

**DIRECTIONS IN DEVELOPMENT**

# Evaluating the Impact of Development Projects on Poverty

A Handbook for Practitioners

Judy L. Baker

*The World Bank  
Washington, D.C.*

© 2000 The International Bank for Reconstruction  
and Development/THE WORLD BANK  
1818 H Street, N.W.  
Washington, D.C. 20433

All rights reserved  
Manufactured in the United States of America  
First printing May 2000

The findings, interpretations, and conclusions expressed in this paper are entirely those of the author(s) and should not be attributed in any manner to the World Bank, to its affiliated organizations, or to members of its Board of Executive Directors or the countries they represent. The World Bank does not guarantee the accuracy of the data included in this publication and accepts no responsibility for any consequence of their use.

The material in this publication is copyrighted. The World Bank encourages dissemination of its work and will normally grant permission to reproduce portions of the work promptly.

Permission to *photocopy* items for internal or personal use, for the internal or personal use of specific clients, or for educational classroom use is granted by the World Bank, provided that the appropriate fee is paid directly to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA.; telephone 978-750-8400, fax 978-750-4470. Please contact the Copyright Clearance Center before photocopying items.

For permission to *reprint* individual articles or chapters, please fax a request with complete information to the Republication Department, Copyright Clearance Center, fax 978-750-4470.

All other queries on rights and licenses should be addressed to the Office of the Publisher, World Bank, at the address above or faxed to 202-522-2422.

ISBN 0-8213-4697-0

### Library of Congress Cataloging-in-Publication Data

Baker, Judy L., 1960–

Evaluating the impact of development projects on poverty : a handbook  
for practitioners / Judy L. Baker

p. cm. — (Directions in development)

Includes bibliographical references.

ISBN 0-8213-4697-0

1. Economic development projects—Evaluation—Handbooks, manuals, etc.
2. Poor—Developing countries. I. Title II. Directions in development (Washington, D.C.)

HD75.9 .B35 2000  
338.9'0068'4—dc21

00-028325

---

---

# Contents

|  |            |
|--|------------|
| Foreword   | .vi        |
| Acknowledgments  | .viii      |
| 1 Defining Concepts and Techniques for<br>Impact Evaluation                                      | .1         |
| 2 Key Steps in Designing and Implementing<br>Impact Evaluations                                  | .16        |
| 3 Applying Analytical Methods for Impact Evaluation:<br>A Case Study                             | .40        |
| 4 Drawing on “Good Practice” Impact Evaluations  | .65        |
| Bibliography   | .83        |
| <br><b>Annexes</b>   |            |
| <b>Annex 1: Case Studies</b>   | <b>.94</b> |
| 1.1 Evaluating the Gains to the Poor from Workfare:<br>Argentina’s TRABAJAR Program              | .94        |
| 1.2 Does Microfinance Really Help the Poor? New Evidence<br>from Flagship Programs in Bangladesh | .101       |

|      |  |            |
|------|--|------------|
| 1.3  | Bangladesh Food for Education: Evaluating a Targeted Social Program When Placement Is Decentralized: . . .                                   | 105        |
| 1.4  | Evaluating Bolivia's Social Investment Fund . . . . .  | 109        |
| 1.5  | Impact of Active Labor Programs: Czech Republic . . .  | 114        |
| 1.6  | Impact of Credit with Education on Mothers' and their Young Children's Nutrition: Lower Pra Rural Bank Program in Ghana . . . . .            | 119        |
| 1.7  | Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya . . . . .   | 123        |
| 1.8  | Evaluating Kenya's Agricultural Extension Project . .  | 128        |
| 1.9  | The Impact of Mexico's Retraining Program on Employment and Wages (PROBECAT) . . . . .   | 134        |
| 1.10 | Mexico, National Program of Education, Health, and Nutrition (PROGRESA) . . . . .  | 140        |
| 1.11 | Evaluating Nicaragua's School Reform: A Combined Quantitative-Qualitative Approach . . . . .   | 145        |
| 1.12 | Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement . . . . . | 151        |
| 1.13 | The Impact of Alternative Cost-Recovery Schemes on Access and Equity in Niger . . . . .  | 156        |
| 1.14 | Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments . . . . .                                  | 160        |
| 1.15 | Assessing the Poverty Impact of Rural Roads Projects in Vietnam . . . . .  | 165        |
|      | <b>Annex 2: Sample Terms of Reference . . . . .</b>  | <b>169</b> |
| 2.1  | The Uganda Nutrition and Early Childhood Development Project . . . . .   | 169        |
| 2.2  | Rural Roads Impact Evaluation: Vietnam 1997 Baseline . . . . .   | 188        |
|      | <b>Annex 3: A Sample Budget from an Impact Evaluation of a School Feeding Program . . . . .</b>  | <b>195</b> |
|      | <b>Annex 4: Impact Indicators—Evaluation of Bolivia Social Investment Fund . . . . .</b>   | <b>198</b> |
|      | <b>Annex 5: Template of Log Frame for Project Design Summary for the Project Completion Document or Project Appraisal Document . . . . .</b> | <b>204</b> |
|      | <b>Annex 6: Matrix of Analysis . . . . .</b>   | <b>208</b> |

