case of Lakeesha, the "buddy" system was used in which another child with similar interests (in this case, interest in solving geometry proofs and interest in animal tracks) was trained to help Lakeesha master and use five sets of skills in which she had been coached: (a) sharing ideas, (b) complimenting others, (c) offering help or encouragement, (d) recommending changes nicely, and (e) exercising self-control. The peer buddy was asked to reinforce Lakeesha simply by smiling at her or telling her she was doing a great job when she engaged in one of these skills.

Prior to initiating the program, Lakeesha's behavior was tracked over a four-week baseline period. On one of the indicators, initiating conversations with others, she did not initiate any conversations over the four-week observation period. During the first five-week treatment period, the number of conversations Lakeesha initiated increased steadily to nine per week. During the first withdrawal period, the number dropped steadily to zero after three weeks. During the second treatment period, the number of conversations initiated rose steadily to eight. By chance, the initial peer buddy moved to another school and was no longer able to participate, and a new buddy was trained. The results were similar with the second buddy, thus reinforcing the conclusion that the treatment was effective.

This summary greatly oversimplifies the design and the care that is needed during administration and analysis to control for internal and external threats to validity.

Source: Adapted from Morgan & Morgan (2009).

4.1.4 Longitudinal NED Design (Design 6.2)

A longitudinal NED involves the continued observation of a sample of individuals or groups over a long period of time (but no comparison group). Observations are made on the same sample on a regular basis to observe the process of change. Often the design will include a pretest–posttest comparison to measure changes that have occurred over the life of the project, but the periodic observations also make it possible to study the process of change, focusing on both how the program is implemented and how implementation and the response of the subjects are affected by local contextual factors. As the NED does not include a comparison group, the effects of the intervention are estimated by detailed description of the processes of change.

Longitudinal NED designs can also be used to provide a broad analysis of the contextual factors that influence the operation of a particular institution such as a school. These studies can help define the broader contextual factors that need to be taken into consideration when evaluating the effectiveness of programs designed to improve the performance of institutions such as schools or to increase their accessibility to low-income and vulnerable groups.

Box A11.1-16 describes a seven-year longitudinal study that examined the interactions between school and education for teenage girls and boys from different backgrounds in Australia.

BOX A11.1-16

Design 6.2: Longitudinal Design Without Comparison Group: The 12–18 Project—Making Lives Modern in Australia

The 12–18 Project was a study of subjectivity, schooling, and social change funded by the Australian Research Council. Over a seven-year period (1993–2000), the researchers interviewed and videotaped 26 young Australians (14 girls and 12 boys) as they aged from 12 to 18 years. The young people came from diverse backgrounds and attended four different types of schools. Interviews were undertaken twice annually over the high school years and twice in the year afterwards. In the researchers' words, "We listened to these students talk about their sense of self, their values, attitudes to the future, and their experiences of school. Their individual narratives illuminate the uneven and differentiated impact of contemporary social and gender change, and the profound influence of school, community and culture on the shaping of subjectivity" (McLeod & Yates, 2006, p. 2).

Central to the research design of the study was a focus on *school culture*, and the sample of 26 was carefully constructed to include young people from similar class backgrounds going to different schools, as well as those from a different class background in the same school—avoiding the conflation of the "habitus" of school and family that characterizes so much educational research. So although the study followed individuals over time, it was also a comparative study of institutional culture and the way that institutions shape subjectivities.

Source: McLeod & Thomson (2009).

4.1.5 Interrupted Time Series (Design 6.3)

An interrupted time-series design can be used when there is a continuous series of data that begins well before the intervention being studied and continues during the implementation and for some time afterwards (Shadish et al., 2002, Chapter 6). Typical examples include information on traffic accidents or arrests of drivers, records of malnutrition or low birth weight children, school enrollment and attendance records or school test scores, and recidivism rates (the proportion of criminals released from jail who are rearrested within a certain time, often six months). The design is used to determine whether there is a significant break in the time series or a change in the slope at the point where an intervention, such as a change in the traffic laws or the legal drinking age or an incentive program to increase school enrollment, is introduced. For example, traffic accidents or driving arrests might fall after a new law is introduced, or school enrollment might increase.

Box A11.1-17 presents an example of an interrupted time-series analysis used to assess the effect of raising the legal drinking age on the number of drinking-related traffic accidents.

BOX A11.1-17

Design 6.3: Interrupted Time Series: Estimating the Effects of Raising the Drinking Age

During the early 1980s, many U.S. states raised the minimum drinking age from 18 to 21, especially after passage of the Uniform Drinking Age Act of 1984, which reduced highway construction funds to states that maintained a drinking age younger than 21. Wisconsin raised its drinking age to 19 in 1984 and then to 21 in 1986. To assess the impact of these changes, David Figlio (1995) examined an 18-year time series of monthly observations of alcohol-related traffic accidents, stratified by age, that was available from the Wisconsin Department of Transportation from 1976 to 1993. Statistical time-series models were fit to the data for 18-year-olds (who could legally drink prior to 1984), for 19- and 20-year-olds (who could legally drink prior to 1986), and for those older than 21 (who could legally drink over the whole time period).

The outcome variable in these analyses was the rate of alcohol-related crashes per thousand licensed drivers in the respective age-group.

The results showed that for 18-year-olds, raising the minimum drinking age to 19 reduced the alcohol-related crashes by an estimated 26% from the prior average of 2.2 per month per 1,000 drivers. For 19- and 20-year-olds, raising the minimum age to 21 reduced the monthly crash rate by an estimated 19% from an average of 1.8 per month per 1,000 drivers. By comparison, the estimated effect of the legal changes for the 21 and older group was only 2.5% and statistically nonsignificant. The evaluator's conclusion was that the imposition of increased minimum drinking ages in Wisconsin had immediate and conclusive effects on the number of teenagers involved in drinking-related crashes, resulting in substantially fewer crashes than prelegislation trends would have generated.

Source: Rossi, Lipsey, & Freeman (2004, Exhibit 9-H, p. 293). Adapted from Figlio (1995).

4.1.6 Case Study Designs

"Case studies explore real-life events in a natural setting" (Yin, 2004, p. xii). Chapter 13 described a number of different ways that case study designs can be used in program evaluation and the different sample selection procedures that can be used depending on the purpose of the study. Case studies collect detailed information on a relatively small number of cases (individuals, groups, schools, communities, etc.). It is also possible to conduct a single case: for example, the effects of a mass vaccination program on the American public in the 1970s or a methadone maintenance program in Syracuse (Yin, 2004). The cases may be conducted at one point in time, or they can collect information over time. Normally, the purpose of a case study is to help understand what meaning people give to a program, how they perceive its purpose, their attitudes and expectations, how they respond to it, how their response is affected by contextual factors, and what effects it has on them. Cases can be selected to be representative of the broader population, or they can be selected to include subjects with particular characteristics (those who benefited most or least from the program, outliers with unusual responses, etc.). Due to the costs and time involved in conducting case studies, the number of cases will normally be quite small, so their selection is very important if the purpose is to generalize from these cases to the broader population.

Box A11.1-18 describes a rigorous case study evaluation of a Natural Resources Leadership Program that was selected as an example of an exemplary evaluation design for inclusion in "Evaluation in Action: Interviews With Expert Evaluators" (Fitzpatrick, Christie, & Mark, 2009). The evaluation is based on a detailed analysis of six training courses and combined document review, observation of the training programs, interviews with participants and their managers, monitoring of a one-year practicum that followed the training, and follow-up interviews to assess the effects of the program.

BOX A11.1-18

Example of a More Rigorous Nonexperimental Design: Evaluating the Effectiveness of the Natural Resources Leadership Program

The Natural Resources Leadership Program (NRLP) was designed to introduce new approaches to the resolution of environmental conflicts for natural resources leaders in three southeastern states. These new ideas focused on reframing conflicts as opportunities for progress, rather than as fights to be won or lost, and on re-envisioning leadership as facilitating a consensual agreement rather than as persuasion to a claimed position. The leadership program was offered to approximately 150 leaders from public and private sectors and from environmental activist groups and industry. The program was implemented through a series of five 2½-day sessions of residential instruction, spaced out over a period of six months in communication and conflict-resolution skills, leadership development, and government and public policy

(Continued)

(Continued)

resources, complemented by a trip to Washington and a year-long practicum.

The evaluation design was organized around case studies of each of the six leadership programs. The evaluation began with an intensive case study of the first pilot training program that included a review of all program materials, observation of all instructional sessions, interviews during the sessions with most participants, repeated interviews in person and over the phone with program staff, attendance at Advisory Board meetings, and in later years follow-up phone interviews and surveys of participants. This intensive study provided the evaluation team with a deep understanding of the program as designed, implemented, and experienced and permitted grounded development of later instruments.

The case studies of the other five leadership programs were less intensive as the evaluation resources had to be distributed across states and time. In each case, two training sessions were observed, and follow-up data were collected through phone interviews and surveys. It was also necessary to track the practicum experiences and the longer-term program outcomes for the participants.

An important aspect of the evaluation was to assess how the program affected the ability of the graduates to engage differently in real-life environmental disputes. This was done by surveying Advisory Board members, conducting mini-case studies of selected participant practicum projects, and surveying supervisors and other key individuals in participants' work sites. The "best" of the practicums were sampled to understand how the program had contributed to their success.

The evaluation determined that the NRLP was generally successful in realizing its learning aims. Most participants reported that the program changed their conceptual understanding of environmental conflict and changed their ideas of successful leadership in conflict situations, and they learned new skills and techniques for organizing people and information toward the resolution of conflicts. In terms of effects on practice, only a few participants were able to enact the new lessons learned in the field.

Source: Fitzpatrick et al. (2009).

4.2. Data Collected Only on the Posttest Project Group (Design 7)

4.2.1 When to Use

Design Framework 7 is statistically weak because the design does not include either a baseline or a matched comparison group. The absence of a pretest makes it difficult to know if a change has occurred, and the absence of a nontreatment comparison group makes it difficult to know what would have happened without the project treatment (Shadish et al., 2002, p. 106). It is also difficult to obtain precise QUANT estimates of project outcomes or impacts. Despite these limitations, by default, Design 7 is probably the most widely used RWE scenario, mainly because, all too typically, evaluation was not planned from the beginning of the project, nor was it considered necessary or practical to include a counterfactual. In addition, all too often when asked to conduct an evaluation at the end of a project, the evaluator is given very little time (sometimes as little as one or two weeks) and a very limited budget (sometimes as little as a few thousand dollars).

Nevertheless, it is important to appreciate that a wide range of different evaluation approaches can help to strengthen Design 7, and many qualitative and mixed-method evaluations are considered methodologically sound within their respective paradigms and use nonstatistical approaches to assess outcomes and impacts.

There are three main situations in which this design is used. The first is where the project being evaluated is very small, perhaps operating in only one location, or where this is an exploratory study where the main purpose is to obtain an initial assessment of whether the project "works" and whether it seems potentially able to achieve its stated objectives. In this context, *works* might mean any of the following:

- Are women able and willing to apply for loans and invest the proceeds in a small business?
- Do most residents use the community toilets, and are they maintained in good working order?
- Are teachers able to apply the new teaching tools and methods, and are there preliminary indications that they affect students' behavior and performance?

The second situation is where the project is quite large and clients are interested in obtaining estimates of outcomes and impacts, but there was no life-of-project evaluation plan or even a baseline. In this context, the evaluators are aware that they have to do the best they can with a methodologically weak evaluation design.

The third situation is where the evaluation was planned using a mixed-method or QUAL design and where the focus was on understanding the implementation process, the influence on the project context, and the perspectives and experiences of different groups affected by the project (including groups that did not benefit or might even have been affected negatively). Some of these designs are discussed in Chapters 13 and 14.

When using Design 7, the scoping of the evaluation (Step 1, see Chapter 2) is particularly important to fully understand the client's information needs and how the evaluation results will be used.

4.2.2 Description of the Design

In this design, only the project population is studied, and they are surveyed only after the project has been implemented. Data may be collected from a small, rapid sample survey; from QUAL methods (PRA, focus groups, secondary sources, key informants, etc.); or from a mixed-method design combining QUANT and QUAL methods.

4.2.3 Strengthening the Design

Given its methodological weaknesses when used under severe budget and time constraints, it is important to strengthen the design by using some of the approaches described earlier in this appendix. Maximum use should be made of mixed-method approaches (see Chapter 14). Even when operating under time pressures, efforts should be made to construct at least a simple program theory model using the techniques described in Chapter 10 to obtain an approximate estimate of causality. This permits the use of logical deduction through techniques such as pattern matching (Campbell, 1966) or coherence (Rosenbaum, 1995) and through the nine strategies proposed by Davidson (2000, pps. 21–22), described in Chapter 10. Some of these approaches have been described as being analogous to detective work in which a causal sequence is deducted from observing all the clues, and alternative explanations are eliminated through evidence. Readers should also be aware of the criticisms of the use of program theory to estimate causality (Cook, 2000, pp. 29–32).

4.2.4 Incorporating Mixed-Method Approaches Into the Design

The evaluation of village schools in Eritrea (Box A11.1-19) illustrates the use of a mixed-method approach. To try to make up for the lack of pretest and comparison data, recall was used to assess school attendance prior to the construction of the school, and these estimates were triangulated with key informant interviews and a review of school attendance records. Table A11.1-14 explains the logic of the design.

TABLE A11.1-14 ● Design 7: Posttest Analysis of Project Group Without a Baseline or Comparison Group

Time		T_1 (baseline)	Project Intervention	T_3 (end of project)
	Sample selection procedure			
Project group	[P_0] Baseline data may be reconstructed from project records or from the other techniques discussed in Chapter 5.		X	P_1
Comparison group	[C_0] Small, purposively selected samples might be used.	QUAL techniques such as PRA, in-depth interviews, and secondary data may be used to reconstruct baseline data.		Secondary data or QUAL techniques may be used to ask key informants and even project participants how they think they compare to persons who did not participate in the project.

BOX A11.1-19

Case Study for Design 7: Assessing the Impacts of the Construction of Village Schools in Eritrea

In the evaluation of the Eritrea Social Fund, an end-of-project survey was conducted in 48 communities representing the catchment area for 10 newly constructed primary schools. No comparison group was used. Baseline data on school attendance prior to the construction of the schools were estimated by asking families to recall the situation before the schools were built. Recall data seem to have been reliable because it was easy for families to recall whether their children attended school before the village school was built and also because they did not have any incentive to give wrong information. Triangulation was used to compare estimates from recall with key informant interviews and a review of school attendance records. The analysis focused on the following topics:

- Process evaluation: More than 90% confirmed that the school was a high priority, but only 37% had attended meetings to participate in planning the project.
- Accessibility, impact, and gender: The schools were successful in reaching the poorest sectors of the community; it was more difficult to involve recently returned refugees because they were still unsettled and not motivated to send their children to school; families are equally motivated to send boys and girls to school, but if they have to choose for economic reasons, they normally give priority to a boy.
- Social impact: Local school construction reduced travel time (to other schools) for students by one half to two thirds.
- Sustainability: Despite extreme poverty, almost all households contributed the required 10% of the cost of the school in cash, labor, or materials.

Source: Unpublished national consultant report.

APPENDIX 11.2 THE RWE APPROACH TO THE CLASSIFICATION OF FACTORS AFFECTING THE CHOICE OF EVALUATION DESIGN

The RWE approach distinguishes among the following:

a. The stakeholders and *the key evaluation questions* to which they need answers.

b. The *value orientation of the evaluation*. Many evaluations have an explicit or sometimes implicit value orientation which must be understood as it has an important influence on how the evaluation will be designed and the findings used. For example, the evaluation may seek to promote gender equality, emancipation, or social justice. In many other cases the evaluators may either not mention values or may claim their evaluation is "value-free." It can be argued that the promotion of experimental evaluation designs implies an implicit value orientation as it promotes a belief in a certain kind of evidence, and often excludes or undervalues other kinds of evidence.

c. *The context within which the evaluation will be conducted.* These are scenarios, many of which contain factors beyond the control of the evaluator, that narrow down the range of possible evaluation designs that could be used. These factors include when the evaluation is commissioned, whether or not there will be (or was) a baseline, midterm, end-line, and/or ex-post assessment; the kinds of secondary data that are available; and whether or not and when it will be necessary or possible to select a comparison group. Another factor that can have an important influence on the evaluation design is when the project being evaluated uses a *results framework*. In these instances the evaluation is often required to only evaluate the outputs, outcomes, and impacts included in the results framework. Chapter 17 shows how this requirement can seriously constrain the evaluation of gender outcomes as many of the potentially most important outcomes (such as women's increased control over household resources and decision making) are frequently not included in the results framework. Table 11.3 describes seven design frameworks. In Appendix 11.3 we identify some of the strengths and weaknesses of these designs. They are covered in more detail in Appendix 11.1.

d. *The evaluation design options.* Within each of the evaluation frameworks there are a number of design options that can be considered. The design options include the procedures for selecting the comparison group and defining the counterfactual, the choice between quantitative, qualitative, or mixed-method approaches, whether the intervention is considered complex, and whether it is possible (and useful) to incorporate data science (big data) approaches for data collection and analysis. Within a particular evaluation framework the choice of methods will also vary according to the point in the project/program cycle when the evaluation is commissioned. In Table A11.1-1 (Appendix 11.1) we list a total of 19 more nuanced evaluation designs. These include options for evaluation designs that are commissioned at the start of the project (Option A) and at the end of the project (Option B). They also identify different ways that the comparison group can be selected for experimental and quasi-experimental designs. Each of these design options is discussed in more detail in Appendix 11.1, which also includes cases illustrating most of the designs.

e. *Tools and techniques to strengthen any evaluation design:* Though the experimental and quasi-experimental evaluation designs can address statistical sampling issues such as controlling for selection bias, they frequently do not address the many other threats to validity of the evaluation design (see Appendix 11.3). Similarly, qualitative and mixed-method designs face different sets of threats to validity. The techniques described in that table can be used to strengthen all evaluation designs.

Note: All of the evaluation design frameworks and scenarios are applicable to any of the approaches and data-collection methods. But there are major differences in how data (or evidence) is obtained and processed for each approach.

APPENDIX 11.3 THE STRENGTHS AND WEAKNESSES OF THE SEVEN RWE EVALUATION DESIGN FRAMEWORKS

Design	Advantages	Disadvantages
1. Comprehensive longitudinal design with pre-, mid-, post-, and ex-post observations on the project and comparison groups	This is the strongest design framework, studying both the implementation process and sustainability. May be required for research testing new project innovation that, if sustainable impact can be proven, will be expanded to a much greater scale.	• Requiring multiple evaluation events or observations during and after the life of a program, it is the most expensive, the most time-consuming, and the most difficult to implement.
2. Pretest–posttest project and comparison groups	This is a strong, general-purpose experimental or quasi-experimental design. With a well-selected comparison group, it provides good estimates of project impacts.	• Assumes the comparison group is reasonably similar to the project group and willing to participate in two surveys even though they receive no benefits. • Does not assess project implementation or sustainability.
3. Truncated pretest–posttest project and comparison group design (beginning at midterm, no pre-project baseline)	• Observes implementation process as well as impacts. Reasonably robust model, particularly for projects in which implementation begins slowly, so that not too much is missed by starting the evaluation late.	• Does not begin until around project midterm, so the project startup and initial implementation period are not captured.
4. 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 comparison groups, at least at the end of the project. • 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.
5. 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 measure the exact magnitude of project impacts or even changes over time. • Does not control for local history. • Does not assess potential for replication on a larger scale. • Does not study project implementation process.

Design	Advantages	Disadvantages
6. Pretest–posttest project group	• Provides an approximate estimate of project impacts, or at least changes in outcome indicators during the life of a program.	• Does not compare changes in project population with other communities (i.e., counterfactual). • Does not control for local history. • Does not control for the effect of intervening variables through the use of multivariate analysis.
7. Posttest only, only on project group	• Useful for exploratory studies to get a general idea of how well the project model works. • Provides a first, approximate estimate of results, particularly for small or isolated projects.	• Though recall or other methods could be used, this scenario or framework does not directly measure change occurring during the life of the project. • Hard to feel confident that any purported changes are due to the project and not to other factors or interventions. • Does not directly control for external events; nor does it obtain comparative data to estimate attributable impact.

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 10 and with some of the RWE techniques discussed later in this chapter. For Designs 1, 2, 3, 4, and 5, which use comparison groups, the analysis can be greatly strengthened by using multiple regression analysis to statistically control for differences in the characteristics of the project and control groups. Where appropriate secondary data are available, these designs can also be strengthened through statistical matching techniques such as propensity scores and instrumental variables.

APPENDIX 11.4 CHALLENGES FACING THE USE OF EXPERIMENTAL AND OTHER STATISTICAL DESIGNS IN REALWORLD EVALUATION CONTEXTS

Chapter 12, Section 2, discusses some of the challenges facing the application of experimental and quasi-experimental designs in real-world contexts (see Box 12.3 in Chapter 12) and identifies some of the challenges facing the application of experimental and other statistical designs in many RealWorld Evaluation contexts. These factors must be taken into consideration when selecting the appropriate evaluation design.

In the real world, evaluators trying to approximate the most robust designs typically face one or more of the following problems:

- It is rarely possible to randomly assign subjects to experimental and control groups. For logistical, administrative, political, and sometimes ethical reasons, most projects are accessible to or affect everyone in a given community or area. For example, a school, water supply system, or road will usually be accessible to all families in the community, and it is difficult to tell some families they cannot send their children to the school or use the water (assuming they are willing to pay or that they have participated in the construction of the water system).
- Some projects use a self-selection process when, for example, people decide if they wish to apply for microcredits, enroll in a literacy class, or plant new varieties of seed. In these cases, it is likely that the people who do decide to participate will be different in important ways from those who do not participate. Typically, people who take the initiative to participate are economically better off, are better educated, and have more self-confidence. Consequently, it is difficult to know whether observed changes in income, reading skills, health, and so on are due to the effects of the project or to the differences in initial conditions and capabilities of participants and nonparticipants.
- It is very difficult to find a comparison group closely matching the experimental group on the key indicators. Project communities are often selected because of special characteristics. In some cases, project planners choose the poorest communities; in other cases, they choose communities that have the greatest likelihood of success. In either case, it will be difficult to find a comparison group that closely matches the project population.
- In many cases, it is difficult to use any kind of comparison group at all for political or ethical reasons. Frequently, politicians and community leaders in a designated comparison group area will pressure for their community to be included in the project. From the ethical perspective, one would not want to withhold a service (such as oral rehydration treatment for severely malnourished children) just to prove the efficacy of the treatment. Another ethical (or practical) consideration is that it is often considered inappropriate to ask families in comparison communities to spend a long time responding to surveys if they will not receive any benefit. In some cases, the fact that families are being interviewed creates false expectations that they will be eligible to participate in this or a later phase of the project.

BOX A11.4-1

Potential Methodological Weaknesses in Many Statistically Strong Evaluation Designs

Many evaluation designs that are commonly referred to in the evaluation literature as experimental design are in fact only strong with respect to their ability to control for sources of statistical selection bias. The reasons that they are statistically strong (e.g., randomization or statistical matching of samples, strict and inflexible rules concerning how data are collected, and administration of the same survey instrument to the same or equivalent samples before and after the project implementation) makes these quantitative designs potentially weak in other respects, including

- *Weak construct validity:* Many statistical designs are not based on a program theory model

(although it is perfectly possible to incorporate a program theory model).

- *No analysis of the project implementation process:* The pretest–posttest evaluation designs do not study how a project is implemented, and consequently they are not able to assess the extent to which a failure to achieve intended outcomes is due to design failures compared to weaknesses in how the project is actually implemented.
- *No consideration of contextual variables that can explain differences in outcomes in different project locations:* Programs with identical designs and implementation procedures may have significantly different outcomes due to the influence of contextual factors such as local politics, local economies, and factors such as migration or rainfall patterns.
- *Mono-method bias:* Many quantitative designs collect all of their data from a single instrument, most commonly a structured questionnaire. This increases the risk of bias or incomplete information, as it is not possible to compare estimates obtained from different independent sources.
- *Difficulties in collecting information on sensitive topics:* Many quantitative data-collection methods use structured questionnaires and a sampling protocol that requires a predetermined sample frame.
- *Inflexibility and difficulty in adapting the design to changes in the project design or the context in which it is implemented:* Probably the majority of evaluation designs are based on the assumption of a relatively static project design and context within which it is implemented, and they do not address the challenges of "emergence." Over the life of a project there are frequently significant changes in the external environment, the organizational and administrative arrangements, the nature and responses of the target population, and how the project is implemented (and even its objectives). Many evaluations do not take these changes into account, and the program is evaluated as if it had remained unchanged over its lifetime.

Two conclusions result from these potential methodological weaknesses in strong statistical designs. First, it is important when discussing the merits of different evaluation designs to always distinguish between "strong statistical designs" and "methodologically strong designs." While statistical evaluations can be designed to ensure all-round methodological strength, this is frequently not done, leaving many statistically strong designs to be vulnerable in other ways. Similarly, it is possible to have qualitative or mixed-method designs that might be considered weak in terms of conventional quantitative terms but that may use designs that are methodologically sound in other ways. Second, it is almost always possible to strengthen the methodology of all evaluation designs—quantitative, qualitative, and mixed method—by incorporating the "tools and techniques to strengthen any of the basic impact evaluation designs" discussed in this appendix (see Table 11.4 in Chapter 11).

- It is also difficult to ensure that treatments (services) are administered in exactly the same way to all project sites and families. Sometimes the delivery of materials and equipment is delayed; in other cases, there are major differences with respect to the organization of the project and delivery of services in different sites. In one microcredit program, the local administrator may speak the local language and may create a welcoming atmosphere that encourages families to visit the project to discuss loans. In another site, the administrator may not speak the local language and the project may be seen as hostile to the community, causing fewer people to visit the center. For all these reasons, it is difficult to determine whether differences in project performance are due to differences in the responsiveness of different communities or whether the differences are due to the way the project was administered in different sites.
- Finally, each project operates within a unique economic and political context and must interact with a number of government or nongovernmental organizations (NGOs), each of which has its own particular characteristics. Also, the social, economic, and cultural characteristics of the target population may vary significantly among project sites. All these contextual factors can have an important influence on the outcome of the project. Consequently, even when a project is administered in exactly the same way in each site, there may be significant differences in the outcomes as a result of these contextual factors.

APPENDIX 11.5: EXAMPLES OF RANDOMIZED CONTROL TRIALS

BOX 11.4

Examples of Randomized Control Trials

*In June 2003 the Massachusetts Institute of Technology (MIT) launched a program of **randomized trials** to assess the impacts of development programs in developing countries and the United States. The main justification for doing this was the belief that nonrandomized, quasi-experimental evaluation designs can often come to erroneous conclusions about program effectiveness because of the "omitted variable" problem (important factors that might explain apparent program effects have been excluded from the analysis). Some of the randomized control trials reported by the MIT Abdul Latif Jameel Poverty Action Lab (J-PAL) include:*

- Women as policymakers—the impact of female political leaders on policy decisions
- The Balsakhi (Mumbai, India) program—charting the effects of remedial education programs on school quality and test scores
- Measuring the impacts of school inputs in Kenya—the case of flip charts (see Box 11.5)
- Primary school deworming project (Kenya)—the impacts of child health gains due to preschool health and nutrition projects on preschool participation
- Incentives to learn—the impacts of scholarships for girls in Kenya
- School choice in Colombia—measuring the impacts of vouchers for private schooling
- Peer effects, alcohol, and college roommates in the United States—the impact of randomly assigned college roommates on drinking behavior
- A study of racial discrimination in the job market in Chicago and Boston
- The Balwadi health program (New Delhi, India)—the impacts of child health gains due to a preschool health and nutrition project on preschool participation
- Interest rate and consumer credit in South Africa—the effect of changing interest rates on loan acceptance
- Understanding technology adoption—fertilizers in Western Kenya: Why do farmers not use fertilizer even though it appears to have the potential to increase yields considerably?

Source: Adapted by the authors from the MIT Poverty Action Lab Web site (www.povertyactionlab.org).

In 2009 the World Bank produced a volume of case studies on evaluations that influenced policy. The following are some examples in which randomized designs were used:

- The Progresa/Oportunidades conditional cash transfer program in Mexico used randomized selection of communities for each phase of the project. A pipeline design (see Appendix F) with subjects who would not receive benefits until the next phase was used as the comparison group for beneficiaries of the first phase.
- In Kenya an evaluation was designed to assess the effects of charging for insecticide-treated bednets for malaria prevention compared to providing them at no cost. Randomization was used to determine which clinics would provide bednets free and which would charge.
- In Kenya a deworming program for schools was to be introduced in three phases to accommodate financial and administrative constraints. Schools were randomly assigned to the three phases with the schools not yet phased-in acting as a control group.
- In Morocco a rural microfinance program developed a national RCT to evaluate the effects of microfinance on agricultural and nonagricultural activities, and on income, expenditures, and household security. The RCT was introduced in the second phase of the project to increase the credibility of the evaluation findings.

Source: Bamberger, M., and A. Kirk, eds. (2009), "Making Smart Policy: Using Impact Evaluation for Policy Making. Case Studies on Evaluations That Influenced Policy." Washington, DC: World Bank.

Note: For additional examples, see evaluations conducted by the International Initiative for Impact Evaluation (3ie) at www.3ieimpact.org

APPENDICES FOR CHAPTER 12 QUANTITATIVE EVALUATION METHODS

12.1 The Main Types of Questions That Can Be Included in Quantitative Surveys

12.2 Useful Sources of Secondary Data for QUANT Evaluations

12.3 Large-Scale Compilations of the Findings of Randomized Control Trials (RCTs)

12.4 Data Analysis for Quantitative Evaluations

Chapter 12 presents an overview of quantitative evaluation methods. It covers experimental and quasi-experimental designs, with an extensive discussion of the strengths and limitations of randomized control trials and other designs described as "strong" by their advocates; examples of applications of quantitative methods; data-collection methods; the management of data collection; and an extended discussion of a wide range of data analysis methods.

The four appendices review the main kinds of questions used in quantitative evaluations (Appendix 12.1), a guide to useful sources of secondary data that can be used to strengthen the analysis (Appendix 12.2), links to a number of large-scale compilations of randomized control studies (Appendix 12.3), and a detailed review of survey analysis methods for quantitative evaluations (Appendix 12.4).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 12.1 THE MAIN TYPES OF QUESTIONS THAT CAN BE INCLUDED IN QUANTITATIVE SURVEYS

The following are commonly used types of question (adapted from Gray, 2004, pp. 191–198). Some of the categories, such as open questions, are normally considered qualitative approaches, but questions of this type are also frequently included in QUANT surveys as probes or for clarification.

- *Classification questions* collect information on questions such as age, sex, education, whether attending school, type of housing, access to services, amount of land owned, ecological characteristics of the land and types of crops produced, and other relevant characteristics of individuals, communities, or groups. In household surveys, a *household roster* will often compile information on each household member. This information serves two main purposes for the evaluation. First, it is used to select people who are involved in different ways in the project being evaluated. For example, an irrigation project will have different effects depending on the amount and type of land and the crops produced, and evaluating the effects of different kinds of programs to increase school enrollment and reduce drop-outs will have different effects on children attending school and their regularity of attendance. Second, the information is used to select control variables to include in the statistical analysis (see Section 7 of Chapter 12).
- *Open questions* do not present a predefined menu or list of options. The responses should be recorded in full. Examples include "Why did you move to this community?" "How do you feel about . . . ?" "What do parents think about the new teaching programs introduced into the school this year?"[16]
- *Closed questions* can ask for Yes/No answers or can ask respondents to choose one answer from a multiple-choice menu.
- *Lists* allow respondents to select as many responses as they wish.
- *Category (ordinal) questions* are a variety of closed questions where numerical information is put into a series of categories. Instead of asking an open question such as "How much did you earn last month?" or "How frequently do you and your spouse go out together?" the question is asked in this form:
    - Several times a week
    - Once a week
    - Once a month
    - At least once a year
    - Never

Although category (ordinal) questions can simplify the analysis, they require careful field testing. If there is insufficient pre-testing, the wrong range of categories may be selected and everyone may select the same response category (e.g., all spouses go out together several times a week, or everyone falls into the lowest or highest income category), making the information of very little use. The range of categories can also influence the response. For example, if the question asks, "How many times did you and your husband have an argument during the past week?" some respondents will consider this includes minor arguments, such as which TV program to watch. However, if the question uses the same wording but includes "during the past year," many respondents will only report on major family disputes, perhaps mentioning only one or two.

[16] Although open-ended questions should technically be considered QUAL, a few such questions are often included in structured questionnaires (e.g., as follow-up in case of people responding "Other"), so they are mentioned.

- *Ranking questions* ask respondents to rank a set of options in order of their importance, seriousness, and the like. Questions may cover what they like or do not like about the local school or a community organization, the main causes of worker absenteeism, important features of a public transport system, and so on.
- *Scales* are designed to measure the degree or intensity of opinion or experience on a particular topic. The following example illustrates one of the many ways in which scales can be presented. Respondents are asked to indicate how strongly they agree or disagree with a statement such as "The community has become a much safer place since the police post was opened last year."
    - Agree strongly (or very satisfied)
    - Agree (or satisfied)
    - Neither agree nor disagree (or neither satisfied nor dissatisfied)
    - Disagree (or dissatisfied)
    - Strongly disagree (or very dissatisfied)

A common form of a scaled question asks the respondent to rate his or her judgment or opinion on something on an ordinal scale that typically has between 5 and 10 points. Some scales ask respondents to indicate how strongly they "agree" or "disagree," while others may range from, for example, *poor or not at all* (1) to *excellent or always* (5). Some scales may use simple computation of the average ratings (Fink, 2008, 2009), while others may use more complex forms of analysis, such as multidimensional scaling where different items are combined (e.g., Kane & Trochim, 2007, Chapters 4–6; Litwin, 2003; Spector, 1991). Wikipedia provides an excellent overview of widely used scales such as Lickert, Thurstone, and Guttman. Although scales are a useful way to measure how people feel about, for example, community organizations or public service agencies that affect their lives, the development and testing of scale items and the structure of the whole scale is a time-consuming and specialized skill. Consequently, when scales are constructed rapidly and items are not carefully selected or pretested, there is a danger that the results will be either meaningless or misleading. Even when scales are carefully designed, there is still the danger that they may provide distorted or misleading information that leads to invalid findings.

Most scales are based on the assumption (sometimes not explicitly stated) that attitudes are unidimensional. Respondents either like everything about the school or they don't like anything about it. They find that everyone in the government agency is helpful or that everyone is unhelpful. Unfortunately, life is rarely so simple. If, instead of using a scale, respondents had been asked an open question—"What do you like and dislike about the school?"—it would probably have been found that there are some aspects of the new programs that they like and other aspects that they do not like, some teachers and administrators that they like, and others that they do not.

Useful sources for a further discussion of ways to ask questions include the following: Fink (2009), *How to Conduct Surveys*; Fowler and Cosenza (2009), "Design and Evaluation of Survey Questions"; Presser et al. (2004a, 2004b), *Methods for Testing and Evaluating Survey Questionnaires* and "Methods for Testing and Evaluating Survey Questions," respectively; and Sudman and Bradburn (1982), *Asking Questions* (still worth reading). For international development evaluations, particularly those focusing on the evaluation of poverty reduction programs, two of the most comprehensive sources are Grosh and Glewwe (2000), *Designing Household Survey Questionnaires for Developing Countries: Lessons From 15 Years of the Living Standards Measurement Study* (3 vols.); and Klugman (2002), *A Sourcebook for Poverty Reduction Strategies* (2 vols.).

APPENDIX 12.2 USEFUL SOURCES OF SECONDARY DATA FOR QUANT EVALUATIONS

Source	Types of Data	Comments
National household surveys	• Income, expenditure, and consumption data • Access to public services (education, health) • Educational enrollment and performance • Poverty • Household demographic characteristics	• In some countries, these have been conducted several times a year for a number of years. • The National Center for Education Statistics (NCES) in the United States has been surveying and making longitudinal data available for decades. • These normally use sound sampling techniques but may not cover all the informal sector population (which often represents an important part of the project population), or may not be disaggregated to the population targeted by a particular project.
Social sector ministries and departments (health, education, water, transport) in developed countries such as in North America and Europe, and in some developing countries	• Use of services (school attendance, use of health facilities) • Amount paid by users	• Some data sources are comprehensive and well designed, but the reliability and coverage of some studies, particularly in some developing countries, can vary. • Many surveys focus only on quantitative indicators such as access to services (e.g., water, education, and health) but do not include much information on the quality of these services.
Social service facilities (schools, health centers)	• Attendance and utilization rates • Common diseases and their incidence	• Data can be good and comprehensive, but quality varies greatly. • May be problems of under- or misreporting.
Bilateral and multilateral donor agencies (U.S. Agency for International Development, World Bank, U.S. foundations)	• Extensive information on the programs and geographical areas where they operate	• Donors have promoted some of the most comprehensive socioeconomic databases (examples include the World Bank's Living Standards Measurement Studies and USAID's Demographic and Health Surveys). • Data sometimes criticized for being too narrowly quantitative (e.g., the definition of poverty). • Samples often drawn for national statistics; not statistically significant for a smaller target population.

Source	Types of Data	Comments
Universities and research institutions	• In many countries, these are the technically best studies available • Often include both QUANT and QUAL approaches	• Although some university studies, particularly in the United States and Europe, may cover large populations, in developing countries many such studies cover only relatively small areas and samples. For example, many graduate dissertations contain valuable information on the topics of interest to an evaluation, but often they study only relatively small populations.
Government-, donor-, and foundation-supported programs and reports	• Detailed information on the characteristics of the target population, their access to the program, and program performance	• Many studies cover only program beneficiaries and do not include a comparison group. • In other cases, the target population is not clearly defined.
Nongovernment organizations (NGOs)	• In-depth information on populations covered by the agency	• Often cover only relatively small populations and may be more qualitative. • There may be questions on the representativity of the data, particularly for organizations conducting studies on a small budget.
Cooperatives and microfinance programs	• Information on the size and use of loans and repayment rates • Sometimes information on the socioeconomic characteristics of program participants	• Quality of the data is quite variable.
Geographical information systems (GIS)	• GIS identifies the physical location of, for example: • Public services, commercial establishments • Areas with particular characteristics such as high crime, traffic accidents, information mortality Infrastructure (roads, waterpipes, etc.); the information is provided in the form of electronic maps that permit different kinds of information to be overlaid	• Extensive GIS data, much of it free, is now available in the United States, and it is starting to become available in many developing countries. • Mobile phones with GPS capacity now provide a cost-effective way to develop GIS maps in developing countries.

APPENDIX 12.3 LARGE-SCALE COMPILATIONS OF THE FINDINGS OF RANDOMIZED CONTROL TRIALS

The Cochrane Collaboration is an international, independent, not-for-profit organization of over 28,000 contributors from more than 100 countries, dedicated to making up-to-date, accurate information about the effects of health care readily available worldwide. Contributors work together to produce systematic reviews of health care interventions, known as Cochrane Reviews, which are published online in the Cochrane Library. Cochrane Reviews are intended to help providers, practitioners, and patients make informed decisions about health care, and are the most comprehensive, reliable, and relevant source of evidence on which to base these decisions. The reviews seek to achieve the highest possible level of evidence-based findings, in most cases using randomized control trials (www.thecochranelibrary.com/view/0/index.html).

The Campbell Collaboration (C2) helps people make well-informed decisions by preparing, maintaining, and disseminating systematic reviews in education, crime and justice, and social welfare. The Campbell Collaboration is an international research network that produces systematic reviews of the effects of social interventions. Again, most are based on randomized control trials (www.campbellcollaboration.org/systematic_reviews/index.php).

The Abdul Latif Jameel Poverty Action Lab (J-PAL) is a network of 44 affiliated professors around the world who are united by their use of randomized evaluations (Res) to answer questions critical to poverty alleviation. There are more than 170 evaluations that either have been completed or are ongoing (www.povertyactionlab.org).

The International Initiative for Impact Evaluation (3ie) promotes evidence-informed equitable, inclusive, and sustainable development. It supports the generation and effective use of high-quality evidence to inform decision making and improve the lives of people living in poverty in low- and middle-income countries. It provides guidance and support to produce, synthesize, and quality assure evidence of what works, for whom, how, why, and at what cost. Since its founding in 2008, 3ie has awarded over 300 grants (243 impact evaluations, 38 systematic reviews, and 23 other studies) in over 50 countries, with a total value of US$104,328,729. https://www.3ieimpact.org/.

The World Bank Poverty website includes various catalogues of impact evaluations. The following is one source where most but not all of the evaluations use RCTs: http://econ.worldbank.org/external/default/main?pagePK=64166018&piPK=64167664&menuPK=477165&theSitePK=469372&docTY=620265&colTitle=impact%2520evaluation%2520series.

APPENDIX 12.4 DATA ANALYSIS FOR QUANTITATIVE EVALUATIONS

1. Creating a Data Analysis Plan

An important part of data analysis is the design of a data management plan spelling out the objectives of the analysis, the key questions to be addressed, and the hypotheses to be tested. The plan should refer to the scoping phase during which the client's information needs were defined. The analysis plan is particularly important for RWE to ensure that the limited resources and time are focused on the critical issues and questions of concern to clients. The data analysis plan involves the following stages:

- Drafting an analysis plan (see Table A12.4-1)
- *Developing and testing the codebook*. If there are open-ended questions, the responses must be reviewed in the preliminary stage of the analysis to define the categories that will be used in the final analysis. 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 and to ensure responses are distributed across all categories and not just concentrated in one or two.
- *Ensuring reliable coding*. This involves both ensuring that the codebook is comprehensive and logically consistent, and monitoring the data coding process to ensure accuracy and consistency between coders.
- *Reviewing surveys for missing data and deciding how to treat missing data.* 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 in the analysis. For example, there may be differences between sexes, age, economic status, or levels of education of respondents 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 to determine the frequency of missing data for key variables. For variables with significant levels of missing data, an exploratory analysis should be conducted 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
  - Deciding how to treat missing data and the application of missing data policies
  - Identifying any variables that may require recoding
- Full documentation of how data were cleaned, how missing data were treated, and how any indices were created.

TABLE A12.4-1 ● Example of an Analysis Plan for an Evaluation of the Impacts of Microcredit on Female Borrowers

Evaluation Objective 1: To assess the impacts of the program on women's earned income
Hypothesis: Women who participate in the program will have higher earned income than those who do not.

(Continued)

(Continued)

Variables: These might include women who have received loans and women who have not, earned income, age, education, and prior experience in running a business.
Analysis Stage 1: Comparing the mean earned income of women who have and have not received loans through the project (t-test for difference of means).
Analysis Stage 2: Multiple regression analysis testing whether there is a difference in earned income for participants and nonparticipants after controlling for age, education, and prior experience in running a business.
Evaluation Objective 2: To assess the impact of the program on women's feeling of personal empowerment
Hypothesis: Women who have participated in the program will have a stronger feeling of personal empowerment than women who have not participated.
Variables: Women who have participated in the program and those who have not (Note: Participation will be defined both as dichotomous Yes/No variables and also in terms of the number of different services received—loans, training courses, technical support, group meetings, etc.), scale of personal empowerment.
Analysis 1: Two-way table comparing participation/nonparticipation with the score on a 5-point empowerment scale. Chi-square or similar contingency tests will be used.
Analysis 2: Two-way table comparing two ordinal variables: the score on the 5-point empowerment scale and the number of services (between 1 and 5) received from the program. A contingency test for comparing two ordinal variables (e.g., Goodman and Kruksal's Gamma)[17] will be used.

Source: Adapted from Fink (2003c, ex. 1.1).

2. Descriptive Data Analysis

Descriptive data analysis describes important characteristics of the populations studied through measures of central tendency—means, modes, and medians—or the distribution (spread) of the data. The purpose is to obtain an initial understanding of the characteristics of the population studied and to identify similarities and differences among different sectors of the population. These kinds of analysis are almost always conducted before planning more detailed analysis. The types of analysis to be conducted and the statistics to be used will depend on what kind of variable is appropriate:

- *Nominal variables.* For these types of variables, the frequencies of each category can be counted, but the categories do not have any numerical order (i.e., one category is not greater or lesser than another on a scale). Examples include economic sectors in which persons work, regions of birth, reasons for migrating to a city, and favorite subjects in school. While the distribution of responses among the different categories can be described, it is not possible with a nominal variable to calculate, for example, a mean or average.
- *Ordinal variables.* The values of these variables have an inherent order and can be ranked from lesser to greater. For example, relative satisfaction with local schools or health facilities can be ranked by asking respondents to indicate their satisfaction levels on a Likert scale (i.e., one that offers response choices such as *strongly agree*,

[17] The Goodman and Kruksal Gamma is an example of a nonparametric statistical test, which is used to test the correlation between two variables, such as rankings, where the variables are not interval. The test compares the number of pairs that are "concordant" (both have the same ranking) with the number that are "discordant" (the number of pairs in which the rankings are different). See Sirkin (1999, pp. 358–362).

agree, *disagree*, or *strongly disagree*). However, because the intervals between the different categories (e.g., *strongly agree* and *agree* compared with *disagree* and *strongly disagree*) cannot be assumed to be equal, it is not possible to calculate the means and standard deviations (see below). Sometimes, to simplify the analysis, interval variables such as income or age may be transformed into ordinal variables by creating categories such as "under 5 years of age," "5–10 years," "11–20 years," or "over 20 years." This reclassification results in a considerable loss of data but may be justified due to budget and time constraints and in order to make the findings easier to understand for readers with no background in statistics.

- *Interval (numerical) variables.* These are variables such as weight, age, income, time traveling to work, and number of children that can be measured on a scale where the distance between each category is equal. Numerical ordering from largest, most frequent, or longest, for example, to smallest, rarest, or shortest is possible—and the distances between each two numbers is the same on an arithmetic number line. With interval variables, a much wider range of statistical indicators and tests can be used (e.g., mean, standard deviation, statistical significance tests).

The analysis will often begin by presenting measures of central tendency and distribution and will then compare these values for different groups to identify similarities and differences. Let us take the example of household income. The analysis might begin by presenting one or more of the following indicators of central tendency:

- *The mean (average) income of all households.* For example, the mean household income may be 350 pesos per month.
- *A frequency distribution in which income is classified into groups.* The preliminary frequency distribution would give the frequency of each value, for example: "less than 50 pesos," "51–100 pesos," "101–150 pesos," and so on, and the number of families in each category is shown in a table.
- *The mode*. This is the category with the highest frequency, for example: "150–200 pesos."
- *The median*. Assume there were 150 interviews. If these are arranged in a frequency distribution from lowest to highest, the median is the 75th value, for example, 175 pesos. In most distributions, the mode and the median will be fairly close to each other. However, there are some distributions where they can be quite different. For example, a bimodal distribution may have many values concentrated at the lower end of the distribution (many poor households) and many near the top of the distribution (relatively wealthy households) and relatively few in the middle. In this case, the mode (or modes) would be quite different from the median.

The next stage will often be the analysis of dispersion—whether values are similar for most subjects (e.g., most families have similar incomes) or widely dispersed (e.g., some families have very low incomes and others have much higher incomes). The following are indicators of dispersion:

- *Range.* This is the difference between the highest and lowest value. For example, the lowest income may be 75 pesos and the highest may be 1,025 pesos (range = 950).
- *Standard deviation.* This is based on the average difference between each value and the mean. This average is divided by the mean, and the square root is calculated. One of the great advantages of the standard deviation for many kinds of statistical analysis (such as the statistical significance tests discussed in Chapter 15) is that approximately 65% of the scores in any approximately normal population will be within one standard deviation of the mean and 95% will fall within two standard deviations. In our earlier example, the mean income was 350. If the standard deviation was 25, then we would expect that approximately two-thirds of families (65%) would have incomes between 325 pesos (one standard deviation below the mean) and 375 pesos (one standard deviation above the mean). Similarly, around 95% of families would have incomes between 300 and 400 pesos.
- *Standardized Z score.* The standard deviation can be transformed into a standardized (Z) score by subtracting the mean and dividing by the standard deviation, so that the value of the standard deviation can be compared for populations with different means.[18]

[18] Statistical significance tests and the calculation of the standard deviation are discussed in Chapter 15.

3. Comparisons and Relationships Between Groups

Once the characteristics of the population have been studied, the next stage will usually be to examine similarities and differences between groups on the variables of interest to the evaluation. For example, boys and girls might be compared on school enrollment rates or school test scores, or fishermen and farmers may be compared on income.

The simplest comparisons involve two-way tables. Table A12.4-2 shows a hypothetical two-way table comparing the frequency with which men and women attend community meetings. From the table, it appears that women attend meetings more frequently than men: 71.1% of women attend either once a week or at least once a month compared with only 27.3% of men. However, when the number of observations is small, a large-percentage difference may not be statistically significant. As was discussed in Chapter 15, a statistical significance test calculates the probability that the difference between the two groups (71.1% and 27.3%) could occur by chance if the two groups actually came from the same population. This is often expressed by saying there is a statistically significant difference between the two groups "beyond a reasonable doubt." A number of statistical tests (such as Chi-square and the t-test) are available to assess the statistical significance of differences. It is important to ensure that the correct statistical test is used to avoid incorrect conclusions about statistically significant differences between groups. In our discussion of threats to validity in Chapter 7, the incorrect application of statistical tests was given as one of the main threats to statistical conclusion validity. It is always a good idea to consult with a statistical specialist if there is any doubt.

Comparisons between more than two variables may use more sophisticated statistical tests of association such as analysis of variance (ANOVA), simple and multiple correlation, and multiple regression.

4. Statistical Procedures for Assessing Program Effects

Table A12.4-6 at the end of this appendix lists some of the most common statistical procedures for testing the different kinds of hypotheses discussed in the following sections.

4.1 The Logic of Hypothesis Testing

Most QUANT evaluations involve the testing of hypotheses to determine whether the predicted or desired program effects have been achieved and whether the magnitude of the observed change (effect size) shows beyond a reasonable doubt that

TABLE A12.4-2 ● Example of a Tallying Chart of the Frequency With Which Men and Women Attend Community Meetings

Frequency of Attending Community Meetings	Men		Women		All Adults	
	Number	% All Men	Number	% All Women	Number	% All Adults
1. Every week	25	11.3	60	38.8	85	22.7
2. At least once a month	35	16.0	50	32.3	85	22.7
3. Several times a year	80	36.3	25	16.1	105	28.0
4. Once a year	70	31.9	10	6.4	80	21.3
5. Never	10	4.5	10	6.4	20	5.3
Total	**220**	**100**	**155**	**100**	**375**	**100**

the program intervention is associated with this outcome (effect). In some cases, the hypotheses to be tested are derived from a program theory model (see Chapter 10). In other cases, the *null hypothesis* (often represented as H_0)—that there is no difference between project and comparison groups—is tested. The reason for using a null hypothesis is that it is never possible to prove that a program has produced a certain effect. What a statistical significance test does is to indicate the probability that the observed difference between the project and comparison groups could have occurred if the project participants and comparison groups are drawn from the same population. For example, let us assume that the study finds that the average household income of farmers who have used the new seed varieties is 8% higher than the income of farmers who have not used the new seeds, and let us also assume that the analysis finds that there is only a 4 in 100 (4%) chance of a difference as large as 8% occurring if there really is no difference between the two groups. The conventional practice is to assume that if the probability is less than 5 in 100 (5%), then there is a statistically significant difference between the two groups. In situations in which it is important to avoid wrongly assuming that treatments are effective (e.g., the testing of new drugs), a higher level of precision (e.g., 1 in 100 or 1 in 1,000) may be used. These issues are discussed further in Chapter 15 (Section 5). For readers interested in statistical analysis, Table A12.4-6 summarizes some of the common statistical procedures for testing different types of hypotheses and models with nominal, ordinal, and interval variables.

The comparison to be tested is most commonly between samples selected from the project and comparison groups. In this case the null hypothesis would state that there is no significant difference between the project group and the comparison group with respect to the outcome measure (aptitude test score, household income, proportion of girls attending secondary schools, etc.). The null hypothesis (H_0) is specified as follows:

$$H_0: x_0 = x_1$$

where

x_0 = mean or other outcome measure for the total population or for the comparison group

x_1 = mean or other outcome measure for the project population

Chapter 15 discusses the use and interpretation of significance tests and the concept of Type I (false positive) and Type II (false negative) errors in the interpretation of significance tests.

5. Tests Involving the Comparison of Two Means (the t-test)

Most statistical analyses of project impacts compare the mean value of an outcome indicator for the project group with the corresponding mean value for the comparison group. If there is found to be a statistically significant difference between the two means, this will be considered evidence of a potential project effect. The comparison group could be

- *T-test for a single sample:* The total population of interest (for example, all children attending secondary school in a particular region)
- *T-test for dependent means:* A comparison of the mean score for the project group at two points in time (usually before and after the project treatment has been implemented)
- *T-test for independent means:* A comparison group selected to be representative of the total population of interest (e.g., a sample selected from all high school students not selected for the project)
- *Double-difference analysis:* A comparison of the mean score of the project group before and after the project implementation with the mean score for the comparison group before and after the project. The analysis actually compares two means by calculating the *change score* for the project and comparison groups and then applying the t-test to compare the means of the two change scores.
- *T-test for independent means:* A comparison group that represents an alternative treatment with which the project is being compared (e.g., the project may be introducing innovative teaching methods while another program may reduce class size)

The statistical significance of the difference between the mean for the project group and each of these different kinds of comparison groups can be tested by using a t-test. While the exact application of the t-test will vary slightly for each of these comparisons (see Aron & Aron, 2002, Chapters 8 and 9), the basic logic of the test and the stages of the analysis are similar in each case. We assume in the following example that the t-test is being used to compare two independent samples, such as the project group and an independently selected comparison group, so that the form of the test we use is for the comparison of independent means. The key steps are the following:

Step 1. Formulate the research hypothesis. For example: "Scores on the end-of-year math test for sixth graders will be higher for the project group than for the comparison group." When formulating the hypothesis, it is essential to decide whether the research question is to determine *whether there has been an increase in test scores* for the project group or whether the purpose is to test *whether there is a difference between the two groups*. In this latter case, the test must be able to determine whether the mean score for the project group is either significantly higher or significantly lower than for the comparison group. As we will see in the following step, this decision will determine whether a one- or a two-tailed t-test will be required.

Step 2. Determine whether to use a one- or two-tailed test of significance. While most evaluations are interested in measuring a particular direction of change (increased test scores, reduced incidence of waterborne disease), there are cases where the project could produce either a positive or a negative change. For example, an increase in school fees might reduce enrollment if poor families are unable to pay; on the other hand, it could increase enrollment if the fees were used to improve the quality of buildings and equipment or to hire more staff. If the research hypothesis only wishes to test whether change occurred in the predicted direction, a one-tailed test will be used, but if there is interest in studying the direction of change, then a two-tailed test will be used. Clarifying whether a one- or two-tailed test will be used is very important, as it will significantly affect the size of change that has to be produced for it to be found statistically significant. For example, using a one-tailed test, it might be found that a 5% greater increase in enrollment in the project schools might be found statistically significant, whereas for a two-tailed test it might be found that a larger difference would be required for it to be statistically significant.

Step 3. Reformulate the research hypothesis as a null hypothesis that can be tested. Decide whether this should be specified as a one- or two-tailed test.

Step 4: Interpret the t-score. Normally the t-score will be calculated using one of the many statistical packages such as SPSS or SAS. We assume for the purpose of this example that the estimated t-score is 1.75. When using the t-test to determine whether there is a statistically significant difference between the two independent sample means, use the following procedure:

- Determine the degrees of freedom (df). This is $N_1 - 1 + N_2 - 1$ where N_1 and N_2 represent the respective sample sizes for the project and comparison group samples. We will assume in this example that samples of 26 were used for both samples so that df = (26 − 1 + 26 − 1) = 50.
- Determine whether a one- or two-tailed test will be used. We will assume in this example that a one-tailed test is used.
- Define the significance level that will be used (0.01 where a high level of precision is required, 0.05 as the most generally accepted level, or 0.1 where a low level of significance is considered acceptable). We will assume that the 0.05 level is chosen.
- Define the cut-off point in the t-table for determining if there is a statistically significant difference between the two samples. This is determined by selecting the column for the one-tailed test and finding the row for df = 50. We then choose the column for the 0.05 significance level and find that the cut-off t-score is 1.67 for a one-tailed test with df = 50. In our example we have assumed that the t-score = 1.75, and as this is higher than the cut-off score of 1.67 there is a statistically significant difference at the 0.05 level between the project and comparison group.
- If the evaluation were using a two-tailed test the cut-off value for t would be 2.009 and our t-score of 1.75 would not be significant. This illustrates the importance of deciding whether a one- or two-tailed test should be used.

Applying the T-test for Other Types of Comparison Between Means

In the previous example we showed how the t-test would be applied to the comparison of independent means. However, the t-test can also be used to compare a single sample mean (normally the project mean) with the true population mean, to compare two dependent means, or for double-difference analysis. The t-test can be used to test differences between means for each of these types of comparison, and while the logic remains the same, there are certain differences in how the t-test is applied. Aron and Aron (2002, Chapters 8 and 9) explain and provide examples for each of these applications.

6. Comparisons Among Three or More Means (Analysis of Variance)

Sometimes an evaluation requires a comparison of means of three or more groups to determine whether there are differences among them with respect to the project outcome indicator. While it would be possible to conduct separate comparisons between each pair of means using the t-test, analysis of variance (ANOVA) has the advantage that it can compare all of the means at the same time to determine whether there are differences.

For example, low-income families selected to participate in a self-help housing project were all previously living in one of three types of low-income housing: slums, low-income tenement housing, and traditional Spanish-type housing with a large patio that had been converted into a multifamily dwelling with shared toilets and washing facilities. The housing authority wants to know whether there are any differences among families coming from these three types of settlement with respect to the average amount they invest in the construction of their new homes. Analysis of variance is designed to address this type of question.

The basic logic of ANOVA is to define a null hypothesis stating that there are no significant differences between the means of each population. The hypothesis is tested by comparing the variance of the total population with the variance within each group. The population variance is normally not known, so it is calculated by comparing two estimates: the within-group variance and the between-group variance.

If there are no significant differences between the group means, then there will be no significant difference between the estimates of the population variance obtained from the within-group and between-group estimates. If the research hypothesis is true, then the *within-group* variance should be greater than the *between-group* variance. However, if the null hypothesis is true, then the size of the within- and between-group variances should be similar (but never exactly the same, as they are both estimates). The significance of the difference between the two variances is calculated through the F ratio, which is defined as the within-group variances divided by the between-group variance.

The following example illustrates how to find the appropriate F in the F distribution table. Assume that three groups are being compared and that there are five subjects in each group (total 15 subjects).

- The *numerator* column in the table is the number of groups minus 1. In this example, 3 – 1 = 2.
- The *denominator* column in the table is the total number of cases minus the total number of groups. In this example, 15 – 3 = 12.
- In this example, we select the 0.05 significance level.
- Table A12.4-3 illustrates the rows of the F distribution table for denominator = 11 and 12. Select the row for denominator degrees of freedom = 12 and significance level = 0.05, then find the column for numerator degrees of freedom = 2. The critical F value = 3.89. So if the F ratio is greater than 3.89, there is a statistically significant difference between the means of the three groups.

TABLE A12.4-3 ⬢ An Extract From the F Distribution Illustrating for Denominator Degrees of Freedom From 10 to 13

Denominator Degrees of Freedom	Significance Level	Numerator Degrees of Freedom					
		1	2	3	4	5	6
11	0.01	9.65	7.21	6.22	5.67	5.32	5.07
	0.05	4.85	3.98	3.59	3.36	3.20	3.10
	0.10	3.23	2.86	2.66	2.54	2.45	2.39
12	0.01	9.33	6.93	5.95	5.41	5.07	4.82
	0.05	4.75	**3.89**	3.49	3.26	3.11	3.00
	0.10	3.18	2.81	2.61	2.48	2.40	2.33

Source: Salkind (2008, pp. 336–338, Table B-3).

7. Analysis of Cross-Tabulations for Interval, Ordinal, and Nominal Variables (Chi-Square Tests)

The tests used in the previous sections are used for the comparison of means. However, in some evaluations it is necessary to compare differences between groups on *categorical* (also called *nominal*) outcomes such as the type of housing that people currently living in different kinds of urban settlements select in a low-cost housing project; the different ways that children from different types of family structure (female-headed household, male-headed household, single-parent household, etc.) respond to a school feeding program (they eat the meals at school as was intended by the project, they try to take the food home to share with their siblings, or they bring their siblings to school and try to ensure that they also get fed). In these cases, the Chi-square test is widely used. Chi-square tests can in fact be used for any kind of cross-tabulations, including for interval and ordinal data.

The Chi-square test compares the expected distribution of a particular group among the different outcome categories if there was no association between a particular group and the outcomes, with the observed frequencies of outcomes. It then tests for the "goodness of fit" and how closely the actual and expected frequencies match. The Chi-square table shows the cut-off points for each significance level and a given number of degrees of freedom above which there is a statistically significant difference between how each group is distributed among the outcome categories.

In the following example, the sex of household head is compared with three types of response to students with respect to school breakfasts: they eat the breakfast themselves, they try to take it home to share with siblings, or they bring their siblings to school and try to get them breakfast even though they are not enrolled in school. The research question is whether the sex of household head affects the response of students to school breakfasts. The underlying assumption is that households headed by women are on average likely to be poorer, with children knowing from experience the importance of sharing food; therefore, students from female-headed households may be more likely to try to share their school breakfast with their siblings. The Chi-square test assesses the "goodness of fit" of the results to the null hypothesis that there is no

association between sex of household head and attitudes to school breakfast (see Table A12.4-4). It does not directly test the research hypothesis but only determines whether there are differences among the groups.

The Chi-square test involves the following steps:

- Step 1: Determine the observed frequencies in each group.
- Step 2: Determine the expected frequencies in each group if there was no association between groups and outcomes categories.
- Step 3: Compute for each group observed minus expected categories.
- Step 4: Square the differences and divide by the expected group frequency.
- Step 5: Sum the results for each group.
- Step 6: Refer to the Chi-square table to determine the cut-off point for rejecting the null hypothesis for the given degrees of freedom and required significance level.

In this example the value of Chi-square = approximately 62. The degrees of freedom = (no. of columns – 1) × (no. of rows – 1). In this example, df = 2[(3 – 1) × (2 – 1)]. Table A12.4-5 shows a section of a table of cut-off scores for df = 1, 2, and 3. In our example, df = 2 and for the 0.05 significance level the cut-off score is 5.99. The calculated Chi-square score of 62 is clearly much higher than the cut-off, so we can conclude there is a statistically significant difference between the groups with respect to their attitudes to sharing school breakfasts.

TABLE A12.4-4 ⬢ Hypothetical Illustration of a Cross-Tabulation That Could Be Analyzed Using Chi-Square

Sex of Household Head	Student Attitude to Sharing Breakfast With Siblings			Total
	Students eat breakfast	Students try to take food home to share with siblings	Students bring siblings to school and try to get them breakfast	
Female-headed households	20	25	35	80
Male-headed households	80	15	25	120
Total	100	40	60	200

TABLE A12.4-5 ⬢ Section of Table of Cut-Off Scores for the Chi-Square Test

df	Significance Level		
	.01	.05	.10
1	6.64	3.84	2.71
2	9.21	5.99	4.60
3	11.34	7.81	6.25

Source: Sirken (1999, p. 547, app. 4).

8. Controlling for the Effects of Independent Variables That Might Affect the Outcomes Being Studied (Uses of Multiple Regression in Program Evaluation)

The kinds of statistical analysis that we have been discussing so far test for differences between the means of two groups or differences between the distribution of outcomes in different groups. They indicate the probability of finding a particular t-test, F-test, or Chi-square score of all groups that come from the same population. One of the limitations of these tests for the purposes of program evaluation is that they are not able to determine the extent to which these differences are due to the effects of the project intervention or to characteristics of the two groups that might affect outcomes. For example, assume that an evaluation is being conducted of the social and economic effects of providing low-income families with better housing. A pretest–posttest comparison is made of the change in household income of project participants and a matched control group over the first three years of the project. Assume that a t-test for independent means compared the change scores for the project and comparison groups (double-difference analysis) and found that there had been a significantly greater increase in household income for the project than for the comparison group. Based solely on this analysis, we only know that the rate of increase was higher for the project group, but we do not know whether the difference was due solely to the project intervention or whether project households had certain attributes not shared by the comparison group that might have made it more likely that their incomes would have increased more rapidly even if they had not moved to the project. For example, the project group might have had a higher education level, greater household assets, or fewer dependent children (and therefore fewer health and education expenses), all of which might have increased their ability to increase their income.

Questions such as these can be addressed through multiple regression. When the evaluation survey includes information on household characteristics such as education, number of children, income, employment, and household assets (and other characteristics relevant to a particular survey), multiple regression is able to statistically match households in the project and comparison group so as to determine whether differences between the two groups with respect to the project outcomes still exist after controlling for the effect of these households' characteristics. Sometimes the analysis will show that the rate of increase in income is closely associated with education, household assets, or types of employment, and that when the effects of these characteristics are accounted for the differences in outcome indicators are significantly reduced. In other words, much of the difference between the project and comparison group is due to these characteristics and not to the project effect. In other cases it will be found that after controlling for these characteristics there is still a significant difference between the two groups with respect to project outcomes, suggesting that the project intervention seems to have been a major contributor to the observed differences in outcomes.

The following paragraphs indicate some of the ways that multiple regression can be used to reduce sample bias by strengthening the match of the project and comparison group samples.

9. Strengthening the Matching of the Project and Comparison Groups at the Stage of Sample Design (Propensity Score Matching and Instrumental Variables)

Propensity Score Matching

Sometimes the evaluation will be fortunate enough to have access to sample survey data that had already been collected by another agency around the time that the project had begun and that covered the project and comparison group population, used a relatively large sample, included the key information of interest to the evaluation, and interviewed the right people. Examples of such surveys include household income and expenditures; living standard measurement surveys that are conducted periodically in a number of countries; or more specialized surveys or censuses that are conducted by government or international agencies on topics such as health and nutrition, education, or employment.

A logistical regression is run to determine the household characteristics that predict participation in the project. A sample of project households is then selected and each household is then matched with a set of (usually around five) "nearest neighbors" that closely match a particular household with respect to the probability of participation but who are not in the project. So each project household in the sample is matched by its own set of nearest neighbors.

Data are obtained from the secondary data set of the score of each household at the start of the project on the outcome variable being tested (for example, household income). Both project and comparison households are re-interviewed at the end of the project and a *change score* is then computed for each set of households by calculating the change for the project household and then subtracting from this the average change for the set of nearest neighbors. The average change score is then computed for the total sample as the mean of the change scores for each set of households. A statistical significance test is then conducted to determine if the average change score is sufficiently large to reject the null hypothesis.

Instead of using secondary data for the baseline survey it is also possible to conduct a sample survey covering a random sample of project and comparison group households at the start of the project and to use this sample to conduct the propensity score matching.

Instrumental Variables

This is a regression technique to control for sources of bias in estimating outcomes due to factors affecting program participation. When these biases are not addressed the outcome estimates are based on intended participation rather than actual participation. The instrumental variable (IV) approach identifies a variable that is correlated with program participation but not with program outcomes for subjects once they are in the program. The IV approach, by adjusting for factors affecting the likelihood of participation, improves the validity of the analysis of determinants of project outcomes.

Regression Discontinuity

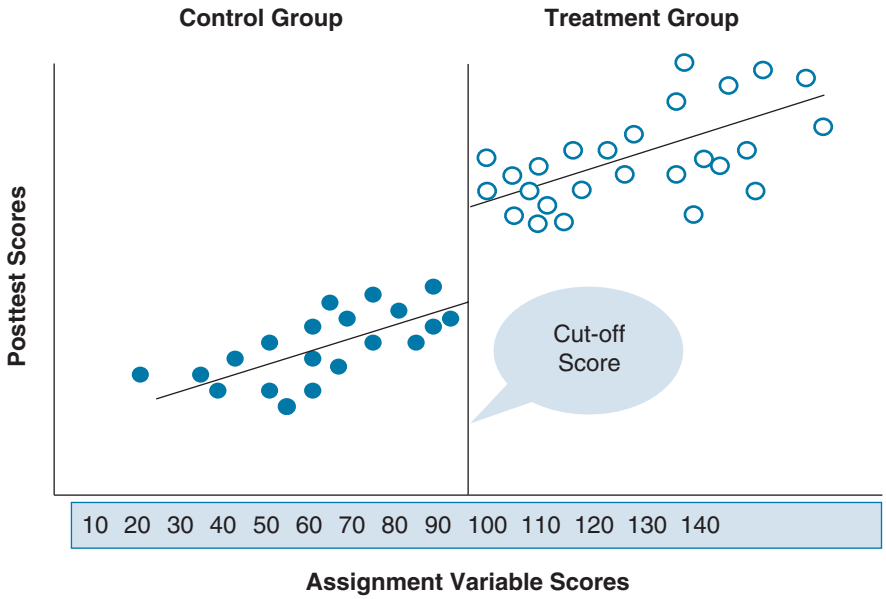
This is a form of multiple regression that provides an unbiased estimate of project impact. The approach requires the definition of an *assignment variable* that is used to determine a cut-off point for determining whether subjects are assigned to the project or the control group. The assignment variable must be either an interval or an ordinal scale. It can be an interval variable such as size of farm, education of the household head, or hours of vocational training received while in jail. It can also be a scale constructed by experts or program managers that defines either the need to participate in the program or the likelihood of success. Examples of such constructed scales, all of which must be ordinal, are scores on a scale of clinical depression based on ratings by psychiatrists, likelihood of success in developing a small business, or a scale of economic and social vulnerability that combines various dimensions. A cut-off point is defined on the scale, and all subjects above the cut-off point will be selected for the project and all below will form part of the control group.

The evaluation design compares subjects just above the cut-off point (who are in the project group) with those just below (who are in the control group). The logic is that by comparing groups just below and just above the cut-off point the two groups are likely to be similar to each other in all respects other than that one group participates in the project and the other does not. After the project treatment a multiple regression analysis is conducted and a regression trend line is computed. If the project had an effect, then the line will jump at the cut-off point. Figure A12.4-1 presents an example of a regression discontinuity analysis where there is a clear discontinuity ("jump") in the regression line. It is, of course, necessary to conduct a multiple regression analysis to determine whether the discontinuity is large enough to be statistically significant.

10. Using Multiple Regression During the Analysis of Project and Comparison Group Data

Often the data analysis will begin by using the t-test to determine whether there are statistically significant differences between the means of the project and comparison groups on key outcome indicators. Multiple regression can then be used to assess whether there is still a significant difference after controlling for key household characteristics. The following is an example of a multiple regression analysis to determine whether there are differences in the change in household incomes between the project and comparison groups over the first three years of the project. The analysis controls for the

FIGURE A12.4-1 ● Hypothetical Example of a Regression Discontinuity Design



Source: Prepared by the present authors.

years of education of the household head, the amount of land owned by the household at the start of the project (a proxy for wealth), and the number of dependent children:

$$Y = f(D_1, x_1, x_2, x_3)$$

where

- Y = the change in household income over the first three years of the project
- D_1 = dummy variable indicating if the household is in the project ($D_1 = 1$) or in the comparison group ($D_1 = 0$)
- x_1 = years of education of the household head
- x_2 = hectares of land owned by the household at the start of the project
- x_3 = number of children in the household under the age of 15

If the regression coefficient of D_1 is statistically significant this shows that there is a difference in the average change in household income over the first three years of the project between the project and comparison groups. If the coefficient is positive this shows that the project income has increased faster, while if it is negative this shows the comparison group income grew faster.

If a statistically significant t-test score was found but at the same time the coefficient of D in this regression equation was not significant, that would show that the apparent project effect revealed by the t-test was in fact due to the difference in household characteristics of the two groups and not to the project effect.

TABLE A12.4-6 ● Examples of Statistical Procedures for Testing the Main Types of Evaluation Hypotheses[19]

Test	Type of Variable	Type of Data	Reference
Contingency tables: Chi-square	Nominal	Two-way tables comparing the distribution of two nominal variables. *Example:* comparing male and female frequency of attending meetings	Frankfort-Nachmias & Leon-Guerrero (2011, Chapter 11); Salkind (2008, Chapter 16)
Contingency tables: for example, Goodman and Kruskal's Gamma	Ordinal	Two-way tables comparing the distribution of two ordinal variables. *Example:* comparing high, medium, and low levels of income with high, medium, and low levels of participation	Sirkin (1999, pp. 358–362)
T-test and Z-test	Interval	Comparing two means. *Example:* comparing the average income of two groups	Frankfort-Nachmias & Leon-Guerrero (2011, Chapter 9); Salkind (2008, Chapters 10–11); Moore & McCabe (1999, Chapter 8)
Analysis of variance	Interval	Comparing differences between three or more means. *Example:* comparing the average income of farmers, self-employed, and wage earners	Frankfort-Nachmias & Leon-Guerrero (2011, Chapter 14); Salkind (2008, Chapter 13)
Multiple regression	Interval sometimes including dichotomous or nominal (dummy) variables	Estimating the magnitude (proportion of the variance) and statistical significance of the association between two variables after controlling for intervening variables	Brief introduction: Aron & Aron (2002, Chapter 12); fuller discussion in Sirkin (1999, Chapters 13 and 14); R. Khandker et al. (2010, various chapters)
Econometric methods for impact evaluation	Nominal, ordinal, and interval	Covers randomization, propensity score matching, double-difference analysis, instrumental variables, regression discontinuity	R. Khandker et al. (2010)

[19] The reader is referred to the recommended readings at the end of Chapter 12, which provide both brief introductory texts and more advanced discussions of statistical analysis.

APPENDICES FOR CHAPTER 14
MIXED-METHOD EVALUATION

14.1 Two Case Studies Illustrating Different Ways in Which Mixed-Method Designs Can Strengthen Impact Evaluations

14.2 Characteristics of Quantitative and Qualitative Approaches to Different Stages of the Evaluation Process

14.3 Common Issues Affecting the Validity of Statistical Impact Evaluation Designs and How Mixed-Method (MM) Designs Can Help Address Them

14.4 Three Case Studies Illustrating the Use of Mixed-Method Evaluations

Chapter 14 reviews mixed-method designs that combine quantitative and qualitative evaluation concepts and methods to combine the strengths of both approaches while addressing their weaknesses. The rationale for the mixed-method approach is discussed. It is emphasized that most evaluations have a predominantly quantitative or predominantly qualitative approach, and that mixed methods have a different purpose for the two approaches. Mixed methods can also incorporate quantitative and qualitative components and can either be incorporated into the design sequentially (e.g., a quantitative survey followed by qualitative case studies) or in parallel (e.g., a participant observation study may be conducted at the same time as the quantitative survey). The chapter also shows how mixed methods can be used to tell a more compelling story about what a program has achieved, for example using Collaborative Outcome Reporting Technique (CORT). Finally, three case studies are presented to illustrate different ways that mixed methods can be incorporated in the evaluation of national level, sectoral, and program level interventions.