**Boxes**

1.1 The Problem of Selection Bias . . . . . 5

1.2 Summary of Quantitative Methods for Evaluating  
Program Impact . . . . . 6

1.3 Summary of Methods Used to Evaluate  
Adjustment Policies . . . . . 11

2.1 Main Steps in Designing and Implementing Impact  
Evaluations . . . . . 17

2.2 Key Points for Identifying Data Resources for Impact  
Evaluation . . . . . 21

3.1 Steps in Propensity Score Matching . . . . . 50

3.2 Sources of Bias in Naïve Estimates of  
PROSCOL’s Impact . . . . . 53

3.3 Doing a Double Difference . . . . . 56

3.4 Poverty Measures . . . . . 59

3.5 Comparing Poverty with and without the Program . . . 60

**Tables**

2.1 Evaluation Methods and Corresponding Data  
Requirements . . . . . 28

2.2 Main Data Collection Instruments for Impact  
Evaluation . . . . . 32

4.1 Summary of "Good-Practice" Impact Evaluations . . . . . 67

4.2 Summary of Estimated Costs from Several World Bank  
Impact Evaluations . . . . . 79

---

---

# Foreword

Despite the billions of dollars spent on development assistance each year, there is still very little known about the actual impact of projects on the poor. There is broad evidence on the benefits of economic growth, investments in human capital, and the provision of safety nets for the poor. But for a specific program or project in a given country, is the intervention producing the intended benefits and what was the overall impact on the population? Could the program or project be better designed to achieve the intended outcomes? Are resources being spent efficiently? These are the types of questions that can only be answered through an impact evaluation, an approach that measures the outcomes of a program intervention in isolation of other possible factors.

Many governments, institutions, and project managers are reluctant to carry out impact evaluations because they are deemed to be expensive, time consuming, and technically complex, and because the findings can be politically sensitive, particularly if they are negative. Many evaluations have also been criticized because the results come too late, do not answer the right questions, or were not carried out with sufficient analytical rigor. A further constraint is often the limited availability and quality of data.

Yet with proper and early planning, the support of policymakers, and a relatively small investment compared with overall project cost, a rigorous evaluation can be very powerful in assessing the appropriateness and effectiveness of programs. Evaluating impact is particularly critical in developing countries where resources are scarce and every dollar spent should aim to maximize its impact on poverty reduction. If programs are poorly designed, do not reach their intended beneficiaries, or are wasteful, with the right information they can be redesigned, improved, or eliminated if deemed necessary. The knowledge gained from impact evaluation studies will also provide critical input to the appropriate design of future programs and projects.

This handbook seeks to provide project managers and policy analysts with the tools needed for evaluating project impact. It is aimed at read-

ers with a general knowledge of statistics. For some of the more in-depth statistical methods discussed, the reader is referred to the technical literature on the topic. Chapter 1 presents an overview of concepts and methods, Chapter 2 discusses key steps and related issues to consider in implementation, Chapter 3 illustrates various analytical techniques through a case study, and Chapter 4 includes a discussion of lessons learned from a rich set of “good practice” evaluations of poverty projects that have been reviewed for this handbook. The case studies, included in Annex I, were selected from a range of evaluations carried out by the World Bank, other donor agencies, research institutions, and private consulting firms. They were chosen for their methodological rigor, in an attempt to cover a broad mix of country settings, types of projects, and evaluation methodologies. Also included in the Annexes are samples of the main components that would be necessary in planning any impact evaluation—sample terms of reference, a budget, impact indicators, a log frame, and a matrix of analysis.

Although the techniques used in impact evaluation are similar across sectors and population subgroups, the illustrations of methodologies and case examples in the handbook focus on assessing the impact of projects targeted to the poor. Poverty impact can include a wide range of projects and evaluation questions, such as measuring the impact of microfinance programs on household income, the impact of a training program on employment, the impact of a school feeding program on student attendance, or the impact of the construction of rural roads on household welfare.

Regardless of the project type or questions being addressed, the design of each impact evaluation will be unique, depending on factors such as the type of data available, local capacity, and timing and budget concerns. Finally, evaluations that will yield high-quality, credible, and generalizable results for policymakers will require strong financial and political support; early and careful planning; participation of stakeholders in the design of the objectives and approach of the study; adequate data; a suitable mix of methodologies, including both quantitative and qualitative techniques; the rigorous application of these techniques; and communication between team members throughout the process.

---

---

# Acknowledgments

The preparation of this book benefited from the invaluable contributions of a core team. I would like to acknowledge both the written input, and helpful comments along the way, from the following team members: Gillette Hall (case studies, lessons learned), Julia Lane (case studies, lessons learned), Martin Ravallion (analytical methods case study), and Laura Rawlings (implementation issues, lessons learned); and the work on impact evaluation carried out by Kene Ezemenari, Gloria Rubio, Anders Rudqvist, and K. Subbarao. Background research was carried out by Matthew Fleming and Samir Stewart. The book was jointly supported by the Latin America and Caribbean Region and the Poverty Reduction and Economic Management Network of the World Bank under the leadership of Norman Hicks, Guillermo Perry, and Michael Walton. The work also benefited greatly from the comments received by Omar Arias, Sabina Alkire, Michael Bamberger, Soniya Carvalho, Wendy Cunningham, Norman Hicks, Shahidur Khandker, Norbert Schady, and Quentin Wodon.

---

---

# Chapter 1

## Defining Concepts and Techniques for Impact Evaluation

A comprehensive evaluation is defined in the literature as an evaluation that includes monitoring, process evaluation, cost-benefit evaluation, and impact evaluation. Yet each of these components is distinctly different. Monitoring will help to assess whether a program is being implemented as was planned. A program monitoring system enables continuous feedback on the status of program implementation, identifying specific problems as they arise. Process evaluation is concerned with how the program operates and focuses on problems in service delivery. Cost-benefit or cost-effectiveness evaluations assess program costs (monetary or non-monetary), in particular their relation to alternative uses of the same resources and to the benefits being produced by the program. And finally, impact evaluation is intended to determine more broadly whether the program had the desired effects on individuals, households, and institutions and whether those effects are attributable to the program intervention. Impact evaluations can also explore unintended consequences, whether positive or negative, on beneficiaries. Of particular interest for this handbook is the extent to which project benefits reach the poor and the impact that these benefits have on their welfare. Some of the questions addressed in impact evaluation include the following: How did the project affect the beneficiaries? Were any improvements a direct result of the project, or would they have improved anyway? Could program design be modified to improve impact? Were the costs justified?

These questions cannot, however, be simply measured by the outcome of a project. There may be other factors or events that are correlated with the outcomes but are not caused by the project. To ensure methodological rigor, an impact evaluation must estimate the counterfactual, that is, what would have happened had the project never taken place or what otherwise would have been true. For example, if a recent graduate of a labor training program becomes employed, is it a direct result of the program or would that individual have found work anyway? To determine the counterfactual, it is necessary to net out the effect of the interventions from other factors—a somewhat complex task. This is accomplished through the use of comparison or control groups (those who do not participate in a program or receive benefits), which are subsequently compared with the treatment group (individuals who do receive the intervention). Control

groups are selected randomly from the same population as the program participants, whereas the comparison group is more simply the group that does not receive the program under investigation. Both the comparison and control groups should resemble the treatment group in every way, the only difference between groups being program participation.

Determining the counterfactual is at the core of evaluation design. This can be accomplished using several methodologies which fall into two broad categories, experimental designs (randomized), and quasi-experimental designs (nonrandomized). It is, however, quite tricky to net out the program impact from the counterfactual conditions that can be affected by history, selection bias, and contamination. Qualitative and participatory methods can also be used to assess impact. These techniques often provide critical insights into beneficiaries' perspectives, the value of programs to beneficiaries, the processes that may have affected outcomes, and a deeper interpretation of results observed in quantitative analysis. The strengths and weaknesses of each of these methods are discussed in more detail below. As the reader will find, no technique is perfect and thus the evaluator must make decisions about the tradeoffs for each method chosen. Early and careful planning will, however, provide many more methodological options in designing the evaluation.

## Experimental Designs

Experimental designs, also known as randomization, are generally considered the most robust of the evaluation methodologies. By randomly allocating the intervention among eligible beneficiaries, the assignment process itself creates comparable treatment and control groups that are statistically equivalent to one another, given appropriate sample sizes. This is a very powerful outcome because, in theory, the control groups generated through random assignment serve as a perfect counterfactual, free from the troublesome selection bias issues that exist in all evaluations. The main benefit of this technique is the simplicity in interpreting results—the program impact on the outcome being evaluated can be measured by the difference between the means of the samples of the treatment group and the control group. One example is the Kenya textbooks evaluation in which evaluators selected a random allocation of program sites, administered a baseline survey, created control groups, and then administered the treatment, which in this case was the delivery of textbooks. Having control and treatment groups then allowed the evaluators to clearly determine the impact of textbooks on student learning.