There are four annexes: Two case studies illustrating how mixed methods could be used to evaluate the same program but with quantitatively or qualitatively focused evaluations (Appendix 14.1); comparing how quantitative and qualitative evaluations typically approach each stage of the evaluation cycle (Appendix 14.2); common issues affecting the methodological rigor (validity) of impact evaluation designs (Appendix 14.3); and three case studies illustrating applications of mixed methods in the evaluation of national, sector, and program level interventions (Appendix 14.4).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 14.1 TWO CASE STUDIES ILLUSTRATING DIFFERENT WAYS IN WHICH MIXED-METHOD DESIGNS CAN STRENGTHEN IMPACT EVALUATIONS

This appendix summarizes two studies illustrating different ways in which unintended outcomes (UOs) were identified through MM designs. The references to the two studies are given. First, in their study of the effects of community-driven development on the levels of local conflict in Indonesia, Barron, Diprose, and Woolcock (2011) spent almost a year reviewing the academic literature on the causes and types of conflict, conducting participant observation in a sample of communities, interviewing local experts and key informants, and reviewing local documentation on the levels and types of conflict. They were able to identify more than 60 forms of community-level conflict that were then built into the evaluation design.

In the second example, the evaluation of a sexual health information campaign using SMS messages in Uganda included an RCT survey asking young men and women about their knowledge and attitudes toward the use of contraceptives and their risky sexual behaviors (Jamison, Karlan, & Raffler, 2013). It was found that risky sexual behavior decreased for both women and men. However, the follow-up, in-depth QUAL study found that as women became more aware of the dangers of risky sexual behavior, they refused to have sex with partners involved in high-risk behavior. An important UO was that men who were refused sex with their regular partners sought sex outside of their regular relationship, thereby significantly increasing their high-risk sexual behaviors. This important finding was only discovered by using an MM design.

APPENDIX 14.2 CHARACTERISTICS OF QUANTITATIVE AND QUALITATIVE 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 that usually generates 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.	• While 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 usually inductive (derived from information gathered during the course of the study).
Selection of subjects or units of analysis	• Random sampling so that findings can be generalized, and to permit statistical testing of differences between groups. • Requires a sampling frame that lists all of the units (houses, individuals, schools, etc.) of the population being 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 procedure varies according to the purpose of the study. • Purposive sampling is often used to collect the most useful and interesting data related to the purpose of the study. • While this is usually not done for QUAL evaluations, sometimes for mixed-method approaches the sample may be selected from 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 study (although the type of data collection and the subjects, groups, or organizations to be studied in the QUAL analysis will usually be different). • Usually a much smaller number of people are interviewed in more depth.

Evaluation Activity	Quantitative Approach	Qualitative Approach
Evaluation design	• Normally one of the quasi-experimental designs described in Chapter 11 will be used (although RCTs will often be used in the relatively small proportion of evaluations where this is possible). A randomly selected sample that represents the project participants, and where possible a comparison or control group, is interviewed at one or more clearly defined points in time during the life of the project. • Where possible, outcomes and impacts are estimated by comparing data collected before and after (and possibly during) the implementation of the project.	• The researcher(s) become immersed in the community, group, or organization over a relatively long period of time. • The effects of the program are studied through collecting information on many different aspects of the community or group and its economic, political, ecological, cultural, and psychological setting. • Normally the evaluation does not try to establish a direct cause-and-effect relationship between the project and the changes that are observed.
Data-collection and recording methods	• Data are usually recorded in structured questionnaires that are administered consistently throughout the study. There is extensive use of precoded, 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 precoded 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. 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 nonparticipant observation. • Study may use photography. • Several qualitative methods are used for multiple perspectives and triangulation.

(Continued)

(Continued)

Evaluation Activity	Quantitative Approach	Qualitative Approach
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 estimates from the evaluation survey.	• 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). • Triangulation of findings from different researchers, data collection, times and location of data collection, and methods of interpretation are a central element of the QUAL approach.

APPENDIX 14.3 COMMON ISSUES AFFECTING THE VALIDITY OF STATISTICAL IMPACT EVALUATION DESIGNS AND HOW MIXED-METHOD (MM) DESIGNS CAN HELP ADDRESS THEM

Issue	Potential Contribution of Mixed Methods
Evaluation Design Issues	
1. Limited construct validity. Many strong evaluations use secondary data sources and must rely on proxy variables that may not adequately capture what is being studied, so findings can be misleading.	• Exploratory qualitative studies can strengthen understanding of the key concepts being studied. • Focus groups and PRA can provide beneficiary perspective on concepts and constructs.
2. Decontextualizing the evaluation. Conventional IE designs ignore the effect of the local political, economic, institutional, sociocultural, historical, and natural environmental context. These factors will often mean that the same project will have different outcomes in different communities or local settings.	• Ethnographers, key informants, and other qualitative techniques can provide information on the local context. Contextual analysis can be incorporated into regression analysis through the creation of dummy variables.
3. Ignores the process of project implementation—the problem of the "black box." Most IEs use a pretest–posttest comparison and do not study how the project is actually implemented. If a project does not achieve its objectives it is not possible to determine if this is due to design failure or implementation failure.	• Qualitative techniques such as participant and nonparticipant observation and key informants can be combined with program monitoring to integrate quantitative and qualitative information on implementation and other project processes.
4. Designs are inflexible and cannot capture or adapt to changes in project design and implementation and in the local contexts. IEs repeat the application of the same data-collection instrument, asking the same questions and using the same definitions of inputs, outputs, outcomes, and impacts. It is very difficult for these designs to adapt to the changes, which frequently occur in the project setting or implementation policy.	• Panel studies, participant observation, key informants, and so on have the flexibility to detect and observe changes in the project or its setting.
5. Hard to assess the adequacy of the sampling frame. Evaluations frequently use the client list of a government agency as the sampling frame. This is easy and cheap to use, but frequently the evaluation ignores the fact that significant numbers of eligible families or communities are left out—and these are usually the poorest or most inaccessible.	• Small-scale, rapid studies of selected areas can be used to assess the adequacy of sampling frames.
6. No clear definition of the time frame over which outcomes and impacts can be measured. The posttest measurement is frequently administered at a time defined by administrative rather than theoretical considerations. Very often the measurement is made when it is too early for impacts to have been achieved, and it may be concluded that the project did not have an impact.	• Program theory models can be used to define the time frame over which outcomes and short-, medium-, and long-term impacts can be expected to occur. This can help define both when the impact evaluation should be conducted and also the initial indicators that a project is on track to achieving its outcomes/impacts.

(Continued)

(Continued)

Issue	Potential Contribution of Mixed Methods
Evaluation Design Issues	
7. Difficult to identify and measure unexpected outcomes. Structured surveys can also measure the expected outcomes and effects and are not able to detect unanticipated outcomes and impacts (positive and negative).	• Program theory models can identify preliminary indicators that can be measured early in the project and that provide evidence that the project is on track. • Qualitative methods such as key informants, participant observation, and focus groups can provide early indicators of whether the project is on track.
Data-Collection Issues	
8. Reliability and validity of indicators. Many statistical designs only use a limited number of indicators of outcomes and impacts, almost all of which are quantitative.	• MM can combine multiple quantitative and qualitative indicators that in combination can enhance validity and capture different dimensions of what is being studied. • MM makes extensive use of triangulation through which estimates obtained from different indicators are systematically compared and refined, and understanding is enhanced by comparing different perspectives.
9. Inability to identify and interview difficult-to-reach groups. Most QUANT data-collection methods are not well suited to identify and gain the confidence of sex workers, drug users, illegal immigrants, and other difficult-to-reach groups.	• Ethnographers and other qualitative researchers have extensive experience in reaching inaccessible groups.
10. Difficult to obtain valid information on sensitive topics. Structured surveys are not well suited to collect information on sensitive topics such as domestic violence, control of household resources, and corruption.	• Case studies, in-depth interviews, focus groups, and participant observation are some of the many qualitative techniques available to study sensitive topics.
11. Lack of attention to contextual clues. Survey enumerators are trained to record what the respondent says and not to look for clues such as household possessions, evidence of wealth, interaction among household members, or the evidence of power relations to validate what is said.	• Observation and key informants are two of the many useful techniques.
12. Often difficult to obtain a good comparison group match. Adequate secondary data for using propensity score matching are only infrequently available, and often control groups must be selected on the basis of judgment and usually very rapid visits to possible control areas.	• Judgmental comparison group selection can be strengthened through rapid diagnostic studies, consultations with key informants, and so on.
13. The vanishing control group. Control groups get integrated into the project, or they may be eradicated through migration, flooding, or urban renewal.	• Panel studies and observation techniques can monitor changes in the size and composition of the comparison group, can help explain the dynamic of the changes, and can provide early warning when corrective actions must be taken.

Issue	Potential Contribution of Mixed Methods
Data-Collection Issues	
14. Lack of adequate baseline data. A high proportion of evaluations are commissioned late in the project and do not have access to baseline data. Many IEs collect baseline data but usually only collect QUANT information.	• There are a range of qualitative techniques that can be used to help "reconstruct" baseline data.
Analysis and Utilization Issues	
15. Long delay in producing findings and recommendations that can be used by policymakers and other stakeholders. Conventional IEs do not produce a report or recommendations until the posttest survey has been completed late in the project cycle or when the project has ended. By the time the report is produced, it is often too late for the information to have any practical utility.	• Formative evaluation can provide periodic feedback to stakeholders throughout the life of a project. Some of this information can be generated by the planning and initial data-collection phases of the impact studies, building up a constituency for the later findings of the quantitative studies.
16. Difficult to generalize to other settings and populations. This is a particular challenge for RCTs and similar designs that estimate average effects by controlling for individual and local variations.	• Techniques such as quota sampling can use small samples to study variations in the population studied, providing a stronger basis for assessing the populations for which program replication is most and least likely to be successful.
17. Identifying and estimating influence of unobservables. Participants who are self-selected, or who are selected by an agency interested in ensuring success, are likely to have unique characteristics that affect, and usually increase, the likelihood of success. Many of these are not captured in structured surveys, and consequently positive outcomes may be due to these preexisting characteristics rather than to the success of the project.	• PRA techniques, in-depth interviews, and key informants can help identify and study "unobservables" that could not be easily addressed through formal surveys.

APPENDIX 14.4 THREE CASE STUDIES ILLUSTRATING THE USE OF MIXED-METHOD EVALUATIONS

This appendix describes three case studies illustrating different ways in which mixed-method designs have been used for the evaluation of development projects. The first case is an evaluation of the effectiveness of a program in Indonesia for targeting small-scale development assistance projects to poor rural communities, the second is a program in India for promoting democratic decentralization, and the third is an evaluation of a program in Eritrea concerning the social and economic impacts of feeder road construction. All three evaluations use sequential mixed-method designs.

1. Indonesia: The Kecamatan Development Project

The Kecamatan Development Project (KDP) in Indonesia is one of the world's largest social development projects. Implemented in the aftermath of the Suharto era and the East Asian financial crisis in 1998, KDP was primarily intended as a more efficient and effective mechanism for getting targeted, small-scale development assistance to poor rural communities, but it was also envisioned as a project that could help to nurture the proto-democratic state at the local level. KDP requires villagers to submit proposals for funding to a committee of their peers, thereby establishing a new (and, by design, inclusive) community forum for decision making on development issues (Guggenheim, 2006). Given the salience of conflict as a political and developmental issue in Indonesia, a key evaluation question is whether these forums are in fact able to complement existing local-level institutions for conflict resolution and in the process help villagers acquire a more diverse, peaceful, and effective set of civic skills for mediating local conflict. Such a question does not lend itself to an orthodox stand-alone quantitative or qualitative evaluation, but rather to an innovative mixed-method approach.

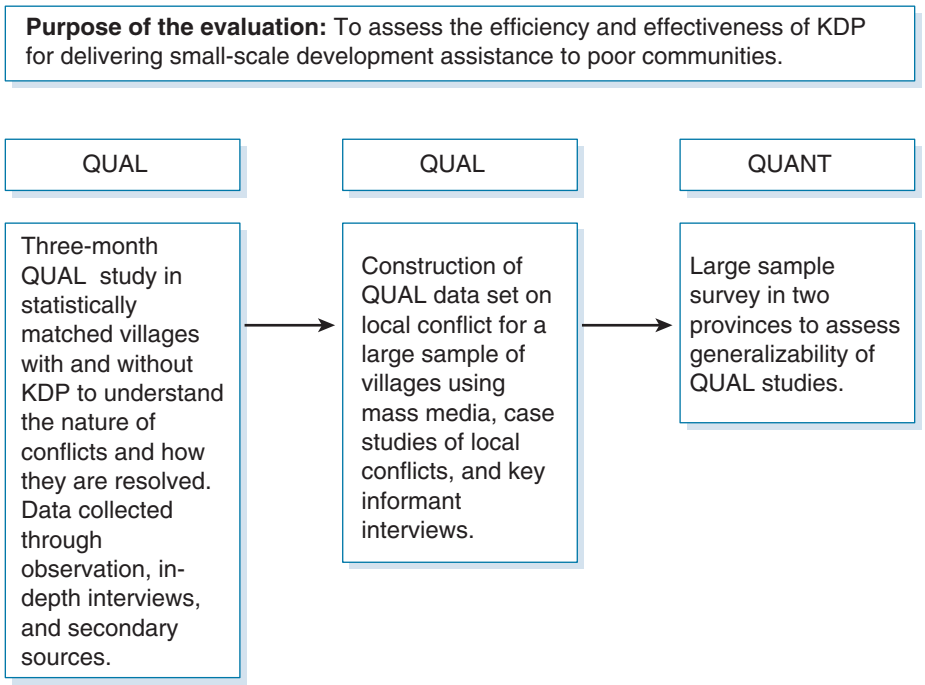
In this instance, the team decided to begin with qualitative work, as there was relatively little quantitative data on conflict in Indonesia and even less on the mechanisms (or local processes) by which conflict is initiated, intensified, or resolved. Selecting a small number of appropriate sites from across Indonesia's 3,500 islands and 350 language groups was not an easy task, but the team decided that work should be done in two provinces that were very different (demographically and economically), in regions within those provinces that (according to local experts) demonstrated both a high and low capacity for conflict resolution, and in villages within those regions that were otherwise comparable (as determined by propensity score matching methods) but that either did or did not participate in KDP. Such a design enabled researchers to be confident that any common themes emerging from across either the program or nonprogram sites were not wholly a product of idiosyncratic regional or institutional capacity factors. Thus, quantitative methods were used to help select the appropriate sites for qualitative investigation, which then entailed three months of intensive fieldwork in each of the eight selected villages (two demographically different regions by two high- or low-capacity provinces by two program or nonprogram villages). The evaluation design is summarized in Figure A14.4-1.

The results from the qualitative work—useful in themselves for understanding process issues and the mechanisms by which local conflicts are created and addressed (see C. Gibson & Woolcock, 2008)—fed into the design of a new quantitative survey instrument that would be administered to a large sample of households from the two provinces and used to test the generality of the hypotheses and propositions emerging from the qualitative work. A data set on local conflict was also assembled from local newspapers. Together, the qualitative research (case studies of local conflict, interviews, and observation), the newspaper evidence, data on conflict from national-level surveys, and key informant questionnaires provided a broad range of evidence that was used to assess the veracity of (and, where necessary, qualify and contextualize) the general hypotheses regarding the conditions under which KDP could and could not be part of the problem and/or solution to local conflict.

2. India: Panchayat Reform

A recent project evaluating the impact of *panchayat* (village government) reform—democratic decentralization in rural India—combines qualitative and quantitative data with a randomized trial. In 1992 the Indian government passed the 73rd amendment to the Indian constitution to give more power to democratically elected village governments (*gram*

FIGURE A14.4-1 ● The Mixed-Method Design for the Evaluation of the Kecamatan Development Project in Indonesia



panchayats—henceforth GPs) by mandating that more funds be transferred to their control and that regular elections be held, with one-third of the seats in the village council reserved for women and another third for "scheduled castes and tribes" (groups who have traditionally been targets of discrimination). It was also mandated that a deliberative space—village meetings (*gram sabhas*)—be held at least two times a year to make important decisions such as the selection of beneficiaries for antipoverty programs and discussing village budgets.

It is widely acknowledged that the state of Kerala has been by far the most effective in implementing the 73rd amendment. There were two elements that contributed to this success. The first was that the state government devolved significant resources to the GPs with 40% of the state's expenditures allocated to them; the second was the "people's campaign," a grassroots training and awareness-raising effort to energize citizens to participate, with knowledge, in the panchayat system. This led to better village plans, widespread and more informed participation, and more accountable government. Kerala is, of course, a special case, with very literate and politically aware citizens (literacy rates are close to 100%). The crucial policy question is whether the Kerala experiment can be replicated in much more challenging and more representative settings.

The northern districts of the neighboring state of Karnataka represent such settings. The literacy rate is about 40%, with high levels of poverty and a feudal social environment with high land inequality. These districts are also known to be beset by corruption and extremely poor governance. If a people's campaign could work in these districts, it could provide an important tool to transform the nature of village democracy in the country by sharply increasing the quality and quantity of citizen participation in the panchayat system and, in turn, have a significant effect on the standard of living. Also, these districts have access to two large national schemes that have substantially increased the funding of GPs, raising the budget

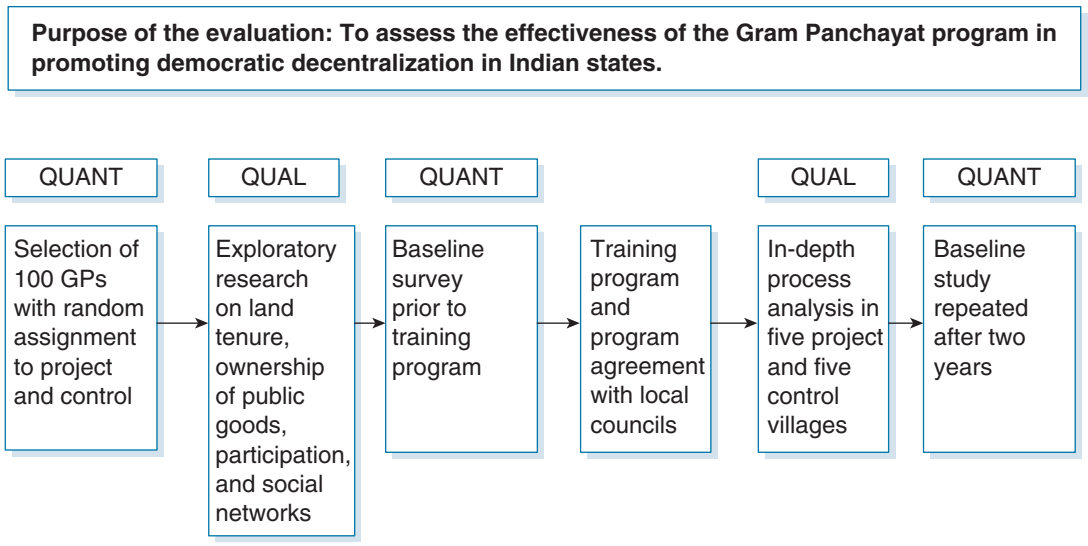
of GPs from about 200,000 Indian rupees a year to approximately 4,000,000 rupees. Thus, GPs in these districts have fulfilled the first element of the Kerala program—high levels of funding. The evaluation focuses on assessing the impact of the people's campaign. It randomly assigns 50 GPs as "treatment." Another set of GPs, matched to belong to the same county as the treatment GPs and with similar levels of literacy and low-caste populations and randomly chosen within this subset, is selected as "control" GPs. (They are also chosen to be at least one GP away from treatment GPs to avoid treatment spill-over problems.)

The "treatment" consists, initially, of a two-week program conducted by the Karnataka State Institute of Rural Development, which is responsible for all panchayat training in the state and has extensive experience in the field. The program trains citizens in participatory planning processes and deliberative decision making, and disseminates information about the programs and procedures of the panchayat. At the end of two weeks, a village meeting is held where priorities are finalized and presented to local bureaucrats. At a meeting with the bureaucrats, an implementation agreement is reached wherein the bureaucrats commit to providing funding and technical support for the selected projects over the course of the year. Following this initial training, the GP is monitored with monthly two-day visits over a period of two years in order to ensure the program's progress.

An extensive quantitative baseline survey was implemented in the 200 treatment and control villages randomly selected from the 100 selected GPs and completed a month prior to the intervention. The survey instruments, developed after several weeks of investigative fieldwork and pretesting, included village-level modules measuring the quality and quantity of public goods, caste and land inequality in the village, and in-depth interviews with village politicians and local officials. Twenty households from each village were also randomly chosen for a household questionnaire assessing socioeconomic status, preferences for public goods, political participation, social networks, and other relevant variables. Two years later, the same sample of villages and households was re-interviewed with identical survey instruments. These pretest and post-test quantitative data provide a gold-standard quantitative assessment of impact using a randomized trial. The design is summarized in Figure A14.4-2.

To understand "process" issues, however, equal attention was given to in-depth qualitative work. A subset of five treatment and five control GPs from the quantitative sample was selected purposively for the qualitative investigation. They were

FIGURE A14.4-2 ● The Mixed-Method Design for the Evaluation of the Gram Panchayat Program in India



selected to compare areas with low and high literacy and different types of administrative variation. A team of qualitative investigators visited these villages for a day or two every week over a two-year period investigating important dimensions of change: political and social dynamics, corruption, economic changes, and network affiliation, among other things. Under the supervision of two sociologists, the investigators wrote monthly reports assessing these dimensions of change. These reports provide a valuable in-depth look at month-to-month changes in the treatment and control areas that allow the assessment of the quality of the treatment, changes introduced by the treatment, and other changes that have taken place that are unrelated to the treatment. Thus the qualitative work provides an independent qualitative evaluation of the people's campaign but also supplements findings of the quantitative data.

An important challenge in understanding the nature of the 73rd amendment is to study participation in public village meetings (*gram sabhas*) held to discuss the problems faced by villagers with members of the governing committee. Increases in the quality of this form of village democracy would be a successful indicator of improvements in participation and accountability. To analyze this, a separate study was conducted on a sample of 300 randomly chosen villages across four South Indian states, including Kerala and Karnataka. Retrospective quantitative data on participation in the meetings are very unreliable, however, because people's memories are limited about what may have transpired at a meeting they may have attended. To address this issue, the team decided to record and transcribe village meetings directly. This tactic provided textual information that was analyzed to observe directly changes in participation (see Ban & Rao, 2009). Another challenge was in collecting information on inequality at the village level. Some recent work has found that sample-based measures of inequality typically have standard errors that are too high to provide reliable estimates. PRAs were therefore held with one or two groups in the village to obtain measures of land distribution within the village. This approach proved to generate excellent measures of land inequality, and since these are primarily agrarian economies, measures of land inequality should be highly correlated with income inequality. Similar methods were used to collect data on the social heterogeneity of the village. All this PRA information has been quantitatively coded, thus demonstrating that qualitative tools can be used to collect quantitative data. In this example, the fundamental impact assessment design was kept intact, and both qualitative and quantitative data were combined to provide insights into different aspects of interest in the evaluation of the intervention.

3. Eritrea: The Community Development Fund

The Eritrean Community Development Fund (CDF) was launched soon after Eritrea gained independence in the early 1990s, and it had two objectives: developing cost-effective models for the provision of community infrastructure (schools, health care centers, water, environmental protection, veterinary clinics, and feeder roads) and strengthening the participation of the local communities in the selection, implementation, and maintenance of the projects. Separate evaluations were conducted to assess the implementation and impacts of each of the six components. This case describes how mixed methods were used to strengthen the evaluation of the feeder roads component (similar approaches were used to assess the health and education components). Three feeder roads were being constructed, each between 50 and 100 kilometers in length and each serving many small villages that currently had no access to roads suitable for vehicular traffic.

The evaluation was not commissioned until work had already begun on each of the three roads, but none of which had yet been completed (planning and construction took on average around one year, with work often interrupted during the rainy season). The evaluation had a relatively modest budget, and no baseline data had been collected prior to the start of road construction. However, the CDF was intended as a pilot project to assess the efficiency and socioeconomic outcomes of each of the six project components, with the view to considering replication in a follow-up project. Consequently, policymakers were very interested in obtaining initial estimates, albeit only tentative, of the quantitative impacts of each component. Given the rapidly changing economic and social environment during the first decade of independence, it was recognized that the changes observed over the life of the different project components could not be assumed to be due to the project intervention. The need for some kind of simple attribution analysis was recognized, despite the absence of a conventional comparison group.

The possibility that was first considered was to try to identify areas with similar socioeconomic characteristics but that did not have access to a feeder road and that could serve as a comparison group. However, it was concluded, as is often the

case with the evaluation of the social and economic impact of roads, that it would be methodologically difficult to identify comparable areas and, in any case, extremely expensive to conduct interviews in these areas, even if they could be found. Consequently, the evaluation used a mixed-method design that combined a number of different data sources and that used triangulation to assess the validity and consistency of information obtained from different sources. The evaluation combined the following elements (see also Figure A14.4-3):

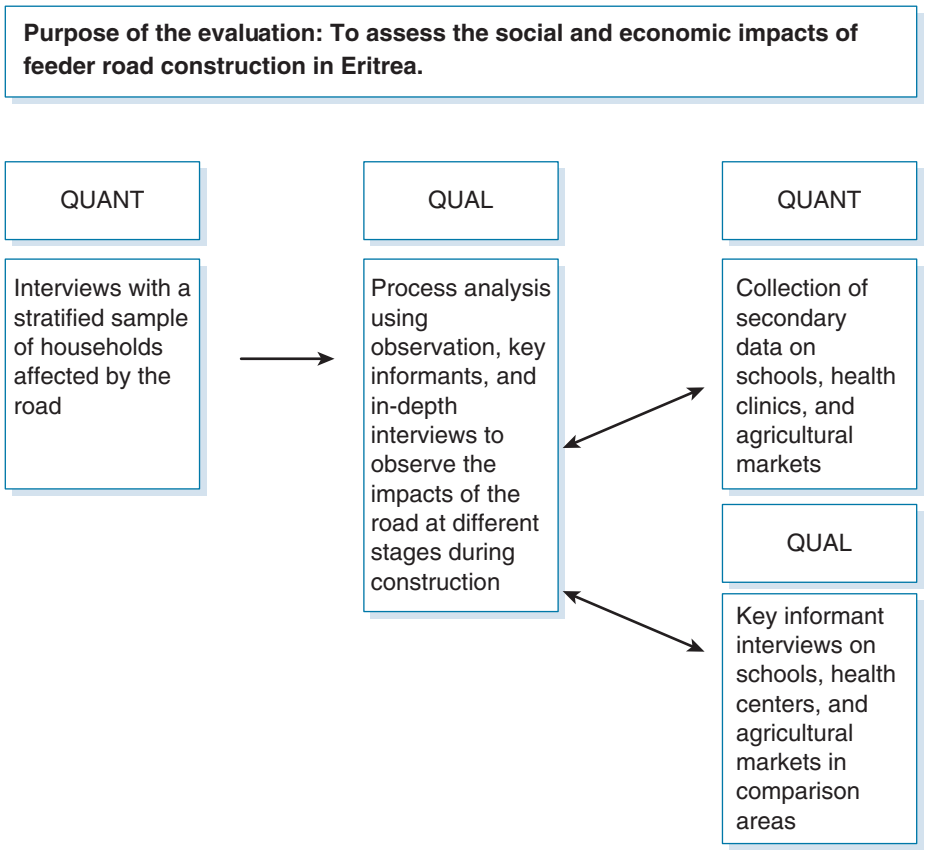
- The evaluation was based on a program theory model that described the steps and processes through which the project was expected to achieve its economic and social impacts and that identified contextual factors that might affect implementation and outcomes.
- The theory model also strengthened construct validity by explaining more fully the wide range of changes that road construction was expected to achieve so that impacts could be assessed on a set of quantitative and qualitative indicators.
- Some of the unanticipated outcomes that were identified in this way included strengthened social relations among relatives and friends living in areas that were previously difficult to reach, and strengthened and widened informal support networks as people were able to draw on financial, in-kind, and other support from a geographically broader network.
- Quantitative survey data were obtained from a stratified sample of households along the road who were interviewed three times during and after the road construction (the evaluation started too late for a pretest measure).
- The baseline conditions of the project population prior to road construction were reconstructed by combining recall of the time and cost to travel to school, reach a health center, transport produce to markets, and visit government agencies in the nearest towns with information from key informants (teachers, health workers, community leaders, etc.) and from secondary sources. Estimates from different sources were triangulated to test for consistency and to strengthen the reliability of the estimates.
- Data on comparison groups before, during, and after road construction were obtained from a number of secondary sources. Information on school attendance by sex and age was obtained from the records of a sample of local schools. In some cases, the data also included the villages from which children came so that it was possible to compare this information with recall from the interviews in project villages. Records from local health clinics were obtained on the number of patients who attended and the medical services that were provided. Unfortunately, the records did not permit an analysis of the frequency of visits of individual patients, so it was not possible to estimate whether there was a relatively small number of patients making frequent use of the clinics or a much larger number making occasional visits. Most of the local agricultural markets were cooperatives that kept records on the volume of sales (by type of produce and price) for each village, so this provided a valuable comparison group. It was planned to use vehicle registration records to estimate the increase in the number and types of vehicles before and after road construction. However, qualitative observations revealed that many drivers "forgot" to register their vehicles, so this source was not very useful.
- Process analysis was used to document the changes that occurred as road construction progressed. This combined periodic observation of the number of small businesses along the road, changes in the numbers of people traveling, and the proportions of people traveling on foot or using animal traction, bicycles, and different kinds of vehicles.
- Country-level data on agricultural production and prices, available over a number of years, provided a broader picture and were used to correct for seasonal variations in temperature and rainfall (both between different regions and over time). This was important in order to avoid the error of measuring trends from only two points in time.

All of the data sources were combined to develop relatively robust estimates of a set of social and economic changes in the project areas over the life of the project and to compare these changes with a counterfactual (what would have been

the condition of the project areas absent the project) constructed through combining data from a number of secondary sources. The credibility of the estimates of changes that could be (at least partially) attributed to the project intervention was then tested through focus groups with project participants, discussions with key informants, and direct observation of the changes that occurred during project implementation.

This evaluation, conducted with a relatively modest budget and drawing on the kinds of secondary data and recall information that are often available, illustrates how mixed-method designs can offer a promising approach to developing an alternative to the conventional statistical counterfactual, thus strengthening our understanding of the potential impacts of the majority of projects where the conventional counterfactual cannot be applied.

FIGURE A14.4-3 ● Mixed-Method Design for the Evaluation of the Eritrea Feeder Roads Project



APPENDIX FOR CHAPTER 15 SAMPLING STRATEGIES FOR REALWORLD EVALUATION

15.1 Using Power Analysis and Effect Size for Estimating the Appropriate Sample Size for an Impact Evaluation

Chapter 15 discusses sampling strategies and provides guidelines for estimating the sample size when working under real-world cost and time constraints. After reviewing different sampling strategies for quantitative, qualitative, and mixed-method evaluations, the chapter provides a detailed explanation of key sampling concepts: statistical significance, the power of the test, one- and two-tailed tests, and effect size. Worked examples are included to show how the required sample size is affected by decisions on the required power of the test and the estimated effect size. Understanding these concepts is essential for real-world evaluations, because when budgets are tight clients often pressure the evaluator to save money by reducing the sample size. The examples show that if sample size is reduced below a certain point, the statistical significance becomes so low that it is not possible to state with any confidence whether the project intervention has actually produced any change in the intended outcomes and impacts.

Appendix 15.1 provides a detailed explanation of the key statistical concepts required to determine the appropriate sample size for a real-world evaluation.

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 15.1 USING POWER ANALYSIS AND EFFECT SIZE FOR ESTIMATING THE APPROPRIATE SAMPLE SIZE FOR AN IMPACT EVALUATION

1. The Importance of Power Analysis for Determining Sample Size for Probability Sampling

Power analysis is a tool for determining the relationship between sample size and the level of statistical precision in a survey or evaluation. It can be used to estimate the sample size required to achieve a specified level of statistical precision, or to estimate the level of statistical precision that can be achieved with a given sample size.

Power analysis is particularly useful for RWE, as reducing the sample size is usually one of the most important options for reducing costs. However, overzealous reductions in sample size can be fatal. Many programs, even if well managed, can only expect to achieve relatively small improvements ("effect size"), and if the sample is too small, the statistical significance tests may commit a "Type II error" (a "false negative") and fail to detect what was in fact a statistically significant project effect. The smaller the effect size, the larger the sample required to detect it. Power analysis is an essential tool for the RWE evaluator, providing precise estimates of the sample size required to achieve the objectives of the evaluation. When the effect size is small, power analysis helps avoid selecting a sample that is too small, and conversely when the effect size is relatively large it can avoid wasting time and money through selecting a sample that is larger than necessary.

This is not just an academic concern, as is illustrated in the following example cited by Lipsey. In a review of 556 evaluations of different kinds of juvenile delinquency interventions, it was found that a high proportion of the estimated effect sizes were around 0.3. Many of the programs were designed to reduce recidivism (the likelihood that a juvenile delinquent after leaving the detention center would be reconvicted within a certain period of time, typically six months). It was found that 57% of the studies found an effect size of 0.3 or less, with about one-sixth having an effect size close to 0.3, which was equivalent to an average reduction of 24% in recidivism (a 38% recidivism rate for the project groups, compared to 50% for the control group). On the face of it this would seem to be a worthwhile program effect. However, in approximately 75% of the studies reviewed an effect size of 0.3 was not found to be statistically significant. Lipsey's review showed that in many cases the reason was that the sample size was too small (the test was "underpowered") to have been able to detect an effect of this size (Lipsey, 1990, Chapter 3, and summarized in Rossi, Lipsey, & Freeman, 2004, Chapter 10). In other words, a high proportion of these evaluation studies were doomed to failure simply because the sample was too small to detect the effect being studied. Table A15.1-4 shows that (with power = 0.8) while an effect size of 0.5 could be detected with a total sample size of 107, a sample of 618 would be required to identify an effect size of 0.2. With an effect size as small as 0.1, the total sample size would increase to 2,472!

While a full understanding of the logic of power analysis requires a solid grounding in statistics, the basic principles are easy to understand. For more complicated evaluation designs, or where high levels of statistical precision are required, it is advisable to consult with a statistical specialist. However, it is important for the evaluator to understand the basic principles of power analysis in order to know what questions to ask the statistician and to make sure these questions are addressed.

2. Estimating Effect Size

The "effect size" is the size of the change or effect that a program produces or is expected to produce. There are different types of effect sizes, including correlation coefficients and difference between means. In this chapter we will only consider estimates for the difference between means (called "d"). Technically, "d" is "*the difference between the outcome measured on program targets receiving the intervention and an estimate of what the outcome for those targets would have been had they not received the intervention*" (italics added; Rossi et al., 2004, p. 302). The larger the difference between the means of the two groups being compared (pretest/posttest project group or project and comparison group), the greater the effect size. Where possible, a *standardized effect size* is used so that comparisons can be made across programs or even across different kinds of effects (see next paragraph). However, it is sometimes necessary to use less precise measures, such as the number

of points increase on a behavior scale or aptitude test where the meaning of the change can be difficult to interpret. For binary variables, an odds ratio is often used (see Rossi et al., 2004, Chapter 10).

To obtain a standardized measure that can be used to compare the findings of different studies, the difference of means is divided by the standard deviation of the population. Thus

$$\text{Standardized effect size} = X_1 - X_2/\sigma X$$

where

$\overline{X_1}$ = the mean score for the project group

$\overline{X_2}$ = mean score for the total population (estimated from the comparison group)

σX = the standard deviation of the total population

For example, assume that after a microcredit program had been operating for two years, the average income of all adult women in the community was 300 pesos, while the average for women who had received loans was 350 pesos, and that the standard deviation for the total population was 100 pesos. The effect size would be calculated as

$$\text{Standardized effect size} = (350 - 300)/100 = 0.5$$

However, if the standard deviation had only been 75 pesos, then the effect size would have been 0.66.

Defining Minimum Acceptable Effect Size (MAES)

The minimum acceptable effect size, also called the critical effect size, is the minimum level of change that the evaluation design must be able to detect. The smaller the effect size that must be detected, the larger the required sample. Table A15.1-1 describes different criteria that can be used to define the MAES. In some cases, the MAES is defined in comparison to an accepted norm or target (for example, average test scores for a particular school grade), in others it is based on a comparison with other similar programs, and in yet other cases policymakers determine what is perceived by politicians and other stakeholders to be the minimum acceptable increase. Also, the MAES may be based on cost-effectiveness calculations. The MAES is normally population-specific, so that the acceptable effect size for a group of young men may be quite different from the acceptable effect size for a group of young women or a group of older people of either sex.

The choice of effect size is a key determinant of the required sample size. Where very small effect sizes must be detected, large samples will be required. Obviously, the MAES cannot be arbitrarily increased just to reduce the sample size. If it is believed that a 10% increase in school enrollment is the most optimistic estimate, it clearly does not make any sense to say, "Let us assume there will be a 25% increase." However, once clients understand the trade-offs between effect size and cost of the evaluation (i.e., sample size), there are sometimes ways to increase effect size. For example, if it is anticipated that enrollment is likely to increase more for girls than for boys, then it would be possible in the first evaluation to study program impact on girls. Obviously, it should be made completely clear to clients and readers that the evaluation does not cover the whole population. Another way to increase effect size is to improve the program design and delivery of services. Assume there is evidence from earlier programs that student math skills increase more when new textbooks are complemented by orientation sessions for teachers. Providing these orientation sessions might improve student performance and hence the effect size. Clearly there are trade-offs, and the client would have to decide whether the additional cost and effort of organizing the orientation sessions were justified.

Another important trade-off is to place less emphasis on statistical significance and greater emphasis on effect size. At least in the developmental stage of a program, it is important to avoid rejecting a potential program effect due to insistence on a high level of statistical significance. This is referred to as the trade-off between practical significance and clinical or statistical significance. At this stage it is important to identify all potentially credible effects. These can then be assessed more

carefully, and more rigorous statistical significance requirements can then be introduced at a later stage. For the present discussion this means that a lower level of statistical power might be used in the program development stage than would be used for full-scale testing of a program at a later stage.

3. Type I and Type II Errors

One of the challenges of sample design is to try, within available resources, to reduce two types of error (see Table A15.1-2):

- **Type I error.** Wrongly concluding that a program has a significant effect on the target variable when, in fact, it does not (error of inclusion or false positive).
- **Type II error.** Wrongly concluding that a program does not have a significant effect on the target variable when, in fact, it does (error of exclusion or false negative).

The relative importance of these two types of error varies according to the research context and the policy objectives. For example, once pilot testing of a new program or treatment (e.g., a conditional cash transfer, secondary school scholarship

TABLE A15.1-1 ⬢ Criteria for Determining the Minimum Acceptable Effect Size (MAES)

Criterion	Examples
1. Difference in the original measurement scale	When the outcome measure has a clearly understood meaning, the MAES may be stated directly in terms of this unit. For example, the dollar value of health services after the introduction of a new program, or the reduced recidivism rate for juvenile offenders.
2. Comparison with test norms or performance of a normative population	For a literacy program the MAES may be defined in terms of reducing the gap below the average grade score in the target scores.
3. Differences between criterion groups	Comparison of school with national grade scores
4. Proportion over a diagnostic or other success threshold	A mental health program might use a well-known test of clinical depression such as the Beck Depression Inventory, which defines a score of 17–20 as borderline clinical depression. The MAES could be defined as the proportion with scores below 17.
5. Proportion over an arbitrary success threshold	Proportion of families in an employment program with incomes above the federal poverty level
6. Comparison with the effects of similar programs	One of the goals of local irrigation programs is the proportion of farmers paying the water service charges required to maintain the system. MAES could be defined as the average repayment rate found in similar projects.
7. Conventional guidelines	Cohen (1988) proposed conventional guidelines, based on meta-evaluations conducted in different sectors of small effects (0.20), medium effects (0.50), and large effects (0.80).
8. Cost-effectiveness	The average unit cost of delivering services is compared with alternative programs or what is considered by stakeholders to be a "reasonable" unit cost.

Source: Adapted by the authors from Rossi, Lipsey, & Freeman (2004, pp. 318–319). Criterion 8 was added by the authors.

TABLE A15.1-2 ⬢ Type I and Type II Errors in the Interpretation of Evaluation Findings

Results of significance test on sample data	The True Population Circumstance	
	Intervention and control means differ	Intervention and control means do not differ
Significant difference found	Correct conclusion (probability = 1 – b). This is equivalent to the statistical power of the test.	**Type I error (false positive)**. Wrongly concluding the project does have a statistically significant impact (probability = a).
No significant difference found	**Type II error (false negative).** Wrongly concluding the project did not have a statistically significant impact (probability = b).	Correct conclusion (probability = 1 – a).

program, or new drug) has shown positive results, before the program or treatment is launched on a national scale, it will probably be necessary to use a higher statistical significance to avoid a false positive and to demonstrate that the program or treatment really does have a positive effect. In this case you would want a very small Type I error rate to minimize the false positive, and your Type I error rate can be lowered by setting a stricter standard of proof (higher significance level). In this case the financial and human costs of wrong decisions are very high. However, for many types of development programs, the primary concern may be to ensure that small but potentially important effects are not overlooked (Type II error). In this case you would want a very small Type II error rate to reduce the risk of a false negative (concluding the program does not have an effect when, in fact, it does). Type II error rate is lowered by increasing power. Meta-analysis studies in many sectors have found that the "effect size" of even the most successful programs is quite small, and consequently it is important to ensure that these are not overlooked by setting too rigorous criteria for accepting a statistically significant effect or having too little power.

The probability of making a Type I error (false positive) is set by the researcher. Increasing the sample size (and the cost) will increase the power and thus reduce the risk of making a Type II error. Consequently, it is essential to determine the relative importance of the two types of error and to set the statistical significance levels, and the resultant sample sizes, accordingly.

4. The Power of the Test

As indicated earlier, statistical power analysis is one of the key tools for estimating how large a sample is required to be able to find a statistically significant project impact if one really does exist. A number of authors have argued that in cases where a project is not expected to produce a very large impact (effect), many evaluation studies have wrongly assumed that a project did not have a (statistically significant) effect when, in fact, the sample was too small to have been able to detect the effect if it did exist (Lipsey, 1990; Rossi et al., 2004). Statistical power is "the probability that an estimate will be statistically significant when, in fact, it represents a real effect of a given magnitude" (Rossi et al., 2004). The normal convention is to set power equal to 0.80, meaning that there is an 80% chance that a particular sample will reject the null hypothesis (i.e., will find a statistically significant difference) if the program really does have an effect (see Box A15.1-1).

Where it is particularly important to avoid Type II errors and to ensure that real program effects are not rejected, it is possible to set power equal to 0.90 or even 0.95 or 0.99. However, the reason why these higher power levels are often not used is that the increase in power requires a significant increase in the sample size. For example, Table A15.1-4 shows that for a one-tailed test, increasing power from 0.8 to 0.9 may require an increase of between 30% and 38% in the sample size, while raising power from 0.8 to 0.95 may require up to a 75% increase. Box A15.1-2 points out that when statistical advice is sought on power and effect size calculations, the statistician will probably not be familiar with the particular field, and consequently the evaluator must do his or her homework and come with estimates of the expected effect size.

BOX A15.1-1

Conventions for Defining the Statistical Power of the Test

- The power of the test is conventionally set at 0.80 (an 80% probability that a sample will find a statistically significant result if the null hypothesis is false).
- Power analysis usually assumes that the 0.05 (a) significance level is being used.
- Where greater precision is required, power can be set at 0.90 or 0.95 (or even higher), and the significance level can be set at 0.01 (or even higher). However, increasing the precision level will significantly increase the required sample size.
- Estimates of power and the sample size required to achieve a certain power are often approximate because power curves are often steep and precise values difficult to estimate.

BOX A15.1-2

Sampling Specialists Are Often Not Familiar With the Fields of Application on Which Their Advice Is Sought

"It is not a minor problem that those who are able to do power calculations readily are generally those who least know the fields of application, and those who best know the fields of application are least able to do power calculations."

Source: Kraemer & Thiemann (1987, p. 99).

An Example: The Statistical Power of an Evaluation of Special Instruction Programs on Aptitude Test Scores[20]

This example is a hypothetical study that was conducted to assess the impact of special instruction programs on aptitude scores of fifth-grade students. It was known from previous studies that the mean score for fifth graders on this test was 200 (see Figure A15.1-1b), and that the population standard deviation (SD) was 48. The standard error (SE) of the distribution of means (some writers use the term *standard deviation of the distribution of means*, e.g., Aron & Aron, 2002, p. 157) for a sample size of 64 is calculated as

$$SE = \sqrt{SD^2/64} = 6$$

This means that for the null hypothesis (that special instruction did not raise aptitude test scores) to be rejected at the 0.05 level, the mean for the treatment group would have to be greater than

$$200 + (6 \times 1.64) = 209.84$$

[20] This example uses the Z-test procedures to illustrate the stages of the analysis. Many researchers and computer programs use the t-test for testing differences between means. The two procedures are similar, although there are some technical differences (see Frankfort-Nachmias & Leon-Guerrero, 2011, pp. 267–268). It is assumed in this example that the conventional power level of 0.8 is used (see Box A15.1-1). The table also shows that the sample size increases dramatically if higher power levels (0.90 or 0.95) are specified. It is assumed in this example that the conventional power level of 0.8 is used (see Box 15.6). The table also shows that the sample size increases dramatically if a higher power level (0.90 or 0.95) is used.

where

1.64 = Z score for the 0.05 significance level.

The shaded area at the right end of the curve in Figure A15.1-1b represents the area for rejecting the null hypothesis.[21] Note that the numbers on the horizontal axis represent the distribution of means for all possible samples—not individual test scores.

It was hypothesized, based on a review of the literature, that the treatment (special instruction) could be expected to raise the mean test score by 8 points, so that the mean of the treatment group would be 208 points.[22] The effect size (ES) is calculated using this equation:

$$ES = X_2 - X_1/SD$$

where

X_2 = the sample mean

X_1 = the population mean

SD = the standard deviation

In this case, ES = (208 – 200)/48 = 0.166.

As there is only one sample mean, this equation can be used. In cases where two sample means are being compared, it is necessary to estimate the adjusted effect size D (see Table A15.1-4).

The shaded area in the upper distribution (Figure 15.1-1a) shows the probability that a sample of the treatment group would have a mean score sufficiently high (i.e., above 209.84) to reject the null hypothesis, even when the treatment really does have an effect. In this case the probability is only 40% that a sample from the treatment group would find a statistically significant difference (even when the project has produced a real increase in test scores). In other words, the risk of a Type II error is very high (60%).

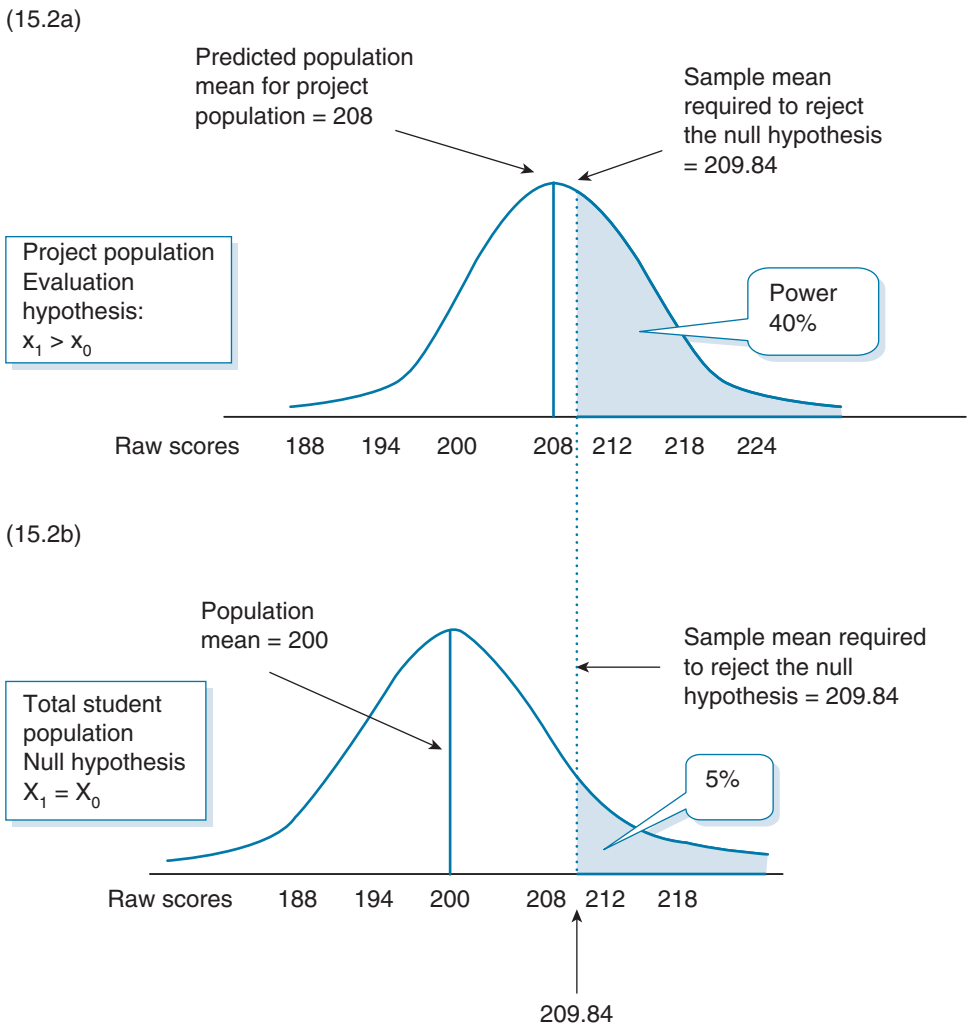
Why is this so? Many people might assume that a carefully selected sample of the treatment group would always find a statistically significant difference (if, as in this case, the treatment really did have an effect). The reason can be seen by comparing the two distributions in Figure A15.1-1. It can be seen that the hypothesized increase in test scores (the effect size) is quite small (0.166), and that there is considerable overlap between the two distributions. This means that if the null hypothesis was true and the treatment had no effect, many samples would by chance have means equal to or greater than the hypothesized treatment mean score of 208. Using the 0.05 significance level, the mean score would have to be greater than 209.84 to reject the null hypothesis—even when the treatment does have an effect.

Figure A15.1-2 shows the effect of an increased effect size on the power of the test. Under this scenario (scenario 2), the average increase in test scores compared to the control group is 16 points (compared to 8 points in the previous example). This represents an effect size of approximately 0.33. It can be seen in the upper half of the figure (Figure A15.1-2a) that 85% of the distribution is now to the right of the sample mean required to reject the null hypothesis. This means that the power of the test has increased to 0.85, and there is only a 15% risk of rejecting the null hypothesis (false negative).

[21] This is equivalent to a Z score of 1.64 (see Aron & Aron, 2002, Chapter 2). The Z score is defined as (sample mean – population mean)/standard deviation of the population mean, in this case (209.84 – 200)/6.

[22] It was assumed, as is usually done unless other information is available, that the SD of the treatment group would be the same as for the total population.

FIGURE A15.1-1 ● Statistical Power Analysis: Testing the Effect of a Program to Raise Mathematical Aptitude Test Scores (Scenario 1)



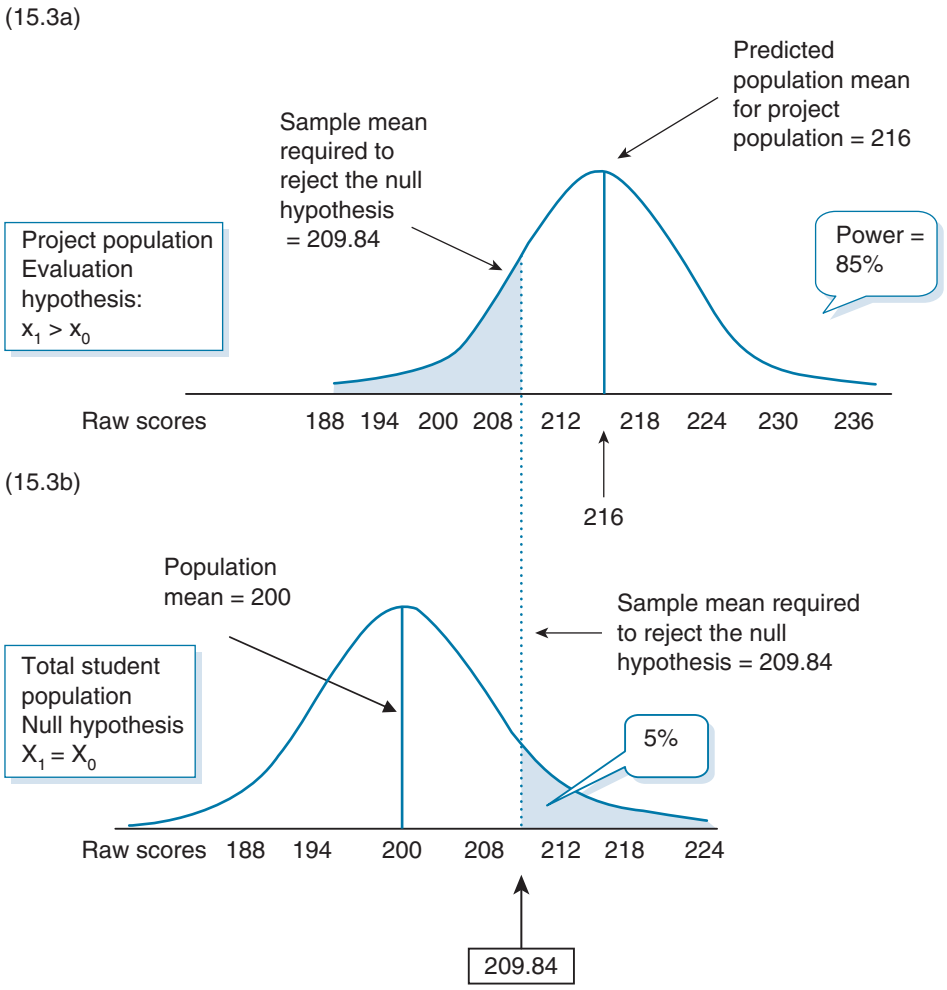
The figure presents the distribution of mean test scores of 64 students on a standardized mathematical aptitude test in a fictional study. The lower curve (Figure 15.1-2b) is based on the known distribution of means for the total student population and has a mean of 200. The upper distribution (Figure 15.1-2a) is based on a predicted distribution assuming the evaluation hypothesis (that the treatment raises aptitude score) is correct. It is hypothesized that the mean for the project group is 208, representing an effect size of approximately 0.16. Shaded sections of both curves indicated the areas in which the null hypothesis will be rejected. Power = 0.40, indicating that there is only a 40% probability that any particular sample will reject the null hypothesis, even though the project treatment did produce a statistically significant increase in test scores.

Source: Adapted from Aron & Aron (2002, Figure 7.1). Some of the figures have been slightly adjusted to make the results consistent with the Kraemer and Thiemann power table.

Calculating Statistical Power

To calculate the statistical power of the test in a particular evaluation, it is necessary to know the effect size (see Table 15.1-3) and the sample size. The statistical power can then be obtained from a power table (see Table 15.1-4 for a simplified power table). In our example the effect size is 0.166, which is quite small. Our sample size is 67. Consulting a more complete

FIGURE A15.1-2 ● Statistical Power Analysis: Testing the Effect of a Program to Raise Mathematical Aptitude Test Scores (Scenario 2)



The figure presents the distribution of mean test scores of 64 students on a standardized mathematical aptitude test in a fictional study. The lower curve (Figure A15.1-2b) is based on the known distribution of means for the total student population and has a mean of 200. The upper distribution (Figure A15.1-2a) is based on a predicted distribution assuming the evaluation hypothesis (that the treatment raises aptitude score) is correct. It is hypothesized that the mean for the project group is 216, representing an effect size of approximately 0.33. Shaded sections of both curves indicate the areas in which the null hypothesis will be rejected. Power = 0.85, indicating that there is an 85% probability that any particular sample will reject the null hypothesis when the project treatment did produce a statistically significant increase in test scores.

Source: Adapted from Aron & Aron (2002, Figure 7.2). Some of the figures have been slightly adjusted to make the results consistent with the Kraemer and Thiemann power table.

power table than Table 15.1-4 (Kraemer & Thiemann, 1987, p. 105) shows that for an effect size of 0.17 (with a one-sample t-test) and a sample size of 64 the statistical power = 0.40.

Scenario 1: Project effect size = 8 points test gain (E = 0.166)

Scenario 2: Project effect size = 16 points test gain (E = 0.33)

If the predicted average increase in test scores had been 16 instead of 8 (an effect size of 0.33), then the power of the test would increase to 0.85. In this case the risk of wrongly rejecting a real project effect would drop from 60% in the earlier case to only 15%. Increasing effect size always increases the power of the test, so it is obviously desirable to have as large an effect size as possible. We saw earlier that there are various ways in which the effect size, and consequently the power of the test, can be increased.

So far the discussion of the statistical power of the test has been based upon a sample size of 64. It is important to understand the dramatic effect that the sample size has on the power of the test and consequently on the statistical and operational utility of the findings. If the sample size was significantly larger, the standard error would decrease and the curves in Figure A15.1-1 would get skinnier. This would mean that there would be less overlap between the top and bottom curves, and the power of the test would increase. This is illustrated in Table A15.1-3. When the estimated effect size is quite low (0.16) the sample size of 64 only gives a statistical power of 0.40, which is too low to be of any practical utility. If the sample size is increased to 124 the statistical power increases to 0.60, which is higher but still too low to be useful. We can see that a sample size of 211 would be required to achieve a statistical power of 0.80, which is normally considered to be the lowest level to indicate a potentially statistically significant result. Increasing the sample size to 292 and 369 would increase power to 0.90 and 0.95, respectively. On the other hand, when the estimated effect size is 0.32 a sample of 64 already produces an operationally useful power level of 0.85, and the sample would only have to increase to 100 to achieve a 0.95 power level.

Deciding How to Set the Power Level

Deciding how to set power involves trade-offs. If a low power level is used, it will probably be possible to reduce the sample size, but it will be difficult to find results that are statistically significant. So setting a low power level means clients want to pay more attention to effect size than to statistical significance. In the exploratory stage of an evaluation, the goal is often to determine whether there are potentially important effects that should be explored further. From this perspective, it may be worth defining a fairly low power level (increasing the risk of a Type I error) so as not to exclude potentially interesting but small effects.

TABLE A15.1-3 ● How Sample Size Affects Statistical Power in the Previous Example When Effect Size = 0.16 and 0.32

Effect Size = 0.16		Effect Size = 0.32	
Sample size	**Power**	**Sample size**	**Power**
64	0.40		
93	0.50		
124	0.60		
161	0.70		
211	0.80		
		64	0.85
292	0.90	79	0.90
369	0.95	100	0.95
539	0.99	145	0.99

Source: Statistical power estimated using Kraemer & Thiemann (1987; master table for 5% level and one-tailed test, pp. 105–108).

5. One- and Two-Tailed Statistical Significance Tests

Most programs have a clearly defined objective to produce a positive outcome, either by reducing a negative indicator (infant mortality, illiteracy, criminal behavior) or increasing a positive indicator (school attendance rates, household income, agricultural output). When the direction of the desired change is known (to increase school enrollment or reduce malnutrition), a "one-tailed" statistical test can be used. Although this is much less common, the direction of the effect is sometimes not known, and it will be necessary to use a "two-tailed test." For example, introducing school fees might increase enrollment (because the quality of education and the maintenance of the facilities might improve), or it might decrease attendance if many poor families are not able to pay the fees. When a two-tailed test is required, the size of the sample will increase from 20 to 27% when power = 0.8 and 15 to 23% when power = 0.9.[23] The smaller increases are associated with larger effect sizes.

6. Determining the Size of the Sample

The Null Hypothesis, the Evaluation Hypothesis, and Deciding the Power and Statistical Significance Level

To evaluate whether an intervention has had a statistically significant impact, it is necessary to start with a *null hypothesis*. We use a null hypothesis because it is never possible to prove statistically that a hypothesis is true, but only to estimate the probability that an effect size as large as the one observed could have occurred if there really was no difference between the project group and the total population. The null hypothesis states that there is no difference between the total population and the project group with respect to the outcome measure (aptitude test score, household income, proportion of girls attending secondary school, etc.). The null hypothesis (H_0) is specified as follows:

$$H_0: x_0 = x_1$$

where

x_0 = mean or other outcome measure for the total population

x_1 = mean or other outcome measure for the project population

The alternative research hypothesis (H_1) is that there is a difference between the project and comparison groups, which is specified as follows:

$$H_1: x_0 \neq x_1$$

A key decision in the evaluation design is the relative importance of (a) ensuring that H_1 is not accepted when it is not true (false positive), and (b) not rejecting H_1 when it is true (false negative). Traditionally, researchers have been concerned to avoid the false positive and have required a high standard for rejecting H_0. Conventionally, a statistical significance level of 0.05 is used, meaning that H_0 will only be rejected if there is less than a 1 in 20 chance of it being true. The higher the statistical significance requirement for rejecting the null hypothesis, the more difficult it is to accept H_1 and to decide that the project intervention did have an effect. Consequently, evaluators and program managers concerned to identify effective interventions will often use a less stringent significance level so as to increase the likelihood of detecting a project effect. This is particularly important in the development field, where many well-designed programs (e.g., health, education, poverty reduction) are expected to only produce a small improvement (the expected effect size is small). Consequently, the use of stringent significance levels means that many small but important effects may not be detected. So an important challenge for the evaluator is to decide on the relative importance of false positives and false negatives, and to select appropriate statistical significance levels for the testing of H_0 and H_1. In a real-world environment, the challenge is to satisfy

[23] Kraemer & Thiemann (1987) include master tables for both one- and two-tailed tests.

the client's needs for acceptably precise estimates for the purposes of decision making while remaining within the available budget and time.

Evaluability Assessment

Many evaluators become overwhelmed by the statistical calculations required to estimate sample size, and it is often forgotten that there are several important preliminary steps in the evaluation before even thinking about testing and sample sizes. The first step involves a thorough review of the existing evidence from the research literature, reports of similar programs, and if possible, discussions with experts. This will help determine whether the proposed program is likely to have an effect. An exploratory field study should then be conducted to understand the program and how it operates, and to determine whether the program effects can be measured at this point in time, for this particular project, and with the available resources. This *evaluability assessment* may suggest various reasons why an evaluation may not be appropriate. The literature may suggest the program model is unlikely to work, it may be considered too early in the project cycle to measure effects, the target population may be too small to permit the kinds of statistical analysis required, or the proposed effects may be too difficult to quantify and measure. Assuming that none of these problems are considered too serious, it is then possible to begin to plan the evaluation. Henry (1990, Chapter 3) and Kraemer and Thiemann (1987, Chapter 2) provide useful overviews of these preparatory stages of the evaluation.

Estimating the Required Sample Size for Power Analysis[24]

The choice of sample size cannot be made in a vacuum, as there are trade-offs concerning the most effective way to use the evaluation resources. For example, spending more on sample size may mean less resources/time for following up on nonresponses to reduce this often important source of nonsampling bias. Each of the steps is illustrated for a hypothetical example in which the required sample size for evaluating the impact of special instructions on student performance on a standardized mathematical aptitude test is estimated. We use the previous example but change several details.

The purpose of the evaluation is to determine whether the special instruction has produced a statistically significant increase in student performance on the standardized mathematical aptitude test. In the present case, the following apply:

- The mean aptitude test score for the total student population is again 200, and the population standard deviation, based on previous studies, is again estimated to be 48.
- The survey will only cover students who received the special instruction, because the mean and standard deviation of test scores for the total student population are known. Consequently, this is a *single sample comparison* with the population mean.
- The total fifth-grade student population is over 20,000, and the number of fifth-grade students who have received the special instructions is 5,000.
- The null hypothesis to be tested is $H_0 = H_1$, where

  H_0 = the mean aptitude test score for the total fifth-grade student population

  H_1 = the mean aptitude test score for students who have received special instruction

The following steps must be followed for determining the efficient sample size:

Step 1: Determine the purpose of the evaluation. Smaller sample sizes are usually needed for exploratory studies than for testing hypotheses concerning project effects.

[24] For a more detailed discussion on the estimation of sample size, see Henry (1990, Chapter 7).

~ The more precise the required results, the greater the required sample size.

In the example, the Ministry of Education stated that the purpose of the evaluation is to determine whether the special instruction program is a cost-effective way to increase student mathematical skills. The ministry also needs to demonstrate that the program is more effective than other approaches. There are many supporters of alternative mathematics teaching programs who will try to challenge the findings, so the ministry requires "acceptable professional levels of significance testing."

Step 2: Determine if it's a one- or two-tailed test. In most cases a one-tailed test will be used, but the evaluator should always check this.

~ Using a two-tailed test will increase the sample size by about 40%.

In the example the purpose is to test whether aptitude scores have increased, so a one-tailed test is appropriate.

Step 3: Estimate the standard deviation (SD) of the population.[25] The size of the SD will affect the estimation of the effect size and consequently the sample size. Sometimes the SD is known from previous studies. In other cases it can be estimated from a small pilot study.[26]

~ The larger the SD, the larger the required sample size.

In the present example, testing has found that the mean test score for fifth-grade students is 200 and the SD is 48. It is assumed that the SD will be similar for the project population.[27]

Step 4: Determine the minimum acceptable effect size (MAES).

~ The smaller the MAES, the larger the required sample size.

A meta-analysis review of other programs to increase math skills has found that effect sizes range between 0.1 and 0.25. In addition, the ministry indicated that a 5% increase in test scores is the minimum effect that could justify funding to continue the program. This would require an increase of 10 points in the average population test score of 200. With a population SD of 48, this is equivalent to the following effect size:

$$\text{MAES} = [210 - 200]/48 = 0.208 \text{ (this is rounded to 0.2 to simplify the calculations)}$$

Step 5: Determine whether the statistical significance of the findings will be tested at the 0.01 level (when a high level of precision is required), at the 0.05 level (the normally accepted level), or at the 0.1 level (for exploratory studies).

~ The higher the required significance level, the larger the sample.

In the example, it is agreed to use the 0.05 significance level.

Step 6: Decide on the required power of the test.

~ The higher the required power, the larger the sample.

In the example, the ministry agrees to use the conventional power = 0.8. It initially wished to use the 0.9 power level to achieve a higher level of statistical precision, but when it was advised that this would increase the sample size by more than 200 interviews (increase from 618 to 854), it decided to accept the 0.8 level.

---

[25] It is important to check that estimates of the SD from previous studies are applicable. If different populations were studied or different questions were asked, earlier studies may not provide a good estimate of the SD for the present study.

[26] Henry (1990, p. 119) suggests that if no other information is available, a rough estimate can be obtained by dividing the range (the difference between the highest and lowest values of the variable) by 4. However, doing a power analysis for proportions requires a different procedure, which does not require the standard deviation.

[27] It is usually assumed that the project group will have the same standard deviation as the total population. However, in nonequivalent control group designs it is possible that the project group could have a different standard deviation. Where sample precision is important, it would be useful to consult with a statistician as to whether an adjustment should be made. For most RWE purposes, this is probably not a major concern.

Step 7: Consult a master statistical power table to calculate the required sample size for a given effect size and power level. Table A15.1-4 is a simplified version of a power table that only covers the 0.05 significance level and only a one-tailed test.[28] The use of the table involves the following steps:

- Determine whether a one- or two-sample comparison is being used. In a one-sample comparison a sample mean is compared with the population mean, as is the case with the present example. When two samples are being compared, as is the case when the project sample is compared with a control group sample, then an adjusted effect size must be used (the second column of Table A15.1-4).
- When a one-sample model is used the effect size (ES) can be read directly from the master power table. This is the first column in Table A15.1-4.
- If a two-sample model is used, the adjusted effect size is obtained from the second column of Table A15.1-4.[29]
- Decide the required significance level and then consult the appropriate power table. In our example, Table A15.1-4 assumes that the significance level is set at 0.05.
- Consult the column for the specified power.
- Consult the row for the effect size or adjusted effect size.
- Locate the intersecting cell to find the required sample size for each group.

In the example:

- The client indicated that normal standards of professional precision should be followed, so the 0.05 significance level will be used.
- The power level is 0.80.
- As this is a one-sample test, the effect size column is consulted and the 0.2 value is used. This is the MAES. The intersection of the power column and 0.2 row shows that 618 interviews would be required.

Step 8: Determine whether to use the finite population correction factor. When the sample represents a high proportion (i.e., more than say 10%) of the total population it is possible to use the finite population correction factor (FPCF) to improve the precision of the estimates (or to produce a slight reduction in the sample size). In the present example, the sample size is only about 3% of the total student population of 20,000, so the use of the FPCF is not required.

Estimating Power for Multiple Regression

In this chapter we only discuss the calculation of power for the comparison of sample means. The calculation of power for multiple regression is a little more complicated, as it is necessary to estimate the number of variables in the model, the size of the partial correlation of the variable of interest and the outcome (controlling for all other variables in the model), and the multiple correlation of all the other variables in the model with the outcome. For a discussion of power analysis for linear regression, see Kraemer and Thiemann (1987, Chapter 6).

7. Factors Affecting the Sample Size

In addition to the factors discussed earlier, the required sample size is affected by the following:

- The reliability of the instrument used to measure effects. In cases where effects cannot be reliably measured, the variance of the estimates increases and the sample size required to achieve a certain power also increases.

---

[28] For more comprehensive master tables covering both one- and two-tailed tests and 0.5 and 0.1 confidence levels, see Kraemer & Thiemann (1987, pp. 105–112).

[29] The following equation is used to obtain the adjusted effect size (Δ):

$\Delta = ES/(ES^2 + 1/pq)^{1/2}$

where ES = effect size; p and q = the proportion of the sample in group 1 and group 2, respectively.

TABLE A15.1-4 ● Approximate Sample Sizes for One- and Two-Group Comparisons of Means to Attain Various Criterion Levels of Power for a Range of Effect Sizes With a One-Tail Test and Significance Level = 0.05

Effect Size		Power			
Effect size (ES) for comparing a single sample with the population mean	Adjusted effect size (Δ) for comparison of two sample means	0.95	0.90	0.80	0.60
0.10	0.05	**4325**	**3423**	**2472**	**1142**
0.20	0.10	**1078**	**854**	**618**	**361**
0.30	0.15	**477**	**378**	**274**	**161**
0.40	0.20	**267**	**212**	**154**	**91**
0.50	0.24	**184**	**146**	**107**	**64**
0.6	0.29	**125**	**100**	**73**	**39**
0.8	0.37	**76**	**60**	**45**	**26**
1.0	0.45	**50**	**40**	**30**	**19**
1.2	0.51	**39**	**26**	**19**	**13**
1.4	0.57	**28**	**22**	**16**	
1.6	0.62	**24**	**19**	**14**	
1.8	0.67	**18**	**15**		
2.0	0.71	**16**	**14**		

Notes:

1. The numbers in the table refer to the required sample size for different combinations of effect size and power. For two-group studies, the number refers to the total sample size for the two groups. It is assumed that the sample size will be the same for both groups. If the sample sizes are different, this may require a slight adjustment to the estimated total sample size (see Note 3).

2. The effect size column is used for one-group analysis (comparing a sample with the total population). This is a slightly simplified estimate, as an adjustment should be made for the single sample t-test (Kraemer & Thiemann, 1987, p. 101, and Chapter 4, Section 4.2), but the effect of the adjustment is very small.

3. The adjusted effect size column is for the comparison of two groups (e.g., comparing project and comparison groups). This uses the adjustment formula for the two-sample t-test: $\Delta = ES / (ES^2 + 1/pq)^{1/2}$ where Δ = adjusted effect size; ES = effect size; p and q = proportion of the sample in group 1 (e.g., project group) and group 2 (e.g., the comparison group) (Kraemer & Thiemann, 1987, p. 101, and Chapter 4, Section 4.3). In the table we have assumed the sample size will be equal for both groups. Sample sizes could be adjusted slightly if the proportions were not equal.

4. The table assumes a one-tailed test at the 0.05 significance level. The required sample size increases between 15% and 30% when a two-tailed test is used (depending on the power level). Required sample sizes increase if the 0.01 significance level is used. When the 1% significance level is used, there is a significant increase in the required sample size (for example, there is an increase of approximately 45% for the 0.95 power level and approximately 60% for the 0.80 power level).

Source: Calculations by the present authors based on Kraemer & Thiemann (1987, master tables).

- The evaluation design. More sophisticated evaluation designs that include intervening (mediator) variables can reduce the variance and consequently the required sample size to achieve a certain statistical power.
- The estimated sample drop-out rate (for panel studies).
- Changes in the size and characteristics of the total study population (for longitudinal studies).
- If effect size can be increased, this will automatically increase statistical power and/or permit sample size to be reduced. As indicated earlier, it may be possible to increase effect size by improving the design and administration of the treatment (special instruction) or by limiting the study to those groups where the effect is expected to be greatest.

APPENDICES FOR CHAPTER 17 GENDER EVALUATION: INTEGRATING GENDER ANALYSIS INTO EVALUATIONS

Part I: Specific Gender Evaluation Tools	
17.1	The Harvard Gender Analysis Framework
17.2	Tools for More In-Depth Gender-Responsive Evaluation (GRE) Designs
17.3	Two Examples of a Women's Empowerment Index
Part II: General Approaches and Methodologies for Gender Analysis	
17.4	The World Bank/IEG Implementation Completion Report: Gender Flag Review
17.5	Recommended Structure for the Evaluation Cooperation Group GRE Reports
17.6	Widely Used Gender Indices and Checklists
17.7	Evaluation Approaches Used in Standard GRE Designs
17.8	Examples of the Application of the Different GRE Designs
17.9	Evaluating the Gender Dimensions of Complex Development Programs
17.10	The Contribution of Gender-Responsive Budgeting (GRB) to National Policy Dialogue
17.11	Feminist Critical Theory
17.12	Gender-Sensitive Data-Collection Methods and Applications Used in the Case Studies in Appendix 17.14 and the Additional Time and Cost Implications (Compared to Standard Evaluation Methods)
Part III: Case Studies Illustrating Different Approaches to Gender Analysis	
17.13	Summary of Design and Data-Collection Methods Used in the Seven Case Studies Described in Appendix 17.14
17.14	Case Studies Illustrating Different Gender Impact Evaluation Methodologies

Chapter 17 addresses how to design gender evaluations under different scenarios. It begins by presenting the case; why it is important to incorporate a gender focus into almost all evaluations of development projects, programs, and national policies; and why, given the importance of gender, so few evaluations and agency evaluation strategies adequately address gender. Many agencies will gradually introduce a gender focus into their evaluations over a period of years. The results, operational utility, and cost-effectiveness of a gender focus will often depend on how well the first evaluations are received. Given this gradualist approach, a continuum of gender evaluation strategies is presented, and it is expected that many evaluation strategies will gradually move along this continuum over a period of years.

(Continued)

(Continued)

A total of 14 appendices are included with this chapter (list above). They are divided into three parts: Specific Gender Evaluation Tools (Part I); General Approaches and Methodologies for Gender Analysis (Part II); and Case Studies Illustrating Different Approaches to Gender Analysis (Part III). The case studies in Part III describe seven different approaches to gender analysis. Cases include a range of quantitative and qualitive methodologies, each of which describes the purpose of the evaluation, the methodology, the types of analysis, and the findings.

Many of the technical terms in these appendices are included in the Glossary in the book.

CHAPTER 17 APPENDICES PART 1

Specific Gender Evaluation Tools

APPENDIX 17.1 THE HARVARD GENDER ANALYSIS FRAMEWORK

TABLE A17.1-1 ⬢ Harvard Gender Analysis: Activity Profile (Adapted by the present authors)

Month/Season to Which Analysis Refers:[1]							
Socioeconomic Activity	**Gender/Age-Groups (hours per day or per week)**						
	Females			**Males**			
	Child	**Adult**	**Elder**	**Child**	**Adult**	**Elder**	**Locus**[1]
1. **Production of goods and services**							
a. Product/Service							
i. Functional activity[1]							
ii. Functional activity							
b. Product/Service							
i. Functional activity							
ii. Functional activity							
2. **Social reproduction and maintenance of human resources**							
a. Product/Service							
i. *Functional activity*							
ii. *Functional activity*							
b. *Product/Service*							
i. *Functional activity*							
ii. *Functional activity*							

(Continued)

(Continued)

Month/Season to Which Analysis Refers:[1]							
Socioeconomic Activity	Gender/Age-Groups (hours per day or per week)						
	Females			Males			
	Child	Adult	Elder	Child	Adult	Elder	Locus[2]
3. ***Community management***							
a. *Community program*							
i. *Functional activity*							
ii. *Functional activity*							
b. Community program							
i. *Functional activity*							
ii. *Functional activity*							

1. The categories in italics were added by the present authors.

2. Within home; family, field, or shop; in the local community; beyond community.

Source: Adapted by the authors from Overholt, Cloud, Anderson, & Austin (1985, Table 1).

TABLE A17.1-2 ⬢ Harvard Gender Analysis: Access and Control Profile (Adapted by the present authors)

	Access				Control			
	Female		Male		Female		Male	
	Child	Adult	Child	Adult	Child	Adult	Child	Adult
Resources								
Land								
Equipment								
Transport (motorized and nonmotorized)								
Animals								
Labor								
Production								
Social reproduction								
Community management								
Capital								

	Access				Control			
	Female		Male		Female		Male	
	Child	Adult	Child	Adult	Child	Adult	Child	Adult
Benefits								
Outside income								
Assets ownership								
In-kind goods (food, clothing, shelter, etc.)								
Education								
Political power, prestige								
Other								

Note: Items in italics were added by the present authors.

Source: Adapted by the authors from Overholt, Anderson, Cloud, & Austin (1985, Table 2).

APPENDIX 17.2 TOOLS FOR MORE IN-DEPTH GENDER-RESPONSIVE EVALUATION (GRE) DESIGNS

Approach	Method	Strengths	Limitations
1. **Experimental and quasi-experimental designs**			
• Post-project comparison designs	a. Sample survey comparing project and comparison group b. Can use propensity score matching to strengthen estimates	• Provides estimate of project impact • Can combine with mixed methods to strengthen estimates	• Risk of selection bias as there is no pretest data
• Natural experiments and pipeline designs	a. Delays due to natural or administrative factors used to compare projects with areas where project has not yet started	• Can provide estimate of project impact that may have bias, but which can be strengthened through mixed methods	• Selection bias • Needs agile evaluation team to be able to detect areas where delays are occurring
• Reconstructing baseline data	a. Baseline can be "reconstructed" using secondary data, key informants, recall, and PRA	• Valuable tool to strengthen evaluation designs	• Sources have potential bias • Many data sources do not have reliable gender data
2. **Statistical designs**			
• Econometric analysis • Public expenditure incidence analysis • Social exclusion	a. Analysis of usually country-level survey data. Often this will be disaggregated into income deciles or quintiles. b. Social exclusion analysis combines multiple data sets to compare access of different demographic groups to public services		
3. **Theory-based methods**			
• Contribution analysis	a. Develop "program story" and collect evidence to support and challenge	• Can estimate contribution of project to outcomes when attribution analysis not possible	• Often rival hypotheses are not identified and tested
• Outcome harvesting	a. At end of project, beneficiaries and other stakeholders are asked to identify most important outcomes during project	• Participatory approach capturing beneficiary perspectives • Larger numbers of outcomes captured, providing multiple perspectives	• May only capture positive outcomes

Approach	Method	Strengths	Limitations
• Concept mapping	a. Experts or stakeholders identify key outcome indicators, which are converted into rating scales b. Scales can assess changes over project life or compare project and comparison groups at end of project	• Helps develop broad-based indicators of project performance • When used on-line can provide economical way to measure project outcomes • Can involve wide range of gender specialists in indicator development	• Requires high level of specialist input and can be difficult to coordinate
• Realist evaluation	a. Asks what works, for whom, in what ways, to what extent, in what contexts, and how? b. Focuses on reasoning of the actors. How do they respond to interventions?	• Provides much broader understanding than just asking "did it work?" • Tries to explain processes of reasoning and behavioral change • Looks upon actors as proactive and not just reactive	
4. **Case-based methods**			
• Exploratory and descriptive	a. Describes projects, processes, and participant attitudes b. Illustrates findings of QUANT studies	• Puts flesh on the survey statistics • Compares lived experience of women and men • Explains context	• Cases often used for advocacy rather than objective reporting • Cases often not representative • Analysis often superficial
• Analytical (QCA)	a. Matrix created with attributes of subjects and outcome b. Identifies configuration of attributes needed to achieve intended outcome c. Uses mixed methods	• Used with small samples • Addresses complexity • Permits attribution analysis • Can combine with other kinds of evaluation	• Only permits small number of attributes in analysis • Attributes must be dichotomous (Yes/No)
5. **Broader applications of qualitative methods**			
• Initial diagnostic studies	a. Spending time in the community or project area prior to project launch	• Understanding the beneficiary perspective and cultural dynamics • Detecting issues the project design may overlook	• Difficult to commission studies before official project launch • Needs ethnographic training

(Continued)

(Continued)

Approach	Method	Strengths	Limitations
• Key informant panel studies	a. Researcher develops friendship with different types of people who are visited periodically to obtain updates on what the community thinks about the project, what they are hearing, and what is happening	• Independent feedback to avoid only getting information from project staff • Identifies unintended outcomes • Identifies vulnerable groups and those excluded	• Risk of bias if researcher mainly develops contacts with particular kinds of people and does not develop relationships with others
• Participatory group consultation methods	a. PRA, most significant change, and other group consultation techniques used to develop social maps, historical timelines, power analysis, and perceptions of causality	• Visual and mapping methods work well with groups with low literacy • Participatory methods give voice to vulnerable groups, including women	• Can be manipulated by researcher (intentionally or unintentionally) • Often used to get quick community feedback without due attention to the methodology
• Storytelling and sense-making	a. Individuals narrate short stories about events in the community; may focus on project or be open b. Sense-making software used to analyze the stories	• Gives voice to beneficiaries and counterbalance to funder's evaluation criteria • Broadens the focus	• Many stories are brief and superficial • Only positive stories • Need to control for bias introduced by facilitator
6. **Review and synthesis**			
• Meta-analysis, narrative synthesis, realist synthesis			
7. **New information technology**			
• Access to existing big data (satellite and remote sensors, social media, electronic transaction data) and ICT (mobile phones) • Smart data analytics • Social media analysis • Creation of integrated data platforms	a. Combining different estimates of poverty and other key social outcomes from satellites, ATM withdrawals, Twitter trends, etc. b. Analysis of social media for social network analysis, real-time warnings on potential problems (conflict, hunger, disease)	• Large volumes of data easily available • Can work with total population, not just small sample • Permits real-time feedback • Access to more sophisticated predictive analytics • Social media permit analysis of behavior and sentiments • Potentially gives voice to vulnerable groups	• Sample selection bias • Can promote top-down "extractive" approaches • Algorithms may include unintended (or intended) racial and economic biases • Focus on correlations but not explanation of how changes are produced

Approach	Method	Strengths	Limitations
8. **Systems and complexity approaches**			
• Systems mapping	a. Visual representation of system within which project operates b. Identifies linkages among components and external factors	• Visualizes interactions among all elements of systems • Helps describe systems of social control that limit gender outcomes	• Difficult to quantify and analyze processes of change
• Social network analysis	a. Analysis and mapping of processes of communication, influence, and power within an organization or community	• Can compare women's and men's communication networks, positions in power structures, and social capital	• Requires fairly large sample and sophisticated data analysis
• System dynamics	a. A map representing stocks and flows among project/systems components. Estimates how complex systems respond to project interventions.	• Helps assess the effectiveness of different project interventions on different parts of a system • Can be used to identify factors limiting the effectiveness of project interventions	• May require more complex measurement and analysis
• Critical systems heuristics	a. Analysis of the factors that determine what issues lie within the boundaries of the evaluation b. Studies how values affect the scope and focus of an evaluation	• Helps understand how feminist values can be incorporated into an evaluation	• May appear very abstract and theoretical to clients • Many clients believe evaluations should be value-free, so difficult to accept premises of this approach

APPENDIX 17.3 TWO EXAMPLES OF A WOMEN'S EMPOWERMENT INDEX

Example 1: An index for assessing the impacts of microcredit on women's empowerment in Bangladesh

In order to assess the impacts of participation in a microcredit program on women's empowerment in Bangladesh, Hashemi, Schuler, and Riley (1996) spent several months in a sample of rural communities, talking to women and observing their behavior to understand their concerns and what they saw as the realistic possibilities to strengthen their empowerment. Combining these conversations and observations with the results of surveys, they identified eight dimensions that women viewed as contributing to their empowerment.

Dimension	Measures	Scoring
1. Mobility	Places visited: market, clinic, movies, outside the village	1 point for each place visited plus one point if visited alone. 3+ = empowered
2. Economic security	Things owned: house or homestead land, productive asset, savings. Extra point if saving used for a business or money-lending.	2+ = empowered
3. Ability to make small purchases	1 point each for purchasing: items for daily use, items for self, ice cream or candies for children; + 1 point for items purchased without husband's permission and 1 point if purchased with money earned by the respondent.	7+ = empowered
4. Ability to make larger purchases	1 point for pots and pans, 2 points for children's clothing, 3 points for saris for oneself, and 4 points for buying family's daily food. An extra point if purchased with money earned by the woman.	5+ = empowered
5. Involvement in major household decisions	1 point for a major decision alone or with husband on household repairs, 1 point for decision to purchase a goat, 3 points for deciding to lease land, and 4 points for purchase of land, a boat, or bicycle. An additional point when purchase made with money earned by the woman.	2+ = empowered
6. Relative freedom from domination by the family	Respondent asked if within the past year money had been taken from her against her will; land, jewelry, or livestock had been taken against her will; she had been forbidden from visiting her natal home; or she had been prevented from working outside the house.	Woman defined as empowered if none of these things had happened during the past year.
7. Political and legal awareness	One point each for knowing the name of a local government official, a member of parliament, or the prime minister. One point each for knowing the significance of registering a marriage and for knowing the laws governing inheritance.	4+ = empowered
8. Participating in public protests and political campaigning	Respondent classified as empowered if she had campaigned for a political candidate, or had gotten together with others to protest a man beating his wife, a man divorcing or abandoning his wife, unfair wages, unfair prices, misappropriation of relief goods, or "high-handedness" of police or government officials.	Participation in any activity = empowered
Composite empowerment score	A woman is classified as "empowered" if she had a positive empowered score on five out of the eight variables.	

Source: Adapted from Hashemi, Schuler, & Riley (1996).

Example 2: Framework for Developing a Gender Empowerment Index for a Village Development Project in Central Asia

Empowerment Dimension	Examples of Indicators	Computing Empowerment Score
A. Direct outcomes (project objectives)		
1. Participation in decision-making and leadership roles	a. Number of groups in which participates b. Number of groups in which assumes leadership positions c. Number of groups in which participates in decisions d. Number of infrastructure projects responding to women's needs	a. Points will be given for each positive response b. Points will be totaled for this dimension c. The precise scoring procedure will be based on an analysis of the frequency distributions d. Recognized that this is an ordinal and not an interval scale
2. Participation in, and access to, resources from private group enterprises	a. Number of private group enterprises in which member participates b. Leadership roles c. Amount borrowed	Same procedure as for dimension 1
Direct outcomes composite score		Total for dimensions 1 and 2
B. Secondary outcomes (not project objectives)		
3. Participation in decision-making groups and organizations not part of project	Similar to 1 a–c	a. Similar to dimension 1
4. Access to and control of productive resources	a. Access to household or community productive resources (equipment, animals, land)	a. A number of items will be identified and points assigned to each one b. Points will be summed
5. Women's position strengthened within the household	Adapted from Bangladesh empowerment index dimensions 1–6	
6. Increased income and economic opportunities	a. Labor market income b. Greater access to labor market c. Income from enterprise or sale of produce d. Income from rent or money-lending	a. Incomes will probably be transformed into ordinal categories b. Access to labor market will be transformed into ordinal categories c. Points will be summed

(Continued)

(Continued)

Empowerment Dimension	Examples of Indicators	Computing Empowerment Score
B. Secondary outcomes (not project objectives)		
7. Social development	a. Number of people known and with whom interacts (social capital) b. Increased mobility c. Political and legal awareness d. Participation in political protests e. Access to cell phone and internet f. Reduced time burden	a. Each item will include several indicators; some will draw on the Bangladesh index b. A score will be calculated for each item c. Scores will be summed to provide a score for this dimension
Secondary outcomes composite score		Sum of scores for dimensions 3–7
TOTAL EMPOWERMENT COMPOSITE SCORE		Sum of two composite scores

Source: Bamberger (2017, Chapter 7).

CHAPTER 17 APPENDICES PART 2

General Approaches and Methodologies for Gender Analysis

APPENDIX 17.4

The World Bank/IEG Implementation Completion Report: Gender Flag Review

The ICR "Gender Flag" has two objectives: to systematically document the presence of gender dimensions in individual World Bank projects and to create incentives to Implementation Completion Report (ICR) authors to report on gender. The drop-down menu includes five main questions:

1. Is gender a relevant aspect of the project development objective?
2. Does the ICR include sex-disaggregated female or male-specific indicators?
3. Are there indicators that could have been sex disaggregated and were not?
4. Does the ICR discuss specific gender issues?
5. Please comment on any other issues regarding gender features of the ICR.

Source: Bardasi & Garcia (2016, Appendix 3).

APPENDIX 17.5

Recommended Structure for the Evaluation Cooperation Group GRE Reports

1. **History and context.** What are the historical events or approaches to gender that influenced the design of the project? What are the major issues the project should address?
2. **Relevance.** How relevant is the project design for addressing important gender issues?
3. **Efficiency.** How efficiently was the project organized to address the gender issues? Were there other approaches that could have achieved the same gender objectives in a more cost-effective way?
4. **Efficacy/effectiveness.** How successful was the project in achieving its short-, medium-, and long-term gender objectives? Were there any serious unintended gender outcomes? Could they have been avoided or were they due to factors beyond the project's control?
5. **Sustainability and resilience.** What evidence is there that the gender outcomes and impacts will be sustained over time? Have potential negative reactions to women's empowerment ("pushback") been taken into consideration? Did the project strengthen the ability of women and implementing agencies to identify and address gender-related shocks and stresses?
6. **Gender-related lessons learned.** Did the project include gender-related learning and dissemination mechanisms, and were they used effectively? What lessons were learned concerning the selection, design, implementation, and sustainability of gender-responsive evaluations?

Source: Bamberger (2017).

APPENDIX 17.6

Widely Used Gender Indices and Checklists

- **The Africa Gender Equality Index.** The index covers three dimensions: (1) equality in economic opportunities, (2) equality in human development, and (3) equality in law and institutions.
- **The SDGs.** Sets of indicators have been proposed for each of the 17 SDGs. For many of the goals, more detailed sets of indicators have been proposed by specialized agencies (e.g., UN Women, UN Habitat) even though these are not included in the official list of SDG indicators.
- **The Gender-related Development Index (GDI).** This adjusts the Human Development Index to take into consideration gender inequalities with respect to life expectancy, education, and income.
- **Gender Empowerment Measure (GEM).** This compares relative female and male representation in political and economic power. While it is a useful starting point, it should be used with caution as it is only based on a small number of indicators.
- **Social Watch's Gender Equity Index (GEI).** This seeks to address the limitations of GDI and GEM by incorporating more indicators on three dimensions: education, participation in the economy, and empowerment (measured by the percentage of women in professional, technical, managerial, and administrative jobs and the number of seats women have in parliament and in decision-making ministerial posts).
- **The World Economic Forum's Gender Gap Index (GGI).** This combines a range of indicators covering economic participation, economic opportunity, political empowerment, educational attainment, and health and well-being.
- **Regional indices.** A large number of regional indices have been developed that seek to capture the unique characteristics of different regions (e.g., the African Women's Progress Scoreboard [AWPS]).
- **Country-level gender indicators.** Following the launch of the MDGs in 2000, many countries have developed very extensive national databases, although these vary in their coverage of gender indicators.
- **Thematic indicators.** Many sector and thematic gender indices have also been developed by sector agencies such as FAO and UNESCO.

Links to the gender indices mentioned above
1. Africa Gender Equality Index
https://www.afdb.org/en/topics-and-sectors/topics/quality-assurance-results/gender-equality-index/
2. SDG Indicators
https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf
3. The Gender-Related Development Index (GDI)
http://hdr.undp.org/en/content/gender-development-index-gdi
4. Gender Empowerment Measure (GEM)
https://en.wikipedia.org/wiki/Gender_Empowerment_Measure
5. Social Watch Gender Equity Index (GGI)
http://www.socialwatch.org/taxonomy/term/527

Links to the gender indices mentioned above
6. World Economic Forum Gender Gap Index (GGI)
https://www.weforum.org/reports/the-global-gender-gap-report-2016
7. Africa Women's Progress Scorecard
http://www.uneca.org/sites/default/files/PublicationFiles/agdi_2011_eng_fin.pdf
8. Thematic Indicators
• **FAO**
http://www.fao.org/fileadmin/templates/ess/ess_test_folder/Workshops_Events/AFCAS_19/AFCAS_05_7_2_b.pdf
• **UNESCO**
http://uis.unesco.org/en/glossary-term/gender-parity-index-gpi

APPENDIX 17.7 EVALUATION APPROACHES USED IN STANDARD GRE DESIGNS

Note: It is recommended that a mixed-method approach should be integrated into all of the GRE designs. Much of the gender-specific data is difficult to collect and interpret, and the different perspectives and interpretations provided by triangulation are valuable. Where possible, the evaluation should be based on a gender analysis framework (see Chapter 17, Box 17.3 of the book).

Approach	Method	Strengths	Limitations	Tips
A. Desk review of project documents, secondary sources, and systematic reviews	a. Review of gender objectives and indicators in project design, implementation, and M&E b. How concepts such as empowerment are used c. Review of reports by other agencies d. Findings and lessons from previous projects	• Ensures the evaluation is addressing project objectives • Detailed information available for projects with a gender objective • Systematic reviews ground the evaluation in what kinds of outcomes can realistically be expected	• For projects without gender objectives, very little information is available • Gender objectives often only address a narrow range of direct outcomes ignoring secondary and tertiary outcomes • Many sources only focus on women	• Check if project teams have any reports not included in the project files • Apply the evaluation design matrix (see Appendix 17.8)
B. Theory of change (TOC)	a. A TOC is often not developed during project design and must be reconstructed by the Independent Evaluation Office (IEO) b. Broaden TOC to model behavioral change, emergence, and factors limiting intended gender outcomes (social control)	• Provides a framework for structuring the evaluation (identifying outcomes, key assumptions, intended processes of change) • Can also identify a counterfactual (rival hypothesis) to assess the contribution of observed outcomes	• TOC should be developed with stakeholders but this is time-consuming and difficult to arrange • TOC developed by consultants may not reflect stakeholder perspectives • Rival hypotheses often not defined	• Include a timeline over which sustainable gender transformation should be measured • Define the steps in the processes of empowerment and transformation • Ensure flexibility to model the emergence and backlash resulting from women's empowerment • Anticipate unintended outcomes

Approach	Method	Strengths	Limitations	Tips
C. Sex disaggregation of key indicators	a. Collect available sex-disaggregated data and identify sources when sex-disaggregated data was collected but has not been analyzed	• Quick way to identify sex differences in access to project services or outcomes and identify areas for further investigation • Useful for demonstrating gender differences to operations staff	• Disaggregated data may not be available or is expensive to extract • Data on sex may not be accurate or complete • Meeting attendance data may overestimate women's level of involvement	• Sex-disaggregated data may not be complete or reliable
E. Focus group discussion (FGD)	a. Groups of 6–10 people are interviewed together; usually members share some characteristics (age, sex, participation in the project), but groups can include different perspectives b. Information is obtained from each participant on every topic	• Economical way to cover all sectors of the target population • Group interactions create synergy and elicit information that may not surface in individual interviews • Participants may be willing to speak more freely in a group setting	• Due to time pressure, member selection may not be carefully controlled • Groups may be dominated by a few influential people or by someone nominated by government • The facilitator may influence the discussion, encouraging people to respond in a particular way	• Important to avoid selection bias or the group being dominated by a few influential participants
F. Case studies	a. In-depth study of certain individuals or groups to illustrate or explain survey findings b. Focus on process, personal experience, and behavior.	• Well-presented cases create greater impact than statistics • Help understand different ways people respond to the project	• Cases may be "cherry-picked" to find quotes or examples that may not be representative • Cases are often selected in an ad-hoc way and do not cover all sectors of the population	• Ensure selection, design, and analysis are coordinated with other parts of the evaluation • Ensure cases are comparable with other parts of the evaluation • For GRE include both women and men from the extended household and, where appropriate, community organizations such as the church

(Continued)

(Continued)

Approach	Method	Strengths	Limitations	Tips
G. Site visits	a. Short (several hours to 1–2 days) visits to project locations to see the project in action and to meet with staff, partners, and beneficiaries	• Useful to validate information obtained from project staff and government officials • Better understanding of the project reality than written reports, which may be incomplete or biased	• Implementing agencies may only arrange visits to the best projects • Difficult to meet with critics or families with complaints • Difficult to meet with women	• Avoid, or at least be aware of, bias in how communities to be visited are selected by the local agency • Avoid only meeting beneficiaries and project agencies; try to meet non-project informants who can give a different perspective • NGOs may not be objective informants if they are contracted to implement parts of the project
H. Beneficiary and other household surveys	a. Survey covering beneficiaries (and nonbeneficiaries) with questions on attitudes and experiences with project b. Can use structured quantitative surveys or more open qualitative interviews	• Representative sample of the project population • Can collect better information than relying on project records and secondary data	• Time-consuming and expensive • Often does not include a comparison group • Difficult to collect many kinds of gender information from QUANT surveys	• It is often possible to conduct an economical and rapid survey using student teachers, nurses, or university students. This is useful to collect basic information on, for example, who knows about the project and who does and does not participate. Women must be interviewed in situations where they can speak openly.

APPENDIX 17.8 EXAMPLES OF THE APPLICATION OF THE DIFFERENT GRE DESIGNS

Design	Variations	Examples/References
1. Experimental and quasi-experimental	RCTs, quasi-experimental designs, natural experiments	• Using RCT to evaluate the impacts of training of cross-border guards in Rwanda to reduce violence against women and improve socioeconomic outcomes for women. (*Source:* World Bank Gender Innovation Lab) • Many of the RCTs conducted by the Poverty Action Lab assess the impact of development interventions on women. (www.povertyactionlab.org)
2. Statistical	Econometrics, public expenditure incidence analysis, public expenditure tracking	• Public expenditure incidence analysis used to assess what proportion of public expenditures in sectors such as health and education go to low-income families including female-headed households. (*Source:* Davoodi, Tiongson, & Asawanuchit, 2003)
3. Theory-based	Theory of change, process tracing, contribution analysis, realist evaluation	• Using theory of change and contribution analysis to assess the effectiveness of a 10-year OXFAM program to reduce violence against women in El Salvador. (*Source:* Davis & Guevara, 2016)
4. Case-based	Naturalistic, grounded theory, ethnography, process tracing, QCA, within-case analysis, simulations, network analysis	• QCA used to assess the effectiveness of UN Women interventions at the national level on women's economic empowerment. The country was used as the unit of analysis. (*Source:* UN Women, 2014)
5. Participatory and qualitative	Empowerment evaluation, feminist evaluation, PRA, most significant change, outcome harvesting, outcome mapping	• Village women design a survey instrument to identify family needs in poor communities in India and then interpret and disseminate the findings. (*Source:* World Bank Social Observatory, India)
6. Review and synthesis	Meta-analysis, narrative synthesis, realist synthesis	• Using a systematic review, covering all of the published literature, to assess the impacts of microcredit on women's economic empowerment. (*Source:* Vaessen, Rivas, & Leeuw, 2016)
7. New information technology (NIT)	Twitter and social media analysis, satellite images, ATM transactions, phone records, analysis of audio and video images	• Tracking trends in gender-based hostility in factories in Indonesia (UN Global Pulse) • Tracking effectiveness of online messaging to promote girls' empowerment
8. Complexity-responsive evaluation	Applying the complexity-responsive evaluation methods described in Chapter 16. See also Appendix 17.9.	• The case study of the evaluation of a 10-year program to reduce violence against women illustrates the different dimensions of complexity that must be addressed in many GRE evaluations. (*Source:* Davis & Guevara, 2016)

Sources: Adapted from Stern et al. (2012) and Bamberger, Vaessen, & Raimondo (2016).

APPENDIX 17.9 EVALUATING THE GENDER DIMENSIONS OF COMPLEX DEVELOPMENT PROGRAMS

> *Development programs that seek to transform gender norms, enhance equality between women and men, and end violence against women are inherently complex. Their evaluations thus need to make sense of this complexity in order to establish whether they contribute to intricate change processes. Moreover, all development programs with gender dimensions, and complexity-responsive evaluations, should be gender-responsive.* (Raimondo & Bamberger, 2016)

As development initiatives become more complex, conventional evaluation approaches are no longer able to fully evaluate how multiple interventions funded, designed, and implemented by multiple stakeholders, and operating in complex environments, contribute to observed changes in multiple (intended and unintended) outcomes. Under these increasingly common scenarios, it becomes necessary to find new evaluation approaches that are "complexity-responsive" and equity-focused and gender responsive. In this appendix we discuss how the five-step complexity-responsive evaluation process described in Chapter 16 can be applied assessing the gender outcomes of development interventions. The framework can be used both to assess the extent to which defined gender objectives have been achieved (and the factors that constrain the achievement of the gender objectives), and to assess gender outcomes of all programs, including those that do not have defined gender objectives.

Step 1: Understanding the context in which development interventions take place and the factors affecting gender outcomes (holistic analysis)

The evaluation begins with a holistic analysis to understand the multiple systems, including mechanisms of social control, within which the program or other intervention operates. This builds on the four dimensions of complexity discussed in Chapter 16:

- The gender dimensions of the intervention.
- The organizational and institutional framework, the arrangements for achieving the gender objectives, and the interactions among stakeholders that affect, directly or indirectly, gender outcomes.
- The contextual factors affecting how the program is designed, implemented, and evaluated. This includes analysis of the mechanisms of social control that both affect the achievement of gender objectives and influence how the programs affect the status of women.
- The complex causal pathways between inputs and gender outcomes.

The concept of boundaries is important. Boundaries define how broad program effects are intended to be and how widely effects will be assessed (these are two separate but related issues). For example, is a program (such as a girls' scholarship program) only designed to benefit girls and the target villages or districts, or it is designed to have spill-over effects in surrounding areas? There are similar decisions to be made with respect to the evaluation. Is the evaluation only designed to assess direct effects or also spill-over effects? Boundaries must also be assessed with respect to time horizons. Will effects only be assessed over one year or over longer periods of time? The narrower the boundaries, the more economical and precise the evaluation. However, there is a trade-off, as potentially important secondary effects (both positive and negative) may not be captured.

Figure A17.9-1 identifies some of the complexity science tools that can be used in this holistic analysis. This analysis determines whether or not the program can be considered sufficiently complex to justify a complexity-responsive evaluation, and what are the main elements of complexity that must be addressed in the evaluation. The checklist given in Chapter 16 illustrates a useful approach for rating the level of complexity of the four dimensions. For the purposes of GRE, the indicators will need to be made more gender specific.

Step 2: Unpacking the program into its main components

The program is "unpacked" into its main components or elements, each of which can be evaluated separately. The unpacking will identify the potential direct and indirect gender outcomes that are planned, or that may potentially occur. A big advantage of this approach is that it is possible to use conventional evaluation designs, as well as more specialized GRE tools, to evaluate the individual components (whereas these designs do not generally work to evaluate the whole complex program).

Step 3: Selecting the appropriate evaluation methodology

All of the standard and in-depth GRE designs discussed in Sections 3 and 4 of Chapter 17 can be used as appropriate. Six conventional evaluation designs (discussed earlier in Chapter 4) can be used, where appropriate, for these evaluations.

Step 4: Reassembling the findings of the individual component evaluations

This is a very important phase, as there are many situations in which each individual component is evaluated positively, but where the whole program makes little contribution to its overall gender objectives. There are at least three reasons for this:

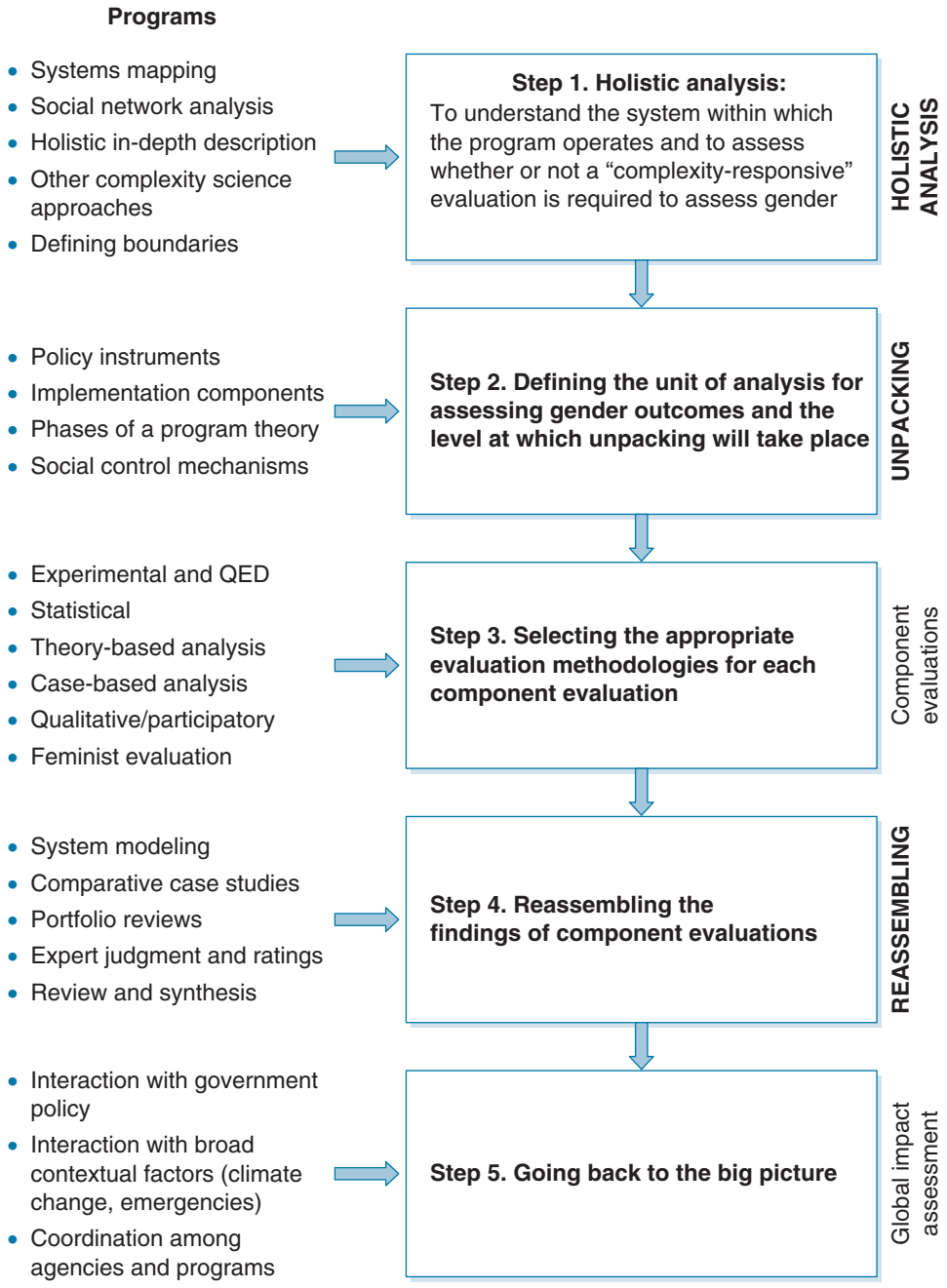
- The program goal is over-ambitious: for example, a training and awareness-raising program for women entrepreneurs may to be too small and limited to address the multiple social, economic, political, legal, and cultural barriers to women's empowerment.
- There may be problems of coordination among the different agencies and program components.
- Unforeseen events, such as a drought, civil war, or change in international markets, may seriously limit the program's effects.

Figure A17.9-1 lists some of the methodologies that can be used for this reassembling analysis.

Step 5: Going back to the "big picture"

Finally, the program and its effects must be assessed within the context of government policies and other programs, the effect of major contextual factors, and the challenges of coordination among the different actors and the programs they manage.

FIGURE A17.9-1 ● A Five-Step Approach for the Evaluation of the Gender Outcomes of Complex Programs



APPENDIX 17.10 THE CONTRIBUTION OF GENDER-RESPONSIVE BUDGETING (GRB) TO NATIONAL POLICY DIALOGUE

The purpose of GRB is twofold:

To ensure that governments incorporate their commitments to women into their budget policies.

To assess the differential impacts of budgetary policies to women and men and to identify any unintended (or intended) biases against women into how budget policies are designed and implemented.

GRB is rapidly expanding around the globe. It is estimated that there are now more than 80 countries that have incorporated some measure of gender budgeting into their budget planning. GRB can be introduced at multiple levels; for instance, integration into Public Finance Management Systems, application at sectoral and local levels, or identifying and financing women's priorities. It is an effective tool for gender policy advocacy across different levels and sectors.

Benefits of GRB

Countries are recognizing that GRB holds governments accountable for achieving their commitments to women but also increases the overall efficiency of budget policies. GRB uses standard budget planning and assessment tools (such as beneficiary expenditure incidence analysis), so it is easy for budget agencies to implement. At least during the early stages, GRB does not require the application of unfamiliar feminist concepts and analytical tools.

Examples of the Benefits of GRB

- A study in 2016 found that women contributed 85% of the savings to the UK government from austerity measures.
- The UK Women's Budget Group estimated that switching 2% of GDP from investment in construction to investment in health care would create 1.5 million jobs compared to 750,000.
- In Rwanda, increased investments in sanitation have significantly increased girls' school enrollment.
- In South Korea, investments in child care have significantly increased women's labor force participation—a critical benefit in a country with an aging population.

Gender budgeting has won the backing of international financial institutions who find this approach much more acceptable than discussions about women's empowerment. It is now accepted by major financial institutions such as the IMF, World Bank, and OECD. Moving forward, GRB faces both challenges and opportunities. Can it be an effective tool for promoting dialogue and action on the challenges addressed in recent women's marches and protests around the world? There may also be potential for GRB to be an advocacy tool or framework for promoting equity and social justice around other critical issues like race, immigration, LGBTQ issues, or homelessness.

Source: Adapted from Bamberger (2017).

APPENDIX 17.11 FEMINIST CRITICAL THEORY

Although not all identify themselves as critical theorists, many feminists have been influenced by critical theory, elements of which can be identified in many feminist evaluation approaches (Gannon & Davies, 2012; Brisolara, Seigart, & Sengupta, 2014). Understanding the approaches and issues raised by critical feminism is essential for understanding most feminist approaches to evaluation, as the fundamental critical theory concepts underpin, explicitly or implicitly, much of feminist thinking about evaluation.

Feminist Critical Theory. Provides a broader framework for thinking about the world and provides a healthy skepticism about many evaluation approaches and their claims to higher levels of rigor and objectivity, which are in fact just one of many ways to think about the programs and processes that are being studied.

Evaluation. Critical theory, within sociology, has its origins in the Frankfurt School, which included writers such as Adorno and Marcuse and is strongly influenced by Marx, the post-Marxists, and Freud. Horkheimer (1937/1972) argued that "critical theory is a social theory oriented towards critiquing and changing society as a whole, in contrast to traditional theory which is oriented only towards understanding and explaining it." Many of the early critical theorists were concerned to understand why society had not evolved as Marx had predicted and how to adapt post-Marxism to the reality of the early 20th century. Feminist critical theory began to evolve in the 1980s and was closely associated with post-modernism and post-structuralism, although there continue to be debates about the relationships between the three (Gannon & Davies, 2012). While feminist critical theory is not a cohesive theory and many elements continue to be widely debated, the writings of authors such as Gannon and Davies (2012), Podems (2010), Brisolara et al. (2014), Benhabib (1986), Kincheloe and McLaren (2003), and Lincoln and Denzin (2003) suggest there is some level of consensus on the key elements of feminist critical theory:

a. A central tenet of critical theory is the importance of being skeptical and of deeply questioning what is assumed to be true.

b. There is a need to reassess the concept of objectivity and to challenge the focus of post-positivism on prediction and causality

c. There is a focus on issues of domination and exploitation (Benhabib, 1986), and a political commitment to gender equality, describing the world in ways that have a greater resonance with women's lived experience.

d. It is recognized that power relations are established and maintained through discourse and through positions assumed in this discourse. An important implication is that agency and emancipation are contingent and limited.

e. Social problems should be politicized by "situating them in historical and cultural contexts, to implicate themselves in the process of collecting and analyzing data, and to relativize the findings" (Lindorf & Taylor, 2002, p. 52). Evaluation should be recognized as a political activity that either accepts or challenges dominant modes of discourse.

f. There is an attention to writing and discursive practices through which versions of the world are created, and to how language is used and to the implications of this usage. Many writers refer to "false consciousness" and how this is structured through the dominant forms of narrative and dialog.

g. Questioning whether it is appropriate to focus on the individual, as the analysis of power structures makes it more relevant to focus on "subjects" and forms of dominance and subjugation that affect broad categories of people.

h. Awareness of how and when one adopts binary categories (such as male and female) that limit perception, understanding, and imagination and the consequences with respect to exclusions and over-simplifications, particularly when categories become closely associated with one another (such as "male" and "rationality").

A goal of critical evaluation is to challenge unjust power structures and to promote social justice through political and social change. It is seen as promoting political activism as opposed to post-modernism and post-structuralism which are considered to be more reflective and sometimes opposed (according to critics) to activism. Lincoln and Denzin (2003) observe:

> "The critique and concern of the critical theorists has been an effort to design a pedagogy of resistance with communities of difference. The pedagogy of resistance, of taking back "voice," or reclaiming narrative for one's own rather than adapting to the narrative of a dominant majority . . . [aims at] overturning oppression and achieving social justice through empowerment of the marginalized, the poor, the nameless, the voiceless." (Lincoln & Denzin, 2003, pp. 625–626)

So these premises and approaches inform, to different degrees, many evaluations that use either a gender analytic or a feminist approach.

APPENDIX 17.12 GENDER-SENSITIVE DATA-COLLECTION METHODS AND APPLICATIONS USED IN THE CASE STUDIES IN APPENDIX 17.14 AND THE ADDITIONAL TIME AND COST IMPLICATIONS (COMPARED TO STANDARD EVALUATION METHODS)

Method	Applications	Ease and Cost of Data Collection and Analysis
A. **Quantitative methods**		
Household surveys	Household composition and household welfare. • ***See Case Studies 1, 3, 4, and 6***	Sex-disaggregated questions can be included at no cost. However, applying submodules to individual household members increases interview time and often requires a second visit.
Attitude studies	Analysis of attitudes toward different organizations or prioritization of needs and projects.	Include on household survey but requires some additional time to administer.
Willingness and capacity to pay	Estimates of how much households are currently paying for services such as health, water, education, and transport and assessment of willingness and capacity to pay for improved services for different household members (boys/girls, etc.).	Questions can be included in household survey but it is essential to interview both women and men. Qualitative methods such as direct observation may be required to check reliability of the information.
Time-use studies	Estimating the time women and men spend collecting water and fuel, traveling to work, domestic activities, and unpaid and paid productive activities. • ***See Case Studies 3 and 5***	Questions can be included in surveys, but where possible this should he combined with focus groups or direct observation. Household diaries can also be used.
B. **Qualitative methods**		
Stakeholder analysis	Identifying main groups affected by or affecting planned or actual policies and determining their interests, influence, and importance. • ***See Case Studies 4, 5, and 7***	This requires individual interviews but often with a relatively small number of respondents.
Institutional analysis	Evaluating the efficiency and client-friendliness of public- and private-sector agencies providing services to the poor. • ***See Case Study 7***	Some questions can be included in household surveys, but where possible these should be combined with focus groups.

Method	Applications	Ease and Cost of Data Collection and Analysis
Focus groups and community forums	Seeking the opinion of community groups on their problems and priority needs, and their experience with the projects and programs being provided. A valuable complement to household surveys. • ***See Case Studies 4 and 5***	Properly conducted focus groups require time to invite the right people and to write up the discussions. A team of two researchers is required. The sessions should be tape-recorded and this further increases the analysis time.
Participatory Rural Appraisal (PRA) and other participatory methods	These methods are used to understand the world of the poor and to listen to their concerns and priorities rather than asking them to respond to a set of survey questions prepared by outside agencies.	Several days and ideally at least one week should be allowed for each community studied. It is important to allow sufficient time to understand the community and to gain the trust of residents before the sessions begin.
Photographs and videos	Photographs provide a dramatic complement to written reports and an effective way to document physical and economic change over time (by taking photographs from the same location at different points in time.	Photographs are quick and easy to take. Videos are also an excellent way to present findings, but they are much more expensive to produce—particularly if editing is required.

CHAPTER 17 APPENDICES PART 3

Case Studies Illustrating Different Approaches to Gender Analysis

APPENDIX 17.13 SUMMARY OF DESIGN AND DATA-COLLECTION METHODS USED IN THE SEVEN CASE STUDIES DESCRIBED IN APPENDIX 17.14

Case Study	Study Design				
	Data-Collection Methods	Ex-Ante		Ex-Post	
		T	C	T	C
1. **Assessing the Gender Impacts of Microcredit Programs in Bangladesh** *Cross-sectoral comparisons of household surveys of borrowers and nonborrowers using statistical controls to adjust for sample selection bias*	1. Household survey 2. Community survey 3. Anthropometric study	No	No	Yes	Yes
2. **Who Takes the Credit? A Different Perspective on the Gender Impacts of Microcredit in Bangladesh** *A purposive sample of women borrowers from four credit programs using recall to assess women's participation in the decision making and control over the use of the loans*	1. Sample of survey of women borrowers 2. Intensive use of recall to review women's level of participation at each stage of loan approval and utilization	Ex-Post Recall	No	Yes	No
3. **Evaluating the Gender and Time-Use Impacts of the Cut-Flower Export Industry in Ecuador** *Quasi-experimental cross-sectoral design with special emphasis on time-use measurement*	1. Household survey 2. Time-use module	No	No	Yes	Yes
4. **Assessing the Gender Impacts of the Bangladesh Rural Roads and Markets Improvement and Maintenance Project** *A quasi-experimental design with ex-post data collection without a control group and with the use of recall to estimate the before-project conditions*	1. Household survey 2. Recall methods to estimate the ex-ante conditions	Ex-post recall	No	Yes	No

Case Study	Data-Collection Methods	Study Design			
		Ex-Ante		Ex-Post	
		T	C	T	C
5. **The Impact of Bicycles on Women's Literacy and Empowerment in India** *Follow-up study to assess the impacts of the introduction of bicycles five years earlier. Data was collected through key informant interviews, a focus group discussion, a village survey, and an activity and time profile with a sample of husbands and wives*	1. (Small) Household survey 2. Activity and time-use profile 3. Focus groups	No	No	Yes	No
6. **Assessing the Impacts on Girls' Educational Enrollment of Private Schools for the Poor in Balochistan, Pakistan** *A longitudinal impact evaluation with random selection of treatment communities and before and after comparisons of the project and control areas and of female and male students*	1. Household survey	Yes	No	Yes	No
7. **Assessing Gender Mainstreaming in the World Bank** *Two institutional assessment studies that combine ratings on the treatment of gender in World Bank projects*	1. Rating of project reports 2. Key informants			Yes	

Note: T = treatment group; C = control group

APPENDIX 17.14[30] CASE STUDIES ILLUSTRATING DIFFERENT GENDER IMPACT EVALUATION METHODOLOGIES

Case Study 1: Assessing the Gender Impacts of Microcredit Programs in Bangladesh

Cross-sectoral comparison of household surveys of borrowers and nonborrowers using statistical controls to adjust for sample selection bias

THE STUDY

The purpose of the study was to examine the gender-differentiated impacts of female and male borrowing from three microcredit programs in Bangladesh on a range of household welfare indicators including income and assets, nutrition, school enrollment, fertility behavior and contraceptive usage, and empowerment. The microcredit programs studied were the Grameen Bank, the Bangladesh Rural Advancement Committee (BRAC), and the Rural Development 12 (RD-12) project of the Bangladesh Rural Development Bank.

METHODOLOGY

The evaluation is based on a 1991–1992 Household Survey conducted by the Bangladesh Institute of Development Studies. The sample covered 29 randomly selected *thanas* from the 391 *thanas* in Bangladesh (with *thanas* affected by the 1991 cyclone being excluded). Twenty-four of the *thanas* had at least one of the three microcredit programs operating while five had none. Several *thanas* had more than one microcredit program operating but no household was a member of more than one. A total of 1,798 households were selected using stratified random sampling. Of these, 1,538 were *target* households (in communities with one of the microcredit programs operating), of whom 905 were participating in one of these programs. The remaining 260 were *nontarget* households.

A detailed household questionnaire covering income, employment, education, consumption, borrowing, asset ownership, savings, children's schooling, fertility behavior, and contraceptive use was administered to all households. For the 315 households included in the nutrition survey, anthropometric data was also collected. A village survey questionnaire was also administered to collect information on crop prices, fertilizers, wages for men, women and children, access to credit markets, and access to roads and public services.

Impact assessments were based on cross-sectional analysis comparing households who did and did not use microcredit programs with respect to the impact indicators (see Table A17.14-1). Econometric methods were used to correct for differences between target villages and non-target villages and between borrowers and nonborrowers with respect to attributes such as wealth, land-holding, and so on likely to be correlated with the impact indicators. The analysis found that target villages were on average wealthier than non-target villages, and adjustment for these differences reduced in many cases the magnitude of the estimated program impacts, although in most cases they remained significant.

FINDINGS OF THE STUDY

Two related studies examine the impact of female and male borrowing—from Grameen Bank, the Bangladesh Rural Advancements Committee (BRAC), and government program RD-12—on such outcomes as per capita household

[30]Adapted from Bamberger (2004).

expenditure (income and girls' and boys' schooling and nutritional status; S. Khandker, 1998; Pitt & Khandker, 1998). The impacts often differ substantially based on whether the borrower is a woman or a man—and often the marginal impacts of borrowing are greater for women than for men.

For all three microfinance programs, the impact of female borrowing on per capita household expenditure (income) is about twice as large as the impact of male borrowing (Table A17.14-1). A 10% increase in female borrowing is associated with a roughly 0.4% increase in per capita expenditure—an effect that is strongly significant statistically. Compare this with a roughly 0.2% increase in per capita expenditure associated with the same percentage increase in male borrowing. Female borrowing also has a greater impact than male borrowing on households' ability to "smooth" consumption over time (S. Khandker, 1998).

Women also benefit from program participation through the cash income generated by self-employment and the assets they acquire in the process. Estimates indicate that microfinance reduces poverty among program participants and reduces aggregate poverty in program villages (even after controlling for observable village characteristics that partially determine the extent of village poverty).

As with other forms of resource control, female borrowing also appears to have a greater impact on children's welfare than male borrowing does. For example, except for BRAC, female borrowing has a greater positive impact on children's school enrollments than male borrowing does. Moreover, in contrast to male borrowing, female borrowing has a large and statistically significant impact on children's nutritional well-being.

At the same time, male borrowing has a greater impact on household net worth than female borrowing. This suggests that while at the margin women seem to invest relatively more than men in the human capital of their children, men appear to invest more than women in physical capital.

Female and male borrowing also have different impacts on household reproductive behavior, suggesting that women and men do not share the same preferences relating to contraception or fertility. For example, female borrowing decreases contraceptive use and, except for Grameen Bank borrowing, increases fertility, whereas male borrowing increases contraceptive use and, except for BRAC borrowing, decreases fertility. At first glance the findings on the impact of female borrowing on contraceptive use may seem counterintuitive, since a body of empirical literature suggests that factors increasing the opportunity cost of women's time tend to reduce fertility. But low-income women in Bangladesh may see additional children as assets capable of assisting them with what are often home-based, self-employment activities.

Increasing women's access to credit also empowers them in other dimensions. For example, female borrowing increases female control of nonland assets (S. Khandker 1998; Pitt & Khandker, 1998).

LESSONS FOR THE EVALUATION OF GENDER IMPACTS

The study provides a good example of how one can plan survey design and data collection to study gender differences in program impacts. It also emphasizes the importance of assessing potential sample selection biases, and shows how this can be done through using econometric methods to control for differences in household characteristics such as income, labor force participation, education, and household size which may be correlated with the outcome (impact) indicators. It should, however, be pointed out this kind of cross-sectional analysis does not address many of the threats to validity of quasi-experimental designs (Valadez & Bamberger, 1994, Box 8.1) such as local history, political interference, interaction between projects, and local context. The design also does not address the specific problems of gender impact assessment discussed in Chapter 17 such as potential biases or omissions concerning information collected from, or about, women.

Acknowledgment: The findings section of this case study is taken directly from World Bank (2001), with additional material from S. Khandker (1998, p. 12).

Sources: Baker (2000, Annex 1.2, "Does Microfinance Really Help the Poor? New Evidence From Flagship Programs in Bangladesh"); Pitt & Khandker (1998); Pitt, Khandker, McKernan, & Latif (1999); World Bank (2001, pp. 160–162).

TABLE A17.14-1 ● Impacts of Female and Male Borrowing on Selected Household Outcomes in Bangladesh

(percentage change for a 10% increase in borrowing)						
	Grameen Bank		BRAC		RD-12	
Household outcome	Male borrowing	Female borrowing	Male borrowing	Female borrowing	Male borrowing	Female borrowing
Per capita spending	0.18	0.43	0.19	0.39*	0.23*	0.40*
Net worth	0.15*	0.14*	0.20*	0.09*	0.22*	0.02
Boys' school enrollment	0.07*	0.61*	−0.08	−.03	0.29	0.79
Girls' school enrollment	0.30	0.47*	0.24	0.12	0.07	0.23
Boys' height for age	−2.98	14.19*	−2.98	14.19	−2.98	14.19
Girls' height for age	−4.92	11.63*	−4.92	11.63*	−.492	11.63*
Contraceptive use	4.25*	−0.91*	0.40	−0.74*	0.84	−1.16
Recent fertility	−0.74*	−0.35	0.54	0.79*	−0.74*	0.50

*Indicates coefficient estimate that is statistically significant at the 10% level or better.

Source: S. Khandker (1998, cited in World Bank, 2001).

Case Study 2: Who Takes the Credit? A Different Perspective on the Gender Impacts of Microcredit in Bangladesh

A purposive sample of women borrowers from four credit programs using recall to assess women's participation in decision making and control over use of the loans

THE STUDY

The purpose of the study was to challenge the frequently stated assumption that women's obtaining and repaying loans is a good indicator of the role of microcredit in promoting women's empowerment. The study sought to estimate the degree of control which women actually exercised over the loans they obtained and the implications this has for a fuller understanding of empowerment.

METHODOLOGY

The evaluation design can be described as follows:

T(1) Intervention (X) T(2)

Project group (P) [P(1)] -------- X ---------- P(2)

Control group I

where

T(1) and T(2) = time periods before the projects began and after women had received loans respectively.

[P(1)] = baseline information re-created through recall.

A purposive sample was selected of women who had obtained loans from four microcredit programs in Bangladesh [N = 253]. The sample was selected to include a variety of group and loan characteristics, such as years of membership in the credit program and size of loan, and the marital status of the women. Loan histories were obtained on all of the women with a range of questions about women's control over the productive process. For example, women were asked what activity they invested in, where the inputs and productive assets came from and who procured them, what they cost, how they were put to use, where outputs were marketed, for what price, what were the problems involved in the productive process, and who the main user of the loan was in terms of labor input and in terms of controlling accounts and general management.

No control group was used because the study was not testing a hypothesis but rather focusing exclusively on women who had received loans.

On the basis of these questions an index of loan control was developed:

- FULL = full control over the entire productive process, including marketing.
- SIGNIFICANT = control over every aspect of the productive process with the sole exception of marketing.
- PARTIAL = loss of managerial control over the productive process, but the provision of substantial inputs of labor.
- VERY LIMITED = minimal input to the production process.
- NO INVOLVEMENT = cases where women provided no labor for activities which are culturally ascribed as masculine.

The study relied heavily on recall to obtain information on how the loans were managed and the authors stress the reliability issues inherent in this method.

FINDINGS OF THE STUDY

The study found the following percentages of control by the women borrowers:

- Full control 17.8%
- Significant control 19.4%
- Partial control 24.1%
- Very limited control 17.0%
- No control 21.7%

The initial conclusion is that women retain full control of less than 20% of the loans (17.8%) and at least significant control in less than 40% (37.2%) of the loans. These figures clearly indicated that borrowing and loan repayment cannot be automatically equated with women's empowerment without a fuller understanding of the dynamics of loan control within the household.

The authors emphasize that the figures must be interpreted within the social context of rural Bangladesh where it is virtually impossible for a woman to retain complete control over all stages of the productive process as social controls limit her geographical mobility and her ability to directly market goods she has produced. This is evidenced by the fact that almost all of the women who retained full control of the loans were either divorced or widowed. They also argue that in a context such as rural Bangladesh, where a woman's economic and social welfare and physical security are almost exclusively defined by her ability to maintain a satisfactory marriage, women must be expected to use a tool such as credit to strengthen their position in the household rather than to seek economic independence.

LESSONS FOR THE EVALUATION OF GENDER IMPACTS

The findings clearly demonstrate the need to broaden the range of indicators used in the evaluation of the impacts of microcredit on the welfare of women.

The technique of historical analysis, in which subjects provide detailed information on how the loan was obtained and managed, is shown to be a useful tool for studying the degree of women's participation at each stage of the loan process.

One potential weakness of the methodology is that the research relies exclusively on information provided by women. Within the cultural context of rural Bangladesh it may be difficult for women to speak freely, particularly with respect to issues such as the control of a loan which could be perceived as a criticism of her husband. Consequently there is some danger of bias in the information provided. The findings could have been strengthened through the use of triangulation whereby other independent sources would be consulted (such as other female household members, neighbors, or members of the credit banks) to provide a consistency check on the information.

Source: Goetz & Gupta (1996).

Case Study 3: Evaluating the Gender and Time-Use Impacts of the Cut-Flower Export Industry in Ecuador

Quasi-experimental cross-sectoral design with special emphasis on time-use measurement

THE STUDY

The purpose of the study was to understand the impacts of women's employment on the allocation of paid and unpaid labor within the household. The study compares household labor allocation in the areas affected by the cut-flower industry, which has a high demand and offers relatively high wages for female labor, with the situation in similar geographical areas that do not have access to this source of employment. While many studies on the time-use allocation impacts of female employment opportunities study female and male time-use decisions independently of each other, the purpose of this study is to understand how female labor market opportunities affect the time allocation between paid employment and unpaid domestic activities for all household members, and specifically how increased earning opportunities for women affect the time allocation of male household members. The cut-flower industry is particularly interesting for this purpose as women can earn as much as, or in some cases more than, males—thus avoiding the common situation where it is economically more rationale for women to devote most of their time to nonpaid domestic activities because men's earning capacity is much greater.

METHODOLOGY

The study uses a quasi-experimental cross-sectional design in which interviews were conducted in May through June 1999 with a sample of 562 households during which observations were obtained on all 2,567 household members over the age of 10. The sample included "treatment" households living in the valley where the cut-flower industry operated and "control" households living in a similar valley at a distance of some 200 km where there was no access to the cut-flower industry.

The data included modules on expenditures, economic activity (including agricultural activity and small businesses), health, education, fertility, credit, and savings. A special instrument was included to obtain detailed accounting of time use. Time-use accounting was obtained both for a 24-hour period (which is considered to be the most reliable) and also for the previous week. The latter was included on the assumption that men's contribution to housework would probably be irregular and could be seriously underestimated if information was only obtained on a 24-hour period. The 24-hour recall estimated the number of minutes dedicated to farm work, paid work, community work, household activities, recreation, and personal care (including sleep). Separate estimates were obtained for married and single men, and married and single women.

FINDINGS OF THE STUDY

The study produced some important, and in some cases unexpected, findings concerning the impacts of women's access to paid employment on the distribution of responsibility for household tasks:

1. All women, irrespective of their labor force participation, continue to shoulder the major burden of domestic chores. All women in the treatment group spent an average of 324 minutes per day on housework compared to an average of 59 minutes for men; the proportions are relatively similar for the control group.
2. When farm work, paid work, community work, and housework are combined, women in the treatment group worked an average of 144 minutes per day longer than men. For the control group the difference was even greater, namely 184 minutes.

3. However, in households where the wife worked, husbands, on average, increased the time they spent on housework. In the treatment group, for all households where the wife worked, the husband devoted an average of 60.36 minutes per day to housework compared to 40.0 minutes when the wife did not work (see Table A17.14-2).
4. Of particular interest is the fact that when the wife worked in the cut-flower industry a husband's time on housework increased to 76.88 minutes (compared to 60.36 minutes for all types of wife's work).
5. The cut-flower industry appears to be unique in that women can earn as much as men, and married women on average earn slightly more than men (6161 sucres compared to 5899 for men). This contrasts with the large gap in average earnings for other sectors where on average married men earn 5337 sucres compared to 2310 for women.
6. An important finding is that when women have paid employment, the hours worked on household chores declines. For married women working in the cut-flower industry, the number of hours per week devoted to domestic work is 24.23 compared to 32.25 for married women not working.

TABLE A17.14-2 ● Average Minutes per Day Spent on Household Tasks for Married Men and Women by Labor Market Status for Cut-Flower and Control Areas

	Flower Growing Area		Control Area
	Works anywhere	Works in cut-flower industry	Works anywhere
Male: Married household head			
Works	57.41	68.79	30.85
And wife works	60.36	76.88	33.06
And wife does not work	40.00	36.46	18.21
Female: Wife of household head			
Works	292.32	229.72	358.15
And husband works	288.32	221.55	359.38
And husband does not work	444.17	253.33	310.00

Source: Summary of Newman (2001, Table 3).

LESSONS FOR THE EVALUATION OF GENDER IMPACTS

Several lessons can be drawn from this case.

1. The selection of a control group is an extremely important element of the design, and great care must be taken to ensure that the characteristics of the control group are as close as possible to the treatment group on the key variables being studied, while at the same time ensuring that it is not affected by the treatment variable.

2. Careful measurement of time use is critical to this and many other gender impact evaluations. This study illustrates the importance of combining an in-depth analysis of a 24-hour period (when recall is likely to be more reliable), with the study of time use over the previous week. This is particularly important for the analysis of male contribution to household chores that is not done on a regular basis. For many studies it is also necessary to cover longer periods of up to a year in order to capture seasonal variations. This is particularly important for the analysis of labor demands during different periods of the agricultural cycle.

3. The study emphasizes the importance for gender analysis of combining aggregate household-level data (for example, on time use and income) with information on individual household members.

4. Even with a carefully selected control group, the cross-sectional design has a fundamental limitation in that it normally does not permit the analysis of changes over time. In the present case study, we do not know how domestic chores were allocated between the husband and wife before the advent of the cut-flower industry. Perhaps these were households with certain special characteristics that meant that men were already assuming a relatively high share of household chores, so that the impact of the cut-flower industry was less than the analysis suggests.

5. This is not to take away from the great value of this kind of cross-sectoral design—particularly as there are many situations where for reasons of cost or time it is not possible to use longitudinal designs. However, the ability to estimate change can often be increased by the use of recall methods, key informants, and in some cases secondary data in order to reconstruct the situation before the intervention which is being studied. For example, spouses could be asked to estimate time-use patterns before the advent of the cut-flower industry. It is often possible to use triangulation methods to assess the reliability of the information given on the before-intervention situation. For example, if men and women (when interviewed separately) provide consistent information, then one can have greater confidence. Information can also be cross-checked through interviews with groups of women and men to discern general patterns of change.

Source: Newman (2001).

Case Study 4: Assessing the Gender Impacts of the Bangladesh Rural Roads and Markets Improvement and Maintenance Project

A quasi-experimental design with ex-post data collection without a control group and with the use of recall to estimate the before project conditions

THE STUDY

The purpose of the study was to evaluate the social and economic impacts of the Second Bangladesh Rural Roads and Markets Improvement and Maintenance Project. A gender impact evaluation was built into this broader study. The overall development objective of the project is to help increase rural employment and incomes and reduce rural poverty by establishing improved, sustainable rural transport and trading infrastructure. More specifically, the project objectives are to help remove physical bottlenecks, improve quality, and reduce costs in rural transport and marketing; create employment and income-generating opportunities among the rural poor, and particularly for disadvantaged women; promote participation of local communities and NGOs in project activities; and increase institutional capacity for efficient rural infrastructure management, including maintenance. The project covers 14 districts in the Dhaka and Rajshahi regions of Bangladesh.

The project incorporates several components to promote the participation of women in project design and in project benefits. These include the promotion of community and user participation (including women) in planning, design, and implementation of road, market, and ghat improvement, as well as construction of culverts and bridges; and a number of "*Lady's Corners*" to provide a legitimized space for women vendors to sell their goods. Traditionally, no women sellers were allowed to sell within the market. They were requested to sell their products at a lower rate in various locations outside the market area. Now, the women vendors are provided with a place within the improved marketplace specifically designed and allocated for them.

METHODOLOGY

The study was based on ex-post data collection once the project was operational. The before project conditions were estimated through individual and group interviews in which beneficiaries were asked to recall their situation before the project began. Triangulation methods were used to check the consistency of the estimates through available secondary data and interviews with key informants. The methodology blended both qualitative and quantitative data with focus on feedback from discussions with the women beneficiaries and relevant stakeholders.

Sample design: The sample covers 8 *thanas* out of the 40 in which the project was being implemented. The sample included the following:

- All of the women beneficiaries of the rural roads (N = 115) from the eight *thanas* were interviewed.
- Thirty focus group discussions (FGDs) were conducted for *rural roads*, with three FGDs per road comprising six participants; one FGD with male participants (villagers/road users); one with only female participants (women laborers of the project); and another comprising both male and female participants (villagers/road users) in equal numbers.
- Six FGDs were conducted for *rural markets.* These covered two markets with three FGDs per market comprising six beneficiaries; one FGD with male participants (men at the markets) and two with female participants (one with women vendors and another with women vendors and sellers).
- Twenty in-depth interviews were carried out with the local government, villagers, and local leaders.

Data collection covered socioeconomic characteristics of the women participants/beneficiaries including age, education, family size, school-going children, sources of earning, farm land owned/rented, housing, ownership of livestock/poultry, and affordability of food, clothing, and medical expenses. Women were asked to compare their condition before and after the project for all of these variables.

FINDINGS OF THE STUDY

The findings indicate that the project has made an overall positive impact on the social and economic conditions of the women and their households with respect to education, income, housing, agricultural capital, food security, access to health services, acquisition of consumer durables, and ability to save and to obtain credit.

The only important area in which the project did not achieve its intended impacts concerned the strengthening of social capital through increased participation in groups and *samities.* The study data indicates that most women discontinued their involvement in groups and *samities* as they had very little spare time after their daily tasks at home and in the workplace.

Benefits Received by People in General From Improvements in Rural Roads Under the Project

In addition to the direct benefits to women discussed above, the survey also identified the following benefits enjoyed by all families in the project areas: (i) quicker and safer movement even at night (28%); (ii) improvement in trading/setting up new shops/industries (16%); (iii) safer movement of schoolchildren/expansion in education/setting of more schools (15%); (iv) availability of employment as vehicle drivers (8%); (v) easy availability of agricultural inputs/improvement in agriculture (14%); and (vi) improvement in the availability of various services, namely, extension, medical, and postal services (19%).

TABLE A17.14-3 ● Estimated Impacts of Rural Roads on Women and Their Families in the Eight Survey *Thanas*

Topic	Before the Project	After the Project
Enrollment of school-aged children	0.41 children per family	0.52 children per family
School drop-outs	0.38 children per family	0.23 children per family
Main source of earning	51.8% domestic servants; 16.7% were daily wage laborers; 14.7% were small traders; and 16.8% were jobless	Project work (vendors, planting trees, and maintenance workers)
Supplementary sources of income	2.6% in wage labor; 5% in domestic work; 77.4% in animal husbandry; 15% in fuel wood collection/making fishing nets	94.4% in animal husbandry and 5.56% in rice processing/fuel wood collection and making fishing nets
Average income of the respondent	47.08 tk./day per respondent	52.23 tk./day per respondent

(Continued)

(Continued)

Topic	Before the Project	After the Project
Ownership of homestead	50% of the total respondents owned their homestead	54% of the total respondents owned their homestead
Type of house	46% lived in thatched houses and 53% in houses made of wood/tin	28% lived in thatched houses and 71% in houses made of wood/tin
Ownership of livestock	One per respondent household	Two or three per respondent household
Ownership of poultry	6 per respondent household	10 per respondent household
Ownership and type of transport owned	0.33% of respondents owned rickshaws and 1% owned bicycles	0.43% of respondents owned rickshaws and 1.8% owned bicycles
Capacity to buy food during normal part of the year	30% of the total respondents had sufficient food, 66% had insufficient food, and 4% had irregular starvation	92% of the total respondents had sufficient food, 8% had insufficient food, and none had irregular starvation
Capacity to buy food during difficult part of the year	18% of respondents had sufficient food, 72% had insufficient food, and 8% had irregular starvation	79% had sufficient food, 21% had insufficient food, and none had irregular starvation
Capacity to buy durables (bed, trunks, radio, etc.)	12% had the capacity to buy durables	42.5% had the capacity to buy durables
Accessing paid medical services for the family and respondent	54% could access minor medical services, but none could afford major services	99% could access minor medical services, and 2% could access major medical services
Repair old houses	17.95% were able repair their old houses	74.36% repaired their old houses
Ability to purchase essentials on credit	14.10% were able to buy essentials on credit	97.44% were able to buy essentials on credit
Ability to save	3% were able to save	94.1% were able to save
Number of pieces of clothing per respondent	5.3% of the women respondents had more than one piece of clothing	57.69% of the women respondents had more than one piece of clothing
Availability of winter clothing (blanket)	8% of families had sufficient winter clothing	44% of families had sufficient winter clothing
Membership in groups/*samities*	42.31% of respondents were members of groups/*samities*	32.05% of respondents were members of groups/*samities*[31]

[31]The major reason identified was that the women had less time to participate in the activities of groups/*samities*.

LESSONS FOR THE EVALUATION OF GENDER IMPACTS

Several lessons can be drawn from this case:

1. Improving transport can produce major social and economic benefits for women and their families.
2. While impact evaluation findings will be much more robust when baseline studies have been conducted, for projects such as roads, which produce a very large and clearly defined impact, it is often possible to obtain usable estimates through retrospective analysis in which respondents are asked to recall their conditions before the project. However, the design should use *triangulation* to provide several independent estimates of each reported change.
3. In cases where baseline data is not available, it is particularly important to include a control group not affected by the project. While it is rarely possible to find a control group that closely matches the project group in all respects, a group with similar characteristics can provide a useful reference for helping to understand the processes of change that are being observed.
4. Projects such as roads can produce a very wide range of social and economic impacts. Consequently, it is important to use focus groups or other participatory approaches so that beneficiaries can identify the types of benefit they have experienced.
5. It is important to distinguish, as this study does, between direct benefits for women (such as improved access to markets) and general benefits affecting the whole community (stimulus to the creation of new businesses, encouragement to government agencies to build more schools and clinics, etc.) from which women also benefit.

Source: Ahmed (2000).

Case Study 5: The Impact of Bicycles on Women's Literacy and Empowerment in India

Follow-up study to assess the impacts of the introduction of bicycles 5 years earlier. Data was collected through key informant interviews, a focus group discussion, a village survey, and an activity and time profile with a sample of husbands and wives.

THE STUDY

This case study assesses the impacts on women's lives of the introduction of bicycles as part of a literacy campaign in the Pudukkottai region of Tamil Nadu State in India. The story of the introduction of bicycles and bicycle riding skills as part of a literacy campaign (by the National Literacy Mission) in the early 1990s in Tamil Nadu is a well-known example of women's increased mobility, independence, and empowerment through a successful intervention: cycling. The author of this study wanted to see how circumstances had evolved more than five years later, and whether the movement in women riding bicycles had been sustained and would remain sustainable. The initial campaign enlisted the help of men to teach women how to cycle. Loans were made available for women to buy cycles and those with a regular income (such as NGO extension workers, child care workers) were quick to take these up. As more women were seen regularly cycling, the opposition and male jokes died away. It became acceptable through the sense of it being a widespread movement.

METHODOLOGY

The study addressed three questions:

1. Though cycles were introduced from the perspective of empowering women rather than meeting their transport needs, have they been able to meet those needs, both for their productive and reproductive activities? Are women able to access bicycles to meet those needs?
2. What has been the impact of women's increased mobility on their self-esteem and confidence, on gender relations in the community?
3. Has providing cycles to women been a sustainable intervention? In particular, has women's investment in cycles continued and do they have control over the use of these cycles?

Data-collection methods: Data was collected through key informant interviews, a focus group discussion, and a village survey covering 49 women in 12 villages. In addition, an activity and time profile was conducted with eight couples.

Only three of the sampled women did not know how to ride a bicycle. Most of the respondents were Scheduled and Backward caste women, half of them barely literate and the others educated up to middle school. They are mostly in the 20- to 30-year age-group, and most of them have children and families to care for, in addition to their income-earning activities. Most are unskilled laborers and their workload is heavy.

Other women reported how taking a sick relative or child to hospital themselves on the bicycle gave them a feeling of independence and usefulness—of being a "useful member of society." The motivation to learn among the women who do not yet know how to cycle is still high today.

FINDINGS OF THE STUDY

Control of bicycles and women's access in a male-dominated culture

In the majority of rural homes of the district, a cycle is now common property. In a door-to-door survey covering 50 households, it was found that 32 of them (64%) owned a cycle; 83 out of 91 men asked knew how to cycle, and 34 out of 100 women. There might perhaps have been three or four prior to the literacy campaign. While access to cycles for women now seems widespread, what is more problematic is the issue of control. Very few women still actually own cycles, hence they are dependent on the cycles of others, and they have to adjust their work according to the needs of the owners. For instance, if a husband owns the bike and has to leave for work at 8 a.m., then the woman has to get up extra early to try to finish as much of her work (water collecting, etc.) before this time. The men in the households generally own the bicycles, and so they get priority in their use.

Only 12 of the 49 women interviewed had easy access to cycles, and another 10 reported that they usually had access to a bicycle when they needed it. The distance of the cycle hire shop was quoted as a problem for the women, reconfirming that the utility of cycles is no longer an issue of debate for the women but is seen as an accepted requirement to meet their needs.

Cycling for women does not seem to have changed gender relations (for more than two-thirds of the sample) in the household significantly. Major decision making (on expenditures, etc.) continues to be vested in the men.

However, there are still some social restrictions that prevent some women from cycling. Husbands say they worry about their wives or daughters being injured, but in many cases women's work is just not a priority for men. Cycles greatly reduce the time and labor inputs for women in several drudgery-ridden tasks that are essential for household maintenance, but as these are unpaid tasks and have no cash value, the owners of the cycles, mostly men, do not see cycles as critical for women in the performance of their tasks.

Economic Impacts of Women's Increased Use of Bicycles

With the greater acceptance of cycling in the district, the profitability of cycle shops as an income-earning enterprise has seen their numbers increase steadily. A cycle shop is now seen as a facility that should be available in a village. With changes in employment patterns and lifestyles, the isolated and self-sufficient village economy is a thing of the past. Mobility and transportation are integral parts of people's lives. Large numbers of girls are cycling to school every day in Pudukkottai; this is indicative of even higher bicycle use in the next generation.

In concluding the comment on the evidence of the survey and interviews, the author states that the primary impact of learning to cycle on women's lives is their perception of independence in terms of their roles in the household and community—productive, reproductive, and community managing roles. The second and related impact has been in terms of improvement in both their self-confidence and self-esteem.

Bicycles and Gender Relations

Looking at gender relations, the picture is more complicated. On the positive side, women cycling has come to be accepted as a normal phenomenon, and rural girls now learn to cycle alongside boys.

An activity and time profile conducted with eight couples revealed that while men and women spent six to eight hours per day on paid work, the women spent a similar amount of time on household maintenance and child care tasks, whereas men spent less than two hours on these. A woman's working day could stretch to between 12 and 18 hours per day.

On the other hand, almost 40% of the women reported that their workloads had actually increased. Tasks that the men would do before, such as marketing, taking the children to school, or whatever involved traveling distances, have all now shifted to the women. Cycles do, however, help them to complete their jobs faster and more easily. Despite their extra burdens, they report having more time for leisure.

On a broader front, the Pudukkottai program has demonstrated that cycling can be one very effective strategy for empowering women. The women themselves have found an efficient, cheap, and easy way of meeting their transport needs, which has also empowered them. The signs are that use of cycles by women in Pudukkottai is a sustained and sustainable phenomenon—an integral and necessary part of their lives.

Widening the Activities for Which Bicycles Are Used

Bicycling is generally viewed as a cheap and efficient means of transport and definitely contributes to meeting the transport needs of women, particularly those in "low access" villages (distant from essential services). The pattern of use and ownership of cycles bears out that better provision of services, such as drinking water, food shops, and health and education facilities, can lead to substantially reducing women's transport burden and needs.

Only four out of the sample of 49 women actually owned their own bicycle, however. Women seemed willing to use hired cycles not only in emergencies but also for use in paid work or when they were able to plan several household tasks together that are located at a distance. Hiring every day would be too expensive, but now that they know how to cycle they can also borrow from neighbors or use one belonging to another member of their own household.

The researchers found that all women who had access to cycles, whether their own or that of a husband, father, or brother, were using them for a range of tasks, related to all areas of their responsibilities. The most common uses were fetching water from the well or tank, taking paddy to the rice mill, collecting fuel and fodder, going to the hospital in an emergency, and going to school (younger girls). A few use the cycle for their productive work, such as selling flowers in the market, purchasing and selling gems to and from the contractor, and maintenance of plants in a government nursery.

A common theme is that with access to bicylces women can be more involved with social, development, and community tasks because they can confidently and independently cycle from village to village. This is the case of one woman: having a bicycle has enhanced her status within the home so she is now a major decision maker in her household. Her husband is quite happy with this, not least because his workload has reduced! This woman bought the bicycle on a loan, as many others have done, and she has already repaid it.

An interesting issue is that while between 30% and 50% of people hiring cycles in the district are women, ladies' cycles can rarely be found in the shops. The women have, in fact, gotten used to riding gent's cycles, and in fact feel that it gives them better balance when carrying loads. Even riding a gent's bicycle in a sari doesn't bother the women anymore; the convenience of this mode of transport outweighs all other considerations.

LESSONS FOR THE EVALUATION OF GENDER IMPACTS

The study shows that it is possible to obtain quick and economical assessment of the use of a new resource such as bicycles. However, in order to assess the importance of the literacy campaigns in promoting the use of bicycles, it would be necessary to use a sample design that includes women who had participated in the literacy campaign as well as women who did not participate in the campaign.

Source: Rao (1999).

Case Study 6: Assessing the Impacts on Girls' Educational Enrollment of Private Schools for the Poor in Balochistan, Pakistan

A longitudinal impact evaluation with random selection of treatment communities and before-and-after comparisons of the project and control areas and of female and male students

THE STUDY

1) It is commonly presumed that cultural taboos against exposing girls to the public further limit incentives for poor parents to send their daughters to school. This paper presents evidence that challenges that view and suggests that parental reticence regarding their children's education can be overcome. In particular, this paper reviews the planning, implementation, current status, and future sustainability of two pilot projects designed to address the constraints to the education of girls in Balochistan. A comparison of enrollment growth between pilot and control communities is used to measure the impact of each experimental program. The aim is to highlight lessons that can be learned from these projects that may be applied in other contexts.

2) Two pilot projects to encourage private school enrollment of poor girls were included in the 1994 Social Action Program (SAP) for Balochistan. One attempted to increase private girls' school supply in Quetta, the capital city of Balochistan. Another was designed to raise private girls' school supply in rural areas. Both these programs assisted communities in setting up private schools and offered financial assistance to defray the initial costs of operation as well as to assist in setting up an endowment. Communities in 10 urban slum areas were invited to set up Parent Education Committees (PECs) and to present proposals for the establishment of a private school which would receive a declining subsidy over a three-year period. Numerous proposals were received from PECs in all 10 areas, and one community was selected randomly in each area. The PECs were able to successfully establish schools in all communities, and in one community two schools were established.

In rural areas the Community Support Program (CSP) had successfully launched 247 schools by 1995, but the participation of girl students was constrained by the difficulties of recruiting either women teachers or male teachers who would be considered suitable by the girls' parents. Consequently, the Rural Girls Fellowship program was established, *using a similar approach to the urban schools program, in which the government* provided resources, including funds to recruit female teachers, and the Village Education Committee (VEC) provided the school. Again, demand exceeded supply and random selection was used.

METHODOLOGY

In the urban areas a longitudinal quasi-experimental design was used with random allocation of communities to treatment and control groups, and with comparisons being made before and after the introduction of the project schools. The dependent variable was school enrollment, which was analyzed for both female and male students. The evaluation design for the rural areas is similar except that no resurvey was conducted in the control areas.

In the urban areas the evaluation approximated an experimental design with the assumption of a perfectly random assignment and consequently assumes that the difference-of-difference of means between the control and experimental groups in the pretest–posttest situations will provide an unbiased estimate of project impacts. However, statistical tests were used to control for individual differences in the characteristics of the experimental and control communities, which might have biased the results.

In the rural areas, given the fact that repeat surveys were not conducted for the control groups, a reflexive evaluation design was used in which the program effect is estimated as the change in the enrollment rates in the project group before and after the project intervention was considered. The authors recognize that due to the lack of a control group this design is not able to control for national trends and other external events, which may lead to the observed changes being erroneously attributed to the project.

FINDINGS OF THE STUDY

1) The Balochistan Province of Pakistan initiated two pilot programs attempting to induce the creation of private schools for the poor. Randomized assignment to treatment and control groups is used to measure program effectiveness. The pilot schools were successful in urban areas but were relative failures in rural areas. Urban schools benefited from larger supplies of children not served by government schools, better availability of teachers, and more educated parents with higher incomes. Use of experienced school operators in the urban pilot was another critical difference. All urban schools appear self-sustaining or else require a modest subsidy, whereas only one rural school may survive as a private school. These pilots show that private schools may offer a viable alternative supply of educational services to poor urban neighborhoods in developing countries. However, they are not likely to offer solutions to undersupply of educational services to rural areas.

Table A17.14-4 shows that the project led to a marked increase in girls' enrollment rates in the 10 urban areas from 45.29% before the program to 76.15% in 1996. There was no comparable increase in girls' enrollment in the control areas. Equally interesting was a comparable increase in boys' enrollment in the project areas with no increase in control areas, despite the fact that the schools received no subsidies for boys and that boys had to pay tuition that was as high as or usually higher than that for girls.

TABLE A17.14-4 ● Comparison of the Effect of the Urban Fellowship Program

Outcome Measure	Treatment		Control	
	Boys	Girls	Boys	Girls
Enrollment Rate Before Program E_0	56.33	45.29	51.06	34.86
Enrollment Rate in 1995 (E_{95})	64.29	63.93	49.68	38.37
Enrollment Rate in 1996 (E_{96})	76.15	71.30	43.50	36.20

Table A17.14-5 shows similar impacts on girls' enrollment in the rural areas. On average, girls' enrollment increased from 41.5% before the project to 51.8% after the project. In contrast, boys' average enrollment declined from 66.6% to 59.8%. The differences were still significant when controlling for child age and birth order, household income, and parents' education. In contrast to the urban areas, no similar increase in boys' enrollment occurred; in fact, boys' enrollment rates declined slightly.

LESSONS FOR THE EVALUATION OF GENDER IMPACTS

The study demonstrates the value of a longitudinal pretest–posttest design with a control group, and with studies being conducted with both groups before and after the project intervention. This is analytically much stronger than the *reflexive evaluation design* used in the rural areas in which the control group is not resurveyed. This design makes it possible to control for historical trends that might have caused the observed changes, in a way that is not possible with a cross-sectional design. The design is even stronger for projects such as this where communities or individuals are randomly assigned to control and experimental groups. Although randomization is usually not possible in most development evaluations,

TABLE A17.14-5 ● Enrollment Effects Using Longitudinal Observations of Rural Girls' Fellowship Villages Enrollment Rates Before and After the Project Implementation

District	Before Implementation		After Implementation	
	Girls	Boys	Girls	Boys
Chagai	50.3%	73.3%	64.9%	72.7%
Mastung/Kalat	61.3%	81.0%	56.0%	60.8%
Gwadar	21.8%	49.1%	43.9%	53.7%
Average	41.5%	66.6%	51.8%	59.8%

situations do occur where some kind of lottery or random selection is used as a way to ensure transparency when demand for services exceeds available resources. However, even when randomization is used, it is still advisable to control statistically for differences in the characteristics of the experimental and control populations that might be correlated with the dependent variable.

The study did not discuss any particular gender issues that might affect the evaluation design or the data-collection methods. One of the gender measurement issues that has been discussed in other evaluation studies, including in Balochistan, concerns possible gender-related reporting biases. For example, during an earlier period in Balochistan there were very few girls' secondary schools and the idea of mixed schools was not widely accepted. However, families sometimes enrolled their daughters in boys' secondary schools, but some of the headmasters, fearing criticism from the education authorities, registered the girls as boys. Some education researchers cited the low reported girls' enrollment as evidence of the lack of interest of parents in sending their daughters to secondary school. On the other hand, in some Muslim countries in Africa, families did not wish their daughters to attend secondary school once they reached puberty, but as education was compulsory they would sometimes bribe the attendance officer to record their daughter as attending school. These are both examples of the need to use consistency checks (triangulation) to check for any possible gender-related reporting biases.

A final lesson, which does not apply just to gender analysis, concerns the need for care in the interpretation of findings based on a small number of sample areas. In the present rural study the results varied among the three districts, with significant increases in girls' enrollment in two districts and very little change in the third district, which already had a high girls' enrollment. If by chance one of the two high-change districts had been replaced by another district that already had a high girls' enrollment, then the overall findings of the study might have shown a much lower level of increase in girls' enrollment. There are two lessons: first, to obtain as much relevant information as possible to be used in the selection of districts, and second, where possible, to include a large number of areas (in this case districts) in the sample.

Sources: Alderman, Kim, & Orazem (2000); Kim, Alderman, & Orazem (1999).

Case Study 7: Assessing Gender Mainstreaming in the World Bank

Two institutional assessment studies that combine ratings on the treatment of gender in World Bank projects

THE STUDIES

The Operations Evaluation Department (OED) of the World Bank has conducted two studies to assess the extent to which gender is mainstreamed in World Bank lending. These studies are described below as examples of ways to conduct a broad-based institutional assessment.

METHODOLOGIES

The 1995 study reviewed the 4,955 investment projects approved by the World Bank between fiscal years 1967 and 1993 and identified 615 that specified some gender-related actions in the project appraisal document. The 1997 study added projects completed during fiscal years 1994 through 1996 that increased the number of projects specifying gender-related actions to 802. All 802 projects with gender-related actions were analysed,[32] with special attention being given to the 58 projects on which more detailed information was available on project outcomes.

FINDINGS OF THE STUDIES

The 1997 study on "Mainstreaming Gender in World Bank Lending"

A key factor in the effectiveness with which gender issues are addressed in development strategies is the adequacy with which agencies integrate ("mainstream") gender within organizations and in the planning and implementation of their development programs. Consequently, most international development agencies have conducted one or more assessments of how adequately gender is addressed within their agency. Two recent World Bank studies (Murphy, 1995, 1997) illustrate the approaches and findings of one such study.

The following are some of the findings of the 1997 study:

- Projects with gender-related actions achieved their overall objectives—that is, they received a satisfactory outcome rating—in relatively greater proportion than similar projects without specified gender actions.
- 74% of 51 completed projects with gender-related actions approved after 1987 in the agriculture and human resources sectors were rated satisfactory for overall outcomes, compared with 65% of 81 projects that did not include gender-related actions.
- Projects that explicitly incorporate gender goals are the most likely to achieve their gender objectives.
- Gender objectives are most likely to be achieved if they are well integrated into the main project objectives.
- The design quality of gender-related actions improved significantly for projects approved in 1994–1995 compared with projects approved earlier.
- The prevalence of projects with gender-related actions remained around 30% in the fiscal 1994–1996 period, having fallen from a high of 45% to 50% in the early 1990s.

[32]Projects were given a 0 rating if there was no reference to gender, 1 if there was some discussion of gender issues, and 2 if specific gender actions were proposed.

- Projects with gender-related actions continued to be concentrated in the "traditional" sectors, with 88% in the agriculture, human resource, and social sectors. In contrast, gender issues received very little attention in many of the Bank's major investment sectors. The 1995 report found that less than 10% of all completed projects in the energy, finance, industry, power, telecommunications, tourism, transport, and water and sanitation sectors included any attention to gender.

LESSONS FOR THE DESIGN OF INSTITUTIONAL GENDER MAINSTREAMING AND IMPACT EVALUATIONS

Several limitations of the methodology used in the 1995 and 1997 studies must be kept in mind in the interpretation of the findings. First, a very broad definition of treatment of gender was used, so that an appraisal report that stated that a high proportion of farmers are women, or that female students experience higher absentee rates than males, would both be rated as having included a discussion of gender. Second, the assessment is mainly based on the statements in the appraisal reports indicating *intended actions* and in most cases it was not possible to obtain information on the degree to which these intended actions were actually implemented. Third, in many countries, particularly in the earlier period covered by the study, gender issues were considered a very sensitive topic, so it is sometimes difficult to know whether the lack of explicit discussion of gender reflected a lack of awareness (or interest) on the part of the project team, or whether in some cases a decision may have been taken to downplay gender in the planning documents—even when there may have been some intention to address these issues during project implementation.[33] Fourth, the study had only limited information on how, and to what extent, gender issues were incorporated into project implementation, monitoring, and evaluation.

Sources: Murphy (1995, 1997).

[33]An earlier study with which the author was involved, using a similar methodology to assess the treatment of participation in Bank projects, found evidence that Bank efforts to promote participation were often downplayed in project documents. At least one of the reasons was that community participation was considered quite controversial in many countries in the 1970s and the 1980s.

Bibliography for Case Studies

Ahmed, N. (2000). *Study on gender dimensions of the Second Bangladesh Rural Roads and Markets Improvement and Maintenance Project*. Available at http://siteresources.worldbank.org/INTGENDERTRANSPORT/Resources/StudyReportonGenderDimensioninRuralRoadsandMarket.htm

Alderman, H., Kim, J., & Orazem, P. (2000). *Design, evaluation and sustainability of private schools for the poor: The Pakistan Urban and Rural Fellowship School Experiments* (mimeo). Report on a study funded by the World Bank Economics of Girls Education and Girls Education Thematic Groups and the World Bank Development Research Group.

Baker, J. (2000). *Evaluating the impact of development projects on poverty: A handbook for practitioners.* Directions in Development. Washington, DC: World Bank.

Bamberger, M. (2000). *Integrating quantitative and qualitative research in development projects.* Directions in Development. Washington, DC: World Bank.

Bamberger, M., Blackden, M., Fort, L., & Manoukian, V. (2002). Gender. In *A sourcebook for poverty reduction strategies.* Washington, DC: World Bank. Available at http://worldbank.org/poverty/strategies/chapters/gender/gender.htm

Barwell, I. (1996). *Transport and the village.* World Bank Discussion Paper No. 344. Washington, DC.

Blackden, M. (2001). *Too much work and too little time: Gender dimensions of transport, water and energy.* World Bank Institute Course on New Agendas for Poverty Reduction Strategies: Integrating Gender and Health. February 20–April 10, 2001.

Chant, S., & Gutmann, M. (2000). *Mainstreaming men into gender and development*. Oxfam Working Papers. Oxford, England.

Fafchamps, M., & Quisumbing, A. (2001a). *Assets at marriage in rural Ethiopia*. IFPRI. Available at www.ifpri.org

Fafchamps, M., & Quisumbing, A. (2001b). *Control and ownership of assets with rural Ethiopian households*. IFPRI. Available at www.ifpri.org

Gross, B., van Wijk, C., & Mukherjee, N. (2000). *Linking sustainability with demand, gender and poverty: A study in community-managed water supply projects in 15 countries.* International Water and Sanitation Center. December 2000.

Hills, J., Le Grand, J., & Piachaud, D. (Eds.). (2002). *Understanding social exclusion*. New York, NY: Oxford University Press.

Khandker, S. (1998). *Fighting poverty with microcredit: Experience in Bangladesh.* New York, NY: Oxford University Press for the World Bank.

Kim, J., Alderman, H., & Orazem, P. (1999). Can private school subsidies increase schooling for the poor? The Quetta Urban Fellowship Program. *World Bank Economic Review, 13*(3), 443–466.

Lee, A., & Hills, J. (1998). *New cycles of disadvantage.* Report of a conference organized by the Centre for Social Exclusion for H.M Treasury. http://sticerd.lse.ac.uk/case

Malmberg-Calvo, C. (1994). *Women in rural transport.* SSTP Working Paper No. 11. World Bank and Economic Commission for Africa.

Moser, C. (1993). *Gender planning and development: Theory, practice and training.* London: Routledge.

Murphy, J. (1995). *Gender issues in World Bank lending.* Operations Evaluation Department. Washington, DC: World Bank.

Murphy, J. (1997). *Mainstreaming gender in World Bank lending: An update.* Operations Evaluation Department. Washington, DC: World Bank.

Narayan, D., with Patel, R., Schafft, K., Rademacher, A., & Schulte, S. K. (2000). *Can anyone hear us? Voices of the poor* (Vol. 1). New York, NY: Oxford University 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 at www.worldbank.org/gender/prr

Overholt, C., Anderson, M., Cloud, K., & Austin, J. (1985a). *Gender roles in development projects: A case book*. West Hartford, CT: Kumarian Press.

Overholt, C., Anderson, M., Cloud, K., & Austin, J. (1985b). Women in development: A framework for project analysis. In C. Overholt, M. Anderson, K. Cloud, & J. Austin (Eds.), *Gender roles in development projects: A case book*. West Hartford, CT: Kumarian Press.

Pitt, M., & Khandker, S. (1998). The impact of group-based credit programs on poor households in Bangladesh: Does the gender of participants matter? *Journal of Political Economy, 106,* 958–996.

Pitt, M., Khandker, S., McKernan, S.-M., & Latif, M. A. (1999). Credit programs for the poor and reproductive behaviour in low-income countries: Are the reported causal relationships the result of heterogeneity bias? *Demography, 36*(1), 1–21.

Quisumbing, A. R., & Maluccio, J. A. (1999). *Intrahousehold allocation and gender relations: New empirical evidence.* Working Paper Series 2, Policy Research Report on Gender and Development, World Bank, Development Research Group/Poverty Reduction and Economic Management Network, Washington, DC.

Sen, A. (1993). Capability and wellbeing. In M. Nussbaum & A. Sen (Eds.), *The quality of life.* Oxford: Clarendon.

United Nations Development Program (UNDP). (1995). *Human development report. Gender and human development.* New York, NY: Oxford University Press.

United Nations Development Program (UNDP). (2000). *Human development report*. New York, NY: Oxford University Press.

Valadez, J., & Bamberger, M. (1994). *Monitoring and evaluating social programs in developing countries*. Economic Development Institute. Washington, DC: World Bank.

Williams, S. (1994). *The Oxfam gender training manual.* Oxford, UK: OXFAM Publications.

World Bank. (2000). *Attacking poverty.* World Development Report 2000/2001. Washington, DC: World Bank.

World Bank. (2001). *Engendering development through gender equality in rights, resources and voice.* Oxford, UK: Oxford University Press.

APPENDICES FOR CHAPTER 19 MANAGING EVALUATIONS

19.1 The Evaluation Framework and Scope of Work

19.2 The Actors Involved in an Evaluation of How Gender Equality Issues Were Addressed in an International Development Program

19.3 Procedures for Contracting Evaluation Consultants

19.4 Guidelines for Strengthening Terms of Reference

19.5 Building in Quality Assurance Procedures

19.6 Evaluation Capacity Development

Chapter 19 provides guidance on how to manage an evaluation. It begins by discussing organizational and political factors that influence the design, implementation, and use of evaluations. Evaluations are a political activity that can have important consequences—both positive and negative—for a range of stakeholders, many of who are interested in influencing the evaluation and its findings. The chapter then identifies the main steps in planning, management, and use of evaluations. The chapter then refers to the Threats-to-Validity Checklist discussed in Chapter 7 and concludes with strategies for the development and institutionalization of sectoral and national evaluation strategies.

There are six appendices: the evaluation framework (a detailed description of the purpose of the evaluation and the context within which it will be implemented) and the scope of work (Appendix 19.1); using an evaluation of an international gender program to illustrate the application of stakeholder analysis (Appendix 19.2); examples of procedures for contracting evaluation consultants (Appendix 19.3); guidelines for strengthening the evaluation terms of reference (Appendix 19.4); building quality assurance procedures into the evaluation (Appendix 19.5); and different approaches to evaluation capacity development (Appendix 19.6).

Many of the technical terms in these appendices are included in the Glossary in the book.

APPENDIX 19.1 THE EVALUATION FRAMEWORK AND SCOPE OF WORK

The Evaluation Framework

Evaluation frameworks are used extensively by the United Nations Development Programme (UNDP; 2009) among many other development agencies, and are considered particularly important for collaborative evaluations that involve more than one development partner (UNDP, 2009, Chapter 3). The framework should clarify the following:

- What is to be evaluated
- The activities needed to set-up and implement the evaluation
- Who is responsible for different evaluation activities
- Timing of the evaluation activities
- The proposed methods
- What resources are required and where do they come from

The UNDP recommends that the evaluation framework normally has three components: a narrative component describing how (international and national) partners will undertake evaluation activities and the accountabilities assigned to each group, a results framework, and a planning matrix (see UNDP, 2009, p. 87 for an example of an evaluation plan/matrix). Some of the considerations for planning the evaluation system include the following:

- Uses, purposes, and timing
- Resources available and required
- The likelihood of future initiatives in the same area
- Anticipated problems
- Need for lessons learned
- Alignment and harmonization (when several development agencies are involved)

It can also be useful to check "evaluability readiness" (UNDP, 2009, p. 148). This involves questions such as the following:

- Does the subject of the evaluation have a clearly defined results map?
- Is there a clearly defined results framework?
- Is there sufficient capacity to provide required data for the evaluation?
- Is the planned evaluation still relevant?
- Will political, social, and economic factors allow for an effective conducting and use of the evaluation?
- Are there sufficient resources?

Additional issues arise when planning a joint evaluation involving a number of different partners. The partners must be selected, and there must be agreement on the scope of work, the funding modality, the process for selecting the evaluators, the reporting and dissemination strategies, and the modality for management response.

The Agency Scope of Work (SoW)

While agencies use the term *Statement of Work* or *Scope of Work* (SoW) to avoid confusion (see Box A19.1), we will refer to this planning document as the *Agency* scope of work. The Agency SoW, as used by organizations such as USAID (2010), can be broader than the evaluation framework as it may lay out a monitoring and evaluation strategy for a complete program lasting several years, which may involve a number of distinct evaluation activities, some conducted internally and some subcontracted. It may also define roles and responsibilities for the management and implementation of the evaluation within the contracting agency, as well as the overall budget and staffing requirements for all of the evaluation activities, including the resources for internal activities as well as for external contracts. For many organizations, the Agency SoW is a standard planning and budgeting mechanism used for all program activities, not just evaluations. This means that evaluation plans must be defined using the same procedures as other program activities, which can introduce some additional constraints and issues in large and complex organizations where administrative departments are processing very large numbers of diverse contracts.

BOX A19.1

How Different Agencies Use the Terms *Evaluation Framework*, *Scope of Work*, and *Evaluation Terms of Reference*

There are some potential confusions concerning the terms *evaluation framework*, *scope of work* (*SoW*), and *terms of reference* (*ToR*) as they are used differently by different agencies.

Evaluation framework. This is an internal planning document used by the UNDP to define the purpose and scope of the evaluation, timing, responsibilities, resource requirements, and proposed methodologies. If this is a collaborative evaluation, the framework will be shared with partners. Once agreement has been reached on the framework, this will be used to define the terms of reference that will be included in the request for proposals that is communicated to interested consultants.

Scope of work (SoW). Agencies such as USAID use the SoW in a similar way to the UNDP's evaluation framework as defining the scope, timing, responsibilities, resource requirements, and proposed methodology. One of the purposes of the SoW is to prepare the budget, time, and other resource requirements. These must be approved before the evaluation begins. Often the SoW covers a broad program of activities that may include a number of different specific evaluations, so often this will lay out the whole evaluation strategy for a multiyear program.

The confusion arises because other agencies use the term *scope of work* to refer to the terms of reference that are given to consultants defining the scope of the evaluation, the questions that must be addressed, and so on.

Terms of reference (ToR). This a general term referring to a document that defines the responsibilities of any consultant contracted for a particular purpose, including but not limited to the design and implementation of an evaluation. In the case of an evaluation ToR, this will describe the purposes of the evaluation, the deliverables to be produced, the timing, and the available resources.

What some agencies call the terms of reference to be given to consultants, other agencies call the scope of work.

How we use the terms: We will use the term *Agency* scope of work to refer to the internal planning document used by an agency to plan and budget an individual evaluation or an evaluation program and the term *evaluator* scope of work to refer to the document that is given to prospective evaluation consultants.

The Agency SoW for an evaluation spells out the mutual obligations between client and the evaluation team—whether internal, external, or mixed. It explains the evaluation parameters and the resources available for conducting it.

- Start with the constraints within which the evaluation must operate:
  - The budget
  - Timing
  - What are other real-world constraints on the evaluation? (lack of baseline or counterfactual data, political pressures, etc.)
- The SoW should be considered as a first step in developing the evaluation utilization plan
- Key elements of the SoW
  - Program implementation: what, where, when (start and end dates)
  - Evaluation fundamentals: purpose, main questions
  - Technical requirements: design, methods, staffing
  - Management information: schedule, budget
  - Present the theory of change (logic model/development hypothesis/results format)
  - Target groups and areas
  - Critical assumptions
  - Information available on the project and its context
- Evaluation fundamentals
  - Primary users of the evaluation
  - How it will be used
  - Technical requirements of the evaluation design
    - Key questions
    - Sources of data
    - Data-collection methods
    - Data quality expected
    - Data analysis
    - How the data will be disaggregated and presented
- Management requirements
  - Team qualifications and size
    - Local personnel
  - Deliverables
  - Schedule
  - Logistics
  - LOE (level of effort)/budget
  - Estimated costs for each step
  - Reporting

Box A19.2 lists the elements of a good evaluation SoW identified by USAID.

BOX A19.2

The Elements of a Good Evaluation Statement of Work (SoW)

1. Describe the activity, program, or process to be evaluated
2. Provide a brief background
3. State the purpose and use of the evaluation
4. Clarify the evaluation questions
5. Identify the evaluation methods
6. Identify existing performance information
7. Specify deliverables and timelines
8. Discuss the composition of the evaluation team
9. Address scheduling, logistics, and other support
10. Clarify requirements for reporting and dissemination
11. Include a budget

Source: USAID (2010).

APPENDIX 19.2 THE ACTORS INVOLVED IN AN EVALUATION OF HOW GENDER EQUALITY ISSUES WERE ADDRESSED IN AN INTERNATIONAL DEVELOPMENT PROGRAM

Two individual consultants were commissioned to conduct an evaluation of the gender policies of an international agency promoting food security in low-income countries around the world. It was agreed that fieldwork would be conducted in a sample of six representative countries in Africa, Asia, and Latin America and that this would be combined with existing documentation from other countries and interviews with agency headquarters staff. The principal actors involved in the evaluation were (a) the two international consultants; (b) local consultants contracted by the international consultants to conduct data collection in several countries; (c) the evaluation department of the development agency, which was actively involved in developing detailed guidelines for the evaluation, working actively with the consultants, and organizing videoconferences to obtain feedback from staff in a wider range of countries than the sample countries selected for fieldwork; (d) country offices in the sample countries, which coordinated the interview schedule, coordinated field trips, and accompanied consultants on some of the field visits; (e) the gender focal points in country offices in sample countries, which coordinated the distribution and collection of a survey administered to agency staff; (f) partner government agencies in sample countries, which provided additional statistical data requested by consultants following field visits; (g) NGO implementing agencies in sample countries, which helped facilitate community meetings in beneficiary communities (including providing interpretation into local languages); (h) key informants from agencies such as the district health department and the police, familiar with the communities but not involved in the project so as to avoid bias from obtaining information only from beneficiaries and agencies directly involved in the project; and (i) the gender unit of the client agency, which provided comments on the evaluation design and the survey instruments and provided background documentation but were not directly involved in the implementation of the evaluation.

The evaluation design sought to ensure impartiality and objectivity by using external evaluators while accepting the practical necessity (due to time and budget constraints) to involve the country offices of the client in coordinating interview schedules, distributing surveys, and, in some cases, facilitating and providing interpretation for community meetings. Consultants were aware of the potential biases that involvement of the client implied and sought to control for this through triangulation (interviews with other stakeholders, review of secondary data, requests for additional data from government and NGO implementing agencies, and through consultation with agencies not involved in the project).

Source: Unpublished personal experience of one of the present authors.

APPENDIX 19.3 PROCEDURES FOR CONTRACTING EVALUATION CONSULTANTS

Different Procedures for Selecting and Contracting Consultants

Different agencies use different procedures for selecting and contracting consultants, and the procedures can also vary depending on the size, duration, and complexity of the contract. Often, simpler procedures will be used for smaller contracts, and in some cases for small contracts, a single consultant or firm can be selected without going through a competitive bidding process. Some of the steps and options, described in more detail in the following sections, include the following:

a. *Using a planning framework.* For many large organizations, the first stage is the use of a broad planning framework such as an evaluation framework or statement of work (SoW) that defines the overall purpose, approach, and resource requirements of the evaluation (discussed earlier in Step 1-A). Often the SoW will cover the total monitoring and evaluation for a multiyear program that might involve a number of different evaluation contracts.

b. *Invitation to submit an expression of interest (EOI).* Often for large contracts, but sometimes also for smaller ones, interested firms and individual consultants can be invited to submit an EOI indicating their interest in submitting a proposal. Sometimes an announcement will be posted, and all interested consultants can respond, while in other cases, the invitation is only sent to a selected short list. The EOI will include background on the firm/consultant and usually initial ideas on the proposed approach.

c. *Request for proposals (RFP).* This document provides detailed information on the purpose and scope of the proposed evaluation and procedures for submitting a proposal. In some cases, all interested firms/consultants can submit proposals, while in other cases, there will be a preliminary screening process, often based on responses to the EOI.

d. *Terms of reference (ToR).* Sometimes the ToR will be completely defined in the RFP, but in other cases, the RFP will invite firms to comment on the draft ToR, and this may be revised and finalized during negotiations with the selected or short-listed firms.

e. *Inception report.* Once consultants have been selected, many agencies require, at least for large contracts, that an inception report be prepared. This provides consultants an opportunity to revise their methodology after having spent some time in the field assessing the feasibility of the proposed methodology, and this may result in revisions to the ToR or SoW.

Contracting the Evaluation Consultants

Evaluators can be contracted directly or through a competitive process. Governments have procedures specifying the required recruiting and contracting processes based on the size and nature of the contract and sometimes the source of funding. Donor agencies also have similar requirements. Working within the required contracting procedures, there are often a number of factors to be considered when recruiting consultants:

- *Fixed cost or level of effort.* A fixed-cost contract specifies the deliverables (evaluation reports, training activities, etc.) to be provided and the time scale for their delivery. This has the advantage that the contracting agency and the consultant know the budget commitment in advance. This procedure works well when the nature of the product is clearly understood and when it can be quantified (e.g., conducting a certain number of interviews). However, a fixed-cost procedure works less well in a new field where the appropriate evaluation design must be developed, tested, and possibly revised and where it is difficult to know in advance how long this will take. The consultant has an incentive to complete the design as quickly and cheaply as possible so as not to lose money on the contract, and consequently there is a danger of producing a poor-quality design.

On the other hand, the level-of-effort contract is more flexible as the number of consultant days can be adjusted as the design requirements become clear. This also provides the flexibility to conduct follow-up interviews if the need arises. However, the potential disadvantage is that the consultant may have the incentive to spend more time than is really required.

- *Contracting a firm versus an individual consultant.* Hiring an individual consultant will usually be much cheaper and may work well if the client is familiar with the field and knows which consultant to hire. The danger is that if the consultant is not able to deliver (she or he is overcommitted or becomes sick, for example), there may not be a backup team member available, and the completion of the contract may be delayed or the quality may suffer. Contracting through an experienced and well-established firm has the advantage that it can draw on a wider range of professional expertise and have more logistical resources for conducting evaluations. However, when working with a firm, it is important to clarify who will actually work on the contract as some firms list a number of well-known professionals in the proposal, but these people may have relatively little direct involvement in the design or implementation of the evaluation.

- *Broadening the range of research expertise.* Many consultants and firms have developed their expertise in particular research fields such as econometrics and quantitative survey research or qualitative and participatory evaluations. With the growing interest in mixed-method evaluations that combine a broad range of quantitative and qualitative techniques, consultants will often be requested to broaden the range of data-collection and analysis techniques they use. This will often result in an evaluation design that is not well integrated. For example, a number of not very well selected focus groups or in-depth case studies may be added to a sample survey, or a rapid and not well designed survey may be added to a set of case studies to permit generalization to a broader population. In these cases, the client may wish to contract a *technical support consultant* to define the requirements for these mixed-method designs and to assess the quality of the proposals.

APPENDIX 19.4 GUIDELINES FOR STRENGTHENING TERMS OF REFERENCE

The draft terms of reference (ToR) will usually be included in the request for proposal (RFP), but it may be revised and finalized later to address comments that consultants include in their proposals or on the basis of feedback from the inception report (see Step 4-B in Figure 19.1 in Chapter 19). According to the Organization for Economic Cooperation and Development (OECD, 2001), a ToR for an evaluation is a written documentation that should present the following:

- The purpose and scope of the evaluation
- The methods to be used
- The standard against which performance is to be assessed or analyses are to be conducted
- The resources and time allocated
- Reporting requirements

Some agencies include a consultant statement of work (SoW) as an element of a solicitation for an evaluation. The SoW for an evaluation spells out the mutual obligations between client and the evaluation team—whether internal, external, or mixed. It explains the evaluation parameters and the resources available for conducting it.

Expanding on the OECD/Development Advisory Committee (DAC) list cited above, the ToR typically includes the following (Morra-Imas & Rist, 2009, p. 445):

- A short descriptive title
- A description of the project or program
- The reasons for and expectations of the evaluation
- A statement of the scope and focus of the evaluation
- Identification of stakeholder involvement
- A description of the evaluation process
- A list of deliverables
- Identification of necessary qualifications of the evaluators
- Cost projection based on activities, time, number of people, professional fees, travel, and other costs

The process of preparing the ToR provides an opportunity to ensure all key stakeholders are involved in the process and that the purposes of the evaluation are clarified. Different stakeholders will frequently have different expectations and requirements, and it will often be necessary to have a process of negotiation. It is important to avoid placing too many demands on the evaluation, and sometimes it will be necessary to prioritize the requirements of different stakeholders.

Table A19.4-1 provides general guidelines for strengthening evaluation ToRs and identifies some of the key areas and issues for developing evaluation ToRs when operating under real-world budget, time, and data constraints. Often the ToRs for RealWorld Evaluation (RWE) are methodologically quite weak as many agencies have come to accept that the quality will be low when resources are limited. However, there are a number of practical ways in which evaluation designs can be strengthened, even when operating under severe budget constraints, and strengthening the ToR is an important element in ensuring better-quality evaluations.

The table identifies a number of preparatory steps that must be taken before the ToR is issued, ways to strengthen the ToR, and follow-up steps once interested consultants have responded to the RFP.

A number of important preparatory activities must be carried out before the ToR is prepared. These include consultations with stakeholders to clarify their information needs and how they plan to use the evaluation findings, as well as conducting a preliminary feasibility (evaluability) analysis to determine whether it will be possible to achieve the stated evaluation objectives with the current resource envelope and timelines. This preliminary analysis, which is often not conducted, will often show that the stated objectives of the evaluation are not feasible, in which case further discussion with stakeholders is required before the ToR is issued.

Some important points to stress in the ToR, particularly for RWE, are as follows:

- Provide a clear and explicit statement of the objectives of the evaluation, how the findings will be used, the required level of precision, and the kinds of management and/or policy decisions to which the findings will contribute.
- Define clearly the required minimum acceptable methodological standards for the evaluation. Some of the minimum requirements for evaluations operating on a limited budget and tight timeline might include (a) the definition of a counterfactual together with an explanation of the sources of data that will be used to test it; (b) a specified minimum number of meetings with nonbeneficiaries; (c) identifying and obtaining information on groups who may be negatively affected by the project; (d) selection of a sample of key informants who are familiar both with the project and with the broader political, economic, and sociocultural context within which it operates; and (e) methodological procedures for selecting participants for focus groups (if these are to be used) and for conducting the discussions and reporting the findings.
- The possibility should be considered of including additional resources to allow the evaluation team to contract a local resource person to assist in the planning and organization of the evaluation. This can include identification and assessment of potential secondary data sources, preparing field trips, identifying key informants, organizing focus groups, and ensuring that the sample of informants will not be limited to project beneficiaries.
- Do not simply state the objectives of the evaluation in technical or process terms. State clearly how the evaluation is expected to help the organization, particularly managers and policymakers.
- State clearly the quality assurance procedures (see Chapter 7) that will be used and the criteria that will be used to monitor and assess the quality of the evaluation design, implementation, and analysis. If a Threats-to-Validity Checklist is used (see Chapter 7), this should be included as an attachment to the ToR so that consultants fully understand the criteria to be used in the quality assurance.
- Avoid choosing too many questions.
- Require the sources to be given for each finding and recommendation in the evaluation report and check on the accuracy and adequacy of the sources. The executive summaries of many evaluation reports include statements like "many respondents stated that . . ." or "most women had encountered problems with respect to . . ." Quite often "many" or "most," in fact, only refer to one or two people attending focus groups or who were included in case studies. In other cases, some findings are not consistent with the evidence presented in the main report, or sometimes there is no evidence to support the statements. It is not unusual for the executive summary to present a more positive assessment of project effects than the evaluation findings actually justify. Many readers only read the executive summary, and if this presents their program in a favorable light, they may not check the validity of the findings[34] or the evidence on which they are based.

At the time of contract negotiation, the evaluators should be advised that all of their findings will be checked for accuracy and all must be documented. It is of course essential that the agency commissioning the evaluation does in fact systematically follow up and actually check the sources. Where the findings are not supported by the evidence, the consultants must be asked to revise the report or even in some cases to return to the field.

[34]An evaluation was commissioned in a South American country to assess the impact of rural roads on access to health, education, and other services. The executive summary reported that the construction of rural roads significantly increased women's utilization of rural health centers. This finding was widely quoted by the Ministry of Transport. In fact, the main report indicated that the impact of rural roads on women's use of health centers was quite limited, partly because the husband controlled the household budget and would often not give his wife the money for the bus fare if she "didn't look sick" or if he would have to mind the children while she was traveling, and partly because many rural communities did not believe in the utility of modern medicine. Unfortunately, not many people actually read the main report, so the positive impacts of rural roads continued to be cited.

TABLE A19.4-1 ⬢ Guidelines for Strengthening Evaluation ToR and Key Issues When Working Under RealWorld Evaluation Constraints

General Guidelines	Key Issues for RealWorld Evaluations
Note: Depending on the scope of the responsibilities given to consultants, it is possible that some of the activities classified as "Preparatory" might be assigned to consultants and included in the ToR.	
A. Preparatory activities	
1. Define the purposes of the evaluation and ensure the participation of key stakeholders. a. Clarify the information needs and expectations of different stakeholders. b. Prioritize information needs and if necessary determine how these could be reduced.	Consider the possibility of reducing the information needs (see Chapter 5) to accommodate budget and time constraints.
2. Define the resources and timelines. a. Define the budget (which may come from several sources). Clarify whether there is flexibility and circumstances in which the budget could be increased (or decreased). b. Define staff, other resources, information, and support to be provided by different stakeholders to the evaluation. c. Define the start and end dates of the evaluation and deadlines for deliverables. Clarify what determines the deadlines and how much (if any) flexibility there is.	Review strategies for addressing budget and time constraints (see Chapters 3 and 4) and assess whether any of these could be applied.
3. Assess the viability of producing the required analysis and deliverables within the available budget and timelines. a. If possible, conduct an assessment of the main sources of secondary data in terms of their availability, quality, and appropriateness for the present evaluation.	This is particularly critical in RWE contexts as budget and time constraints often mean that great reliance must be placed on secondary data. • Consider commissioning an assessment (prior to issuing the ToR) of the availability and adequacy of secondary data. • Use the program theory model (or create a theory model if it does not already exist) to define the time trajectory over which outcomes and impacts are expected to be achieved. Compare this with the evaluation timeline to determine whether it is feasible to generate the estimates of outcomes and impacts specified in the ToR. • If the proposed analysis is not feasible, consider the possibility of either extending the duration of the evaluation or reducing the kinds of analysis that are required.

General Guidelines	Key Issues for RealWorld Evaluations
B. Writing the ToR	
1. State clearly the objectives of the evaluation and define the following: a. The evaluation questions to be addressed b. Key stakeholders and their expected uses of the evaluation c. The overall evaluation approach to be adopted d. The products expected from the evaluation, when each needs to be submitted, and how each will be used e. The expertise required from evaluation team members f. Logistical arrangements	Given the resource constraints, it is important to define clearly the minimum methodological requirements for the evaluation design. For example, the ToR may state that a methodology must be proposed for identifying and collecting data on sectors of the target population that have not had access to the project.
2. Do not simply state the objectives in technical or process terms but make sure they will be understood by all stakeholders. Be clear on how the evaluation is expected to help the organization.	
3. Avoid choosing too many questions. It is better to have an evaluation that examines a few issues in depth than to look into a broad range of issues superficially.	Use the strategies discussed in Chapter 3 to identify any areas where the number or complexity of the questions could be reduced.
C. Follow-up (quality assurance) activities	
1. Encourage consultants to indicate the feasibility of complying with the deliverables and analysis defined in the ToR.	For many RWEs, the stated outputs of the evaluation cannot be achieved within the time limits of the evaluation. It is essential to give consultants an incentive and opportunity to negotiate the ToR and to question the proposed scope of the evaluation without fear of being disqualified.
2. Conduct an evaluability assessment to assess whether the proposed evaluation is feasible and whether it would be able to deliver the proposed outputs and analysis.	The ToRs for many RWEs make unrealistic demands, but many consultants accept the contract knowing that it will not be possible to comply with all of the conditions.
3. Be prepared to renegotiate either the scope of the evaluation and the resources or the timeline.	Renegotiation is required for many RWE ToRs.

Source: Morra-Imas & Rist (2009, pp. 443–444).

APPENDIX 19.5 BUILDING IN QUALITY ASSURANCE PROCEDURES

1. Defining Quality Assurance Procedures

Quality assurance refers to standardized management procedures for assessing the quality of the evaluation design, implementation, analysis, and dissemination. A key concern is to assess the extent to which the findings and recommendations of the evaluation are supported by the evaluation methodology and how it was implemented. Quality assurance procedures can include the following:

- *Preliminary feasibility analysis.* The feasibility analysis is used to ensure that there is a reasonable possibility that the purposes of the evaluation stated in the terms of reference (ToR) could be achieved with the available resources, within the proposed time frame, and at this point in the project cycle. The feasibility analysis should be conducted before the ToR is issued to avoid embarking on an evaluation that could not achieve its stated objectives. A number of factors must be considered. Two common reasons why the stated objectives could not be achieved are that the budget is insufficient to implement the proposed design and collect the required data and that insufficient time is allowed for the different stages of the design, implementation, and analysis of the evaluation. Timing must also take into account factors such as the onset of the rainy season or events such as public holidays, the start of an election campaign, and periods when national agencies will not be able to be actively involved in the evaluation due to pressures to prepare the annual budget or other priority activities. Another common issue is that the evaluation is commissioned when it is still too early in the project cycle to be able to measure the required outcomes or impacts. Finally, funding agencies often underestimate the time required for administrative arrangements such as approval and implementation of the project budget by the host government, the need to complete the hiring of key program staff before work can begin on the evaluation, procurement of local consultants or other local services, or the time required for review and approval of inception reports.

- *Evaluability analysis.* Once evaluation proposals are received (from agency staff if it is to be conducted internally or from consultants), an evaluability assessment may be conducted to determine whether (a) the proposal fully responds to the terms of reference, (b) the proposed methodology is technically sound, (c) it will be possible to implement the methodology within the budget and time constraints, and (d) there is a reasonable likelihood that the required data can be collected (either from secondary sources or through the proposed primary data collection procedures).

- *Application of a threats-to-validity worksheet.* A threats-to-validity worksheet can be used to identify potential threats to the validity of the evaluation findings and recommendations. The worksheet can be used at various points in the evaluation cycle: at the start of the evaluation as part of the evaluability analysis, midway through the evaluation, to assess the draft final report, or when the final report has been submitted. The worksheet can be used by funding agencies, the evaluation team, the project implementation agency, or national planning and policymaking agencies. Threats-to-validity worksheets are discussed later in this section.

The application of a threats-to-validity worksheet is discussed in Chapter 7, where it is pointed out that different approaches are required for assessing the validity of QUANT, QUAL, and mixed-method designs.

Why does the assessment of evaluation validity matter? If the conclusions of the evaluations are not methodologically sound, there is a risk of the following:

- Programs that do not work or that are not efficient and cost-effective may continue or be expanded.
- Good programs may be discontinued.
- Priority target groups may not have access to project benefits.
- Important lessons concerning which aspects of a program do and do not work and under what circumstances may be missed.

Even where methodologically rigorous QUANT, QUAL, or mixed-method evaluation designs are used, there are always factors that pose threats to the validity of the evaluation findings and recommendations (see Bamberger & White, 2007). Unfortunately, in the real world of development evaluation, it is frequently not possible to use the strongest evaluation designs, so the risks of arriving at wrong conclusions and providing wrong or misleading policy advice are much greater.

2. Using the Threats-to-Validity Worksheets

The worksheets for assessing validity for QUANT, QUAL, and mixed-method designs all have three parts, each of which is targeted to a different audience.

- *Part 1.* The cover sheet provides a one-page summary for senior management and for partner agencies. This explains the purpose of the evaluation and the reason for conducting the threats-to-validity assessment. It also summarizes the main conclusions of the validity assessment and the recommended follow-up actions. If the assessment concludes that the evaluation methodology was sound, the recommendation will normally be to accept the findings and recommendations of the evaluation report. However, if methodological weaknesses are identified, the assessment might recommend any of the following:
  - The evaluation report be accepted but that a covering memorandum might flag some of the areas where the findings and recommendations should be treated with caution
  - The consultants be requested to clarify some points
  - The consultants be required to conduct further analysis or in some cases return to the field to verify some of the information or to fill in gaps in the data (e.g., to interview some groups not represented in the focus groups)
  - The report be rejected as methodologically unsound
- *Part 2.* The summary assessment for each component is intended for midlevel management. It presents a half-page text summary of the validity assessment of each of the four or five components and a summary numerical rating (1 = very strong to 5 = serious problems). This provides sufficient detail for midlevel management to understand the main strengths and weaknesses of the evaluation and how these affect the validity of the findings and recommendations. In cases where only a general assessment of the evaluation quality is required, only Parts 1 and 2 of the worksheet may be used. However, when a more rigorous and comprehensive validity assessment is required, Part 3 can also be used.
- *Part 3.* This includes a set of checklists that assess dimensions such as objectivity, internal design validity, statistical conclusion validity, construct validity, external validity, and utilization validity. The checklists vary depending on whether the evaluation used a quantitative, qualitative, or mixed-method design. Rating scales are used to assess the seriousness and importance of each threat to validity. A set of summary scores is included in Part 2 to provide evaluation managers, many of whom are not evaluation specialists, with the key findings from the assessment and the recommended actions that should be taken.

Most quality assurance assessments will only use Parts 1 and 2 described above, as this provides a short, easy-to-understand, and relatively economical assessment of the strengths and weaknesses of an evaluation. The assessment can be prepared internally by an experienced researcher from the evaluation office or by an external consultant.

For a small number of the largest or most important evaluations, or where it is essential to ensure credibility when sensitive or controversial programs are being evaluated, a more comprehensive validity assessment may be required. In these cases, Part 3 can also be used. This provides a more in-depth and technical assessment, and in most cases, an external evaluation specialist will be contracted to conduct the assessment. This is both because of the technical nature of the assessment and also to ensure independence when credibility of the assessment is important.

When and How to Use the Validity Worksheet

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

- During the evaluation design phase to identify potential threats to validity or adequacy. When important problems or threats are identified, it may be possible to modify the design to address them. In other cases, if some of the potential threats could seriously compromise the purpose of the evaluation, further consultations may be required with the client or funding agency to consider increasing either the budget or the duration of the evaluation (where this would mitigate some of the problems) or agreeing to modify the objectives of the evaluation to reflect these limitations. In some extreme cases, the evaluability assessment may conclude that the proposed evaluation is not feasible, and all parties may agree to cancel or postpone the evaluation.
- During the implementation of the evaluation (for example a midterm review). If the worksheet had also been administered at the start of the evaluation, it is possible to assess if progress has been made in addressing the problems. Where serious problems are identified, it may be possible to adjust the evaluation design (e.g., to broaden the sample coverage or to refine or expand some of the questions or survey instruments).
- Toward the end of the evaluation—perhaps when the draft final report is being prepared. This may still allow time to correct some (but obviously not all) of the problems identified.
- When the evaluation has been completed. While it is now too late to make any corrections, a summary of the worksheet findings can be attached to the final report to provide a perspective for readers on how to interpret the evaluation findings and recommendations and to understand what caveats are required.
- For organizations that regularly commission or conduct evaluations, a very useful exercise is to conduct a meta-analysis to compare the ratings for different evaluations to determine whether there is a consistent pattern of methodological weaknesses in all evaluations (or all evaluations in a particular country, region, or sector). We discussed earlier how the worksheet can be used at different points in the evaluation—for example, at the evaluation design stage, during implementation, and when the draft final report is being prepared. When the scale is applied at these different points, it is possible to detect whether any of the threats are corrected or mitigated over time or whether, on the other hand, some of them get worse. Differential sample attrition (between the project and control groups) is a familiar example where a problem may get worse over time as differences between the characteristics of subjects remaining in the project and the control samples may increase.

3. Other Checklists

In addition to the checklists developed by Shadish, Cook, and Campbell (2002), Guba and Lincoln (1989), and Miles and Huberman (1994) (referred to in Appendices 7.1, 7.2, and 7.3), a number of other checklists have been developed for assessing the quality of evaluations or the validity of their conclusions. (See also Chapter 9 for a discussion of evaluation guidelines and standards.) The following are some widely used examples:

The Western Michigan University Evaluation Checklist Project[35]

The purpose of this project is to provide refereed checklists for designing, budgeting, contracting, staffing, managing, and assessing evaluations of programs, personnel, students, and other evaluands; collecting, analyzing, and reporting evaluation information; and determining merit, worth, and significance. Each checklist is a distillation of valuable lessons learned from practice. The site's stated purpose is to improve the quality and consistency of evaluations and enhance evaluation capacity through the promotion and use of high-quality checklists targeted to specific evaluation tasks and approaches. The checklists are classified into the following groups: Evaluation Management, Evaluation Models, Evaluation Values

[35]For information on the Western Michigan University Evaluation Center Checklist Project, see https://wmich.edu/evaluation/checklists.

and Criteria, Meta-Evaluation, and Evaluation Capacity Development and Institutionalization. Many, but not all, of the checklists focus on the education sector.

The site includes a number of widely cited checklists, among which are the following: Michael Scriven's *Key Evaluation Checklist* (2007), Daniel Stufflebeam's *Program Evaluations Metaevaluation Checklist* (1999) and *CIPP Model* (2007), and Michael Patton's *Qualitative Evaluation Checklist* (2003) and *Utilization Focused Evaluation Checklist* (2002c).

The American Evaluation Association's Guiding Principles for Evaluators (2004)

This provides guidance to evaluators in five areas: Systematic Enquiry, Competence, Integrity/Honesty, Respect for People, and Responsibilities for General and Public Welfare.

The Organization for Economic Cooperation and Development/Development Advisory Committee (OECD/DAC) *Quality Standards for Development Evaluation* (2010b). This provides standards for (1) the rationale, purpose, and objectives of an evaluation; (2) evaluation topic; (3) context; (4) evaluation methodology; (5) information sources; (6) independence; (7) evaluation ethics; (8) quality assurance; (9) relevance of the evaluation results; and (10) completeness.

While these three sources are extremely valuable resources, they focus mainly on ensuring quality in the design and implementation of the evaluation and ensuring that evaluators follow appropriate professional standards. There is little direct discussion of threats to validity and how they can be addressed.

APPENDIX 19.6 EVALUATION CAPACITY DEVELOPMENT

1. Target Groups for Evaluation Capacity Building

At least five groups of actors are actively involved in the evaluation process, and the success of most evaluations is dependent on the support and understanding of all of these stakeholder groups, each of which has different roles in the evaluation process and requires different sets of skills or knowledge. An evaluation capacity-building strategy should define ways to strengthen the evaluation skills of each of these groups:

- *Agencies that commission and fund evaluations.* These include donor agencies, foundations, government budget and funding agencies, and national and international nongovernmental organizations (NGOs).
- *Evaluation practitioners who design, implement, analyze, and disseminate evaluations.* These include evaluation units of line ministries, planning and finance ministries, national and international NGOs, foundations and donor agencies, evaluation consultants, and university research groups.
- *Evaluation users.* These include government, donor, and civil society organizations that use the results of evaluations to help formulate policies, allocate resources, and design and implement programs and projects.
- *Groups affected by the programs being evaluated.* These include community organizations, farmers' organizations, trade associations and business groups, trade unions and workers' organizations, and many other groups affected directly or indirectly by the programs and policies being evaluated.
- *Public opinion.* This includes broad categories such as the general public, the academic community, and civil society.

2. Required Evaluation Skills

Each of the different target groups is concerned with different aspects of the evaluation process and requires different skills or knowledge. Some of the broad categories of skills and knowledge include the following:

- Knowing when evaluations are required
- Understanding what evaluations can and cannot achieve and knowing what questions to ask
- Defining what evaluation clients "really want to know"
- Assessing the cost, time, and technical requirements for an evaluation
- How to promote, commission, or finance evaluations
- How to design, conduct, analyze, and disseminate evaluations
- How to use evaluation findings
- Designing evaluations that will be used
- How to assess the quality and utility of evaluations
- Adapting "ideal" evaluation methodology to real-world constraints

Table A19.6-1 summarizes the types of evaluation skills typically required by each of the five target groups. While there is obviously some overlap, the focus of the evaluation capacity needs is significantly different for each group. For example, the *agencies that fund evaluations* (Target Group 1) require skills in how to identify evaluation needs and when evaluations are needed, evaluating and selecting consultants, assessing evaluation proposals, estimating evaluation resource requirements, and understanding what questions an evaluation can and cannot answer.

On the other hand, *the groups affected by programs and policies* (Target Group 4) must be able to define when evaluations are required; negotiate with evaluators and funding agencies on the content, purpose, use, and dissemination of evaluations; ensure that the right questions are asked, that information is collected from the right people, and that the questions are formulated so as to avoid bias; understand, use, and disseminate the evaluation findings; and conduct an independent evaluation if it is necessary to challenge the findings of the "official" evaluation.

TABLE A19.6-1 ⬢ Evaluation Skills Required by Different Groups

Target Group	Examples	Evaluation Skills Needed
1. Funding agencies	• Donor agencies • Ministry of finance and ministry finance departments • Foundations • International NGOs	• Defining when evaluations are required • Evaluating evaluation consultants • Assessing proposals • Estimating evaluation resource requirements (budget, time, human resources)
2. Evaluation practitioners	• Evaluation units of line ministries • Evaluation departments of ministries of planning and finance • Evaluation units of NGOs • Evaluation consultants • Universities	• Defining client needs • Adapting theoretically sound designs to RealWorld budget, time, data, and political constraints • Understanding and selecting among different evaluation designs • Data collection and analysis • Sampling • Supervision • Institutional development • Adapting evaluation methodologies to the real world
3. Evaluation users	• Central government agencies (finance, planning, etc.) • Line ministries • NGOs • Foundations • Donor agencies	• Assessing the validity of quantitative evaluation designs and findings • Assessing the adequacy and validity of qualitative and mixed-method evaluation designs
4. Beneficiary populations (target groups)	• Community organizations • Farmers' organizations • Trade associations and business groups • Trade unions and workers' organizations	• Defining when evaluations are required • Negotiating with evaluators and funding agencies on the content, purpose, use, and dissemination of evaluations • Asking the right questions • Understanding and using evaluation findings • Participatory evaluations
5. Public opinion	• The general public • The academic community • Civil society	• How to get evaluations done • Participatory evaluations • Making sure the right questions get asked • Understanding and using evaluation findings

3. Designing and Delivering Evaluation Capacity Building

Evaluation capacity building can be delivered in many different ways, formally and informally, in long university or training institution programs, or very rapidly. In many cases, the target groups may not consider that they are involved in an evaluation process and may not even recognize that they are in a capacity-building exercise.

Some of the common evaluation capacity-building approaches include the following:

- Formal university or training institute programs. These can range from one or more academic semesters to seminars lasting from several days to several weeks.
- Workshops lasting from less than a day to 2 or 3 days
- Distance learning and other online programs
- Mentoring
- On-the-job training, where evaluation skills are learned as part of a package of work skills
- As part of a community development program
- As part of a community or group empowerment program

Table A19.6-2 illustrates different methods for evaluation capacity development and how they can be applied to different audiences. For example, evaluation users can strengthen their capacity to understand and use evaluation findings through briefings and short workshops, often with additional documentation available on websites.

An example of distance learning is the Brazilian Interlegis program organized by the Brazilian Senate to train municipal government functionaries on how to use social development indicators and evaluation studies to help identify priority areas for action. A particularly important function is helping mayors and other elected officials to use social indicators to compare their municipality with neighboring municipalities on key indicators such as school attendance, infant mortality, unemployment, and crime. This is useful not only to identify priority areas but also to evaluate performance. For functionaries at this level, it is much more meaningful to compare progress with neighboring municipalities than to try to understand and use the overwhelming amounts of information available from national Millennium Development Goal or Human Development Reports.

TABLE A19.6-2 ⬢ Examples of Different Capacity-Building Approaches Customized for Different Audiences

Audience	Type of Training	Duration	Example
Funding agencies	Workshops and seminars	½–3 days	• Chile: Briefings to Ministry of Finance and Parliamentary Budget Committee on findings and recommendations of evaluations of government programs[1]
Evaluation practitioners	Short courses on theory and practice using evaluation tools and techniques	1–2 weeks	• IPDET: 2-week basic course on evaluation principles and review of evaluation methods and approaches • AEA: 1-day workshop on RealWorld Evaluation. How to assess the validity and adequacy of evaluation designs and how to estimate budget, time, and human resources to conduct the evaluation[2]

Audience	Type of Training	Duration	Example
Evaluation users	• Debriefing workshop presenting the findings of an evaluation • Distance learning • Case studies and other short publications on how evaluations are used	0.5–1 day Short sessions once a week over a period of months UNICEF "webinars"	• Briefing workshops on the evaluation of World Bank Gender Policies • Beneficiary assessment studies • National Millennium Development Goals (MDG) workshops. Supported by mass media campaigns, publications, and websites. • Brazil: Interlegis distance learning program for municipal governments on how to use social development indicators and research findings to identify priority areas of action[3] • *Influential Evaluations:* Eight development evaluation case studies where there is convincing evidence the report influenced policy formulation or program design[4]
Affected populations	Participatory assessment and community consultations	0.5 day to 1 week	• Ethiopia: Assessing the impacts of participatory agricultural extension programs[5]
Public opinion	Mass media campaigns to disseminate Citizen Scorecards	Intensive campaign over several weeks consisting of short workshops, briefings, and extensive mass media coverage	• "Holding the State to Account" Citizen Report Card study in Bangalore, India. Mass media reporting on citizen attitudes to the quality of public services supported by workshops organized by NGOs[6]

1. For more information on the Chile program of impact evaluations and evaluations of government programs, see Lopez-Acevedo, Krause, & Mackay (2012).

2. International Program for Development Evaluation Training (IPDET). This program, organized by the World Bank in cooperation with Carleton University and supported by other donor agencies, is probably the most comprehensive training program available for development evaluation practitioners (see IPDET.org).

3. For more information about the Interlegis program (in Portuguese), see http://www.interlegis.gov.br.

4. "Influential Evaluations: Evaluations That Improved Performance and Impacts of Development Programs." 2004. Evaluation Capacity Development, Operations Evaluation Department. World Bank. Available at www.worldbank.org/oed/ecd.

5. A presentation on this study was made at the American Evaluation Association Conference in 2004.

6. A summary of this study is included in "Influential Evaluations" (see Note 4).

Appendix References

Ahmed, N. (2000). Study on gender dimensions of the second Bangladesh Rural Roads and Markets Improvement and Maintenance Project. Available at http://siteresources.worldbank.org/INTGENDERTRANSPORT/Resources/StudyReportonGenderDimensioninRuralRoadsandMarket.htm

Alderman, H., Kim, J., & Orazem, P. (2000). Design, evaluation and sustainability of private schools for the poor: The Pakistan Urban and Rural Fellowship School Experiments. (mimeo). Report on a study funded by the World Bank Economics of Girls Education and Girls Education Thematic Groups and the World Bank Development Research Group.

Alwin, D. (2009). Assessing the validity and reliability of timeline and event history data. In R. Belli, F. Stafford, & D. Alwin (Eds.), *Calendar and time diary methods in life course research* (pp. 277–301). Thousand Oaks, CA: Sage.

Aron, A., & Aron. (2002). *Statistics for the behavioral and social sciences: A brief course* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Baker, J. (2000). *Evaluating the impact of development projects on poverty: A handbook for practitioners*. Annex 1.2, "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Directions in Development. Washington, DC: World Bank.

Bamberger, M. (2004). *Gender evaluation.* Workshop delivered at the International Program for Development Evaluation Training (IPDET). Carleton University, Ottawa. Unpublished.

Bamberger, M., Blackden, M., Fort, L., & Manoukian, V. (2002). Gender. In *A sourcebook for poverty reduction strategies.* Washington, DC: World Bank. Available at http://worldbank.org/poverty/strategies/chapters/gender/gender.htm

Ban, R., & Rao, V. (2009). *Is deliberation equitable? Evidence from transcripts of village meetings in South Asia.* Policy Research Working Paper No. 4928. Washington, DC: World Bank.

Barron, P., Diprose, R., & Woolcock, M. (2011). Contesting development: Participatory projects and local conflict dynamics in Indonesia. Princeton, NJ: Princeton University Press.

Barwell, I. (1996). *Transport and the village.* World Bank Discussion Paper No. 344. Washington, DC.

Benhabib, S. (1995). Feminism and postmodernism. In S. Benhabib, J. Butler, D. Cornell, & N. Fraser (Eds.), *Feminist contentions: A philosophical exchange.* New York, NY: Routledge.

Blackden, M. (2001). *Too much work and too little time: Gender dimensions of transport, water and energy.* World Bank Institute Course on New Agendas for Poverty Reduction Strategies: Integrating Gender and Health. February 20–April 10, 2001.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswick* (pp. 81–106). New York, NY: Holt, Rinehart.

Chant, S., & Gutmann, M. (2000). *Mainstreaming men into gender and development*. Oxfam Working Papers. Oxford, England.

Claypool, L. (2010, December 14). Microbicide gel offers protection against HIV transmission [blog]. USAID. https://blog.usaid.gov/2010/12/microbicide-gel-offers-protection-against-hiv-transmission/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Davoodi, H., Tiongson, E., & Asawanuchit, S. (2003). *How useful are benefit incidence analyses of public education and health spending?* IMF Working Paper. International Monetary Fund. Washington, DC.

Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics, 87*(1), 1–19.

Donaldson, S. (2003). Theory driven program evaluation in the new millennium. In S. Donaldson & M. Scriven (Eds.), *Evaluating social programs and*

problems (pp. 109–141). Mahwah, NJ: Lawrence Erlbaum.

Fafchamps, M., & Quisumbing, A. (2001). *Assets at marriage in rural Ethiopia*. IFPRI. Available at www.ifpri.org

Figlio, D. (1995). The effect of drinking age laws and alcohol-related crashes: Time series evidence from Wisconsin. *Journal of Policy Analysis and Management, 14*(4), 55–66.

Fink, A. (2009). *How to conduct surveys: A step-by-step guide* (4th ed.). Thousand Oaks, CA: Sage.

Fitzpatrick, J., Christie, C., & Mark, M. (2009). *Evaluation in action: Interviews with expert evaluators.* Thousand Oaks, CA: Sage.

Fowler, F. J., & Cosenza, C. (2009). Design and evaluation of survey questions. In L. Bickman & D. Rog (Eds.), *The SAGE handbook of applied social research methods* (2nd ed., pp. 375–412). Thousand Oaks, CA: Sage.

Frankfort-Nachmias, C., & Leon-Guerrero, A. (2011). Social statistics for a diverse society (6th ed.). Thousand Oaks, CA: Sage.

Galasso, E., & Ravallion, M. (2004). Social protection in a crisis: Argentina's Plan Jefes y Jefas. *World Bank Economic Review, 18*(3), 367–400.

Gibson, C., & Woolcock, M. (2008). Empowerment, deliberative development and local level politics in Indonesia: Participatory projects as a source of countervailing power. *Studies in Comparative International Development, 43*(2), 151–180.

Grosh, M., & Glewwe, P. (Eds.). (2000). *Designing household survey questionnaires for developing countries: Lessons from 15 years of the Living Standards Measurement Study.* 3 vols. Washington, DC: World Bank.

Gross, B., van Wijk, C., & Mukherjee, N. (2000, December). *Linking sustainability with demand, gender and poverty: A study in community-managed water supply projects in 15 countries.* World Bank Water and Sanitation Program/IRC. International Water and Sanitation Centre.

Guggenheim, S. E. (2006). Crises and contradictions: Explaining a community development project in Indonesia. In A. Bebbington, S. E. Guggenheim, W. Olsen, & M. Woolcock, *The search for empowerment: Social capital as idea and practice at the World Bank* (pp. 111–144). Bloomfield, CT: Kumarian Press.

Hills, J., Le Grand, J., & Piachaud, D. (Eds.). (2002). *Understanding social exclusion*. New York, NY: Oxford University Press.

Horkheimer, M. (1937/1972). Traditional and critical theory. In *Critical theory: Selected essays* (pp. 188–243). New York, NY: Continuum.

Jamison, J. C, Karlan, D., & & Raffler, P. (2013). *Mixed method evaluation of a passive mhealth sexual information texting service in Uganda.* National Bureau of Economic Research Working Paper 19107. Retrieved from http:/www.nber.org/papers

Jolliffe, D. (2001). Measuring absolute and relative poverty: The sensitivity of estimated household consumption to survey design. *Journal of Economic and Social Measurement,* 27(1), 1–23.

Khandker, R. S., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on impact evaluation: Quantitative methods and practices.* Washington, DC: World Bank.

Khandker, S. (1998). *Fighting poverty with microcredit: Experience in Bangladesh.* Oxford, UK: Oxford University Press.

Kincheloe, J. L., & McLaren, P. (2003). Rethinking critical theory and qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The landscape of qualitative research*. Thousand Oaks, CA: Sage.

Kratochwill, T. R., Hitchcock, J., Homer, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case design technical documentation.* What Works Clearinghouse. Retrieved August 9, 2011, from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Lee, A., & Hills, J. (1998). *New cycles of disadvantage.* Report of a conference organized by the Centre for Social Exclusion for H.M Treasury. http://sticerd.lse.ac.uk/case

Lincoln Y. S., & Denzin, N. K (2003). The seventh moment: Out of the past. In N. K. Denzin &

Y. S. Lincoln (Eds.), *The landscape of qualitative research* (pp. 611–640). Thousand Oaks, CA: Sage.

Lindlof, T. R., & Taylor, B. C. (2002). *Qualitative communication research methods* (2nd ed.). Thousand Oaks, CA: Sage.

Litwin, M. (2003). *How to assess and interpret survey psychometrics* (2nd ed.). The Survey Kit, Vol. 2. Thousand Oaks, CA: Sage.

Lopez-Acevedo, G., Krause, P., & Mackay, K. (Eds.). (2012). *Building better policies: The nuts and bolts of monitoring and evaluation systems.* World Bank Training Series. Washington, DC: World Bank.

Malmberg-Calvo, C. (1994). *Women in rural transport.* SSTP Working Paper No. 11. World Bank and Economic Commission for Africa.

Malone, D. L. (1997). *Namel manmeri: Language and culture maintenance and mother tongue education in the highlands of Papua New Guinea* [PhD dissertation]. Indiana University, Bloomington, IN.

McLeod, J., & Thomson, R. (2009). *Researching social change.* Thousand Oaks, CA: Sage.

McLeod, J., & Yates, L. (2006). *Making modern lives: Subjectivity, schooling and social change.* Albany, NY: State University of New York Press.

Moore, D., & McCabe, G. (1999). *Introduction to the practice of statistics* (3rd ed.). New York, NY: Freeman.

Moser, C. (1993). *Gender planning and development: Theory, practice and training.* London: Routledge.

Murphy, J. (1995). *Gender issues in World Bank lending.* Operations Evaluation Department. Washington, DC: World Bank.

Murphy, J. (1997). *Mainstreaming gender in World Bank lending: An update.* Operations Evaluation Department. Washington, DC: World Bank.

Narayan, D., with Patel, R., Schafft, K., Rademacher, A., & Schulte, S. K. (2000). *Can anyone hear us? Voices of the poor* (Vol. 1). New York, NY: Oxford University 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 at www.worldbank.org/gender/prr

Organization for Economic Cooperation and Development (OECD)/Development Advisory Committee. (2010). *Quality standards for development evaluation.* DAC Guidelines and Reference Series. Paris: OECD. Retrieved August 4, 2011, from http://www.oecd.org/dataoecd/55/0/44798177.pdf

Overholt, C., Anderson, M., Cloud, K., & Austin, J. (1985b). Women in development: A framework for project analysis. In C. Overholt, M. Anderson, K. Cloud, & J. Austin (Eds.), *Gender roles in development projects: A case book*. West Hartford, CT: Kumarian Press.

Patton, M. Q. (2002c). *Utilization focused evaluation: Checklist.* Kalamazoo, MI: Western Michigan University, The Evaluation Center. http://www.wmich.edu/evalctr/archive_checklists/ufe.pdf

Patton, M. Q. (2003). *Qualitative evaluation checklist.* Kalamazoo, MI: Western Michigan University, The Evaluation Center. http://www.wmich.edu/evalctr/archive_checklists/qec.pdf

Pitt, M., & Khandker, S. (1998). The impact of group-based credit programs on poor households in Bangladesh: Does the gender of participants matter? *Journal of Political Economy, 106,* 958–996.

Pitt, M., Khandker, S., McKernan, S.-M., & Latif, M. A. (1999). Credit programs for the poor and reproductive behaviour in low-income countries: Are the reported causal relationships the result of heterogeneity bias?" *Demography, 36*(1), 1–21.

Quisumbing, A. R., & Maluccio, J. A. (1999). *Intrahousehold Allocation and Gender Relations: New Empirical Evidence.* Working Paper Series 2, Policy Research Report on Gender and Development, World Bank, Development Research Group/Poverty Reduction and Economic Management Network, Washington, DC.

Rao, N. (1999). Cycling into the future: The experience of women in Pudukkottai, Tamil Nadu. *Case study presented*

at the International Forum for Rural Transport and Development (IFRTD) workshop in Sri Lanka. June 1999.

Rawlings, L. (2000). Evaluating Nicaragua's school-based management reform. In M. Bamberger (Ed.), *Integrating quantitative and qualitative research in development projects*. Washington, DC: World Bank.

Rosenbaum, P. R. (1995). Design sensitivity in observational Studies. *Biometrika, 91*(1), 153–164.

Salkind, N. (2008). *Statistics for people who (think they) hate statistics.* Thousand Oaks, CA: Sage.

Scott, C., & Amenuvegbe, B. (1991). Recall loss and recall duration: An experimental study in Ghana. *Inter-Stat, 4*(1), 31–55.

Scott, K., & Okrasa, W. (1998). *Analysis of Latvia Diary Experiment.* Washington, DC: World Bank, Development Research Group.

Sen, A. (1993). Capability and wellbeing. In M. Nussbaum & A. Sen (Eds.), *The quality of life.* Oxford: Clarendon.

Sirkin, R. (1999). *Statistics for the social sciences.* Thousand Oaks, CA: Sage.

Spector, P. (1991). *Summated rating scale construction: An introduction.* Quantitative Applications in the Social Sciences, vol. 82. Newbury Park, CA: Sage.

Stufflebeam, D. L. (1999). *Program evaluations metaevaluation checklist (based on the program evaluation standards).* Kalamazoo, MI: Western Michigan University. The Evaluation Center. http://www.wmich.edu/evalctr/archive_checklists/program_metaeval.pdf

Sudman, S., & Bradburn, N. (1982). *Asking questions.* San Francisco, CA: Jossey-Bass.

United Nations Development Program (UNDP). (1995). *Human development report. Gender and human development.* New York, NY: Oxford University Press.

United Nations Development Program (UNDP). (2000). *Human development report*. New York, NY: Oxford University Press.

Williams, S. (1994). *The Oxfam gender training manual.* Oxford, UK: OXFAM Publications.

World Bank. (2000). *Attacking poverty.* World Development Report 2000/2001. Washington, DC: World Bank.

World Bank. (2001). *Engendering development: Through gender equality in rights, resources and voice.* Oxford, UK: Oxford University Press.

Yin, R. (Ed.). (2004). *The case study anthology.* Thousand Oaks, CA: Sage.