While experimental designs are considered the optimum approach to estimating project impact, in practice there are several problems. First, randomization may be unethical owing to the denial of benefits or ser-

vices to otherwise eligible members of the population for the purposes of the study. An extreme example would be the denial of medical treatment that can turn out to be lifesaving to some members of a population. Second, it can be politically difficult to provide an intervention to one group and not another. Third, the scope of the program may mean that there are no nontreatment groups such as with a project or policy change that is broad in scope—examples include an adjustment loan or programs administered at a national level. Fourth, individuals in control groups may change certain identifying characteristics during the experiment that could invalidate or contaminate the results. If, for example, people move in and out of a project area, they may move in and out of the treatment or control group. Alternatively, people who were denied a program benefit may seek it through alternative sources, or those being offered a program may not take up the intervention. Fifth, it may be difficult to ensure that assignment is truly random. An example of this might be administrators who exclude high-risk applicants to achieve better results. And finally, experimental designs can be expensive and time consuming in certain situations, particularly in the collection of new data.

With careful planning, some of these problems can be addressed in the implementation of experimental designs. One way is with the random selection of beneficiaries. This can be used to provide both a politically transparent allocation mechanism and the basis of a sound evaluation design, as budget or information constraints often make it impossible to accurately identify and reach the most eligible beneficiaries. A second way is bringing control groups into the program at a later stage once the evaluation has been designed and initiated. In this technique, the random selection determines *when* the eligible beneficiary receives the program, not *if* they receive it. This was done in the evaluation of a nutrition program in Colombia, which provided the additional benefit of addressing questions regarding the necessary time involved for the program to become effective in reducing malnutrition (McKay 1978). Finally, randomization can be applied within a subset of equally eligible beneficiaries, while reaching all of the most eligible and denying benefits to the least eligible, as was done with education projects in the El Chaco region for the Bolivia social fund evaluation (Pradhan, Rawlings, and Ridder 1998). However, if the latter suggestion is implemented, one must keep in mind that the results produced from the evaluation will be applicable to the group from which the randomly generated sample was selected.

### Quasi-Experimental Designs

Quasi-experimental (nonrandom) methods can be used to carry out an evaluation when it is not possible to construct treatment and comparison

groups through experimental design. These techniques generate comparison groups that resemble the treatment group, at least in observed characteristics, through econometric methodologies, which include matching methods, double difference methods, instrumental variables methods, and reflexive comparisons (see Box 1.2). When these techniques are used, the treatment and comparison groups are usually selected *after* the intervention by using nonrandom methods. Therefore, statistical controls must be applied to address differences between the treatment and comparison groups and sophisticated matching techniques must be used to construct a comparison group that is as similar as possible to the treatment group. In some cases a comparison group is also chosen before the treatment, though the selection is not randomized.

The main benefit of quasi-experimental designs is that they can draw on existing data sources and are thus often quicker and cheaper to implement, and they can be performed after a program has been implemented, given sufficient existing data. The principal disadvantages of quasi-experimental techniques are that (a) the reliability of the results is often reduced as the methodology is less robust statistically; (b) the methods can be statistically complex; and (c) there is a problem of selection bias. In generating a comparison group rather than randomly assigning one, many factors can affect the reliability of results. Statistical complexity requires considerable expertise in the design of the evaluation and in analysis and interpretation of the results. This may not always be possible, particularly in some developing country circumstances.

The third problem of bias relates to the extent to which a program is participated in differentially by subgroups of a target population, thus affecting the sample and ultimately the results. There are two types of bias: those due to differences in observables or something in the data, and those due to differences in unobservables (not in the data), often called selection bias (Box 1.1). An observable bias could include the selection criteria through which an individual is targeted, such as geographic location, school attendance, or participation in the labor market. Unobservables that may bias program outcomes could include individual ability, willingness to work, family connections, and a subjective (often politically driven) process of selecting individuals for a program. Both types of biases can yield inaccurate results, including under- and overestimates of actual program impacts, negative impacts when actual program impacts are positive (and vice versa), and statistically insignificant impacts when actual program impacts are significant and vice versa. (See, for example, LaLonde 1986, Fraker and Maynard 1987, LaLonde and Maynard 1987, and Friedlander and Robins 1995.) It is possible to control for bias through statistical techniques such as matching and instrumental variables, but it is very difficult to fully remove them

which remains a major challenge for researchers in the field of impact analysis.

Among quasi-experimental design techniques, matched-comparison techniques are generally considered a second-best alternative to experimental design. The majority of the literature on evaluation methodology is centered around the use of this type of evaluation, reflecting both the frequency of use of matched comparisons and the many challenges posed by having less-than-ideal comparison groups. In recent years there have been substantial advances in propensity score matching techniques (Rosenbaum and Rubin 1985; Jalan and Ravallion 1998). This method is

### **Box 1.1 The Problem of Selection Bias**

Selection bias relates to unobservables that may bias outcomes (for example, individual ability, preexisting conditions). Randomized experiments solve the problem of selection bias by generating an experimental control group of people who would have participated in a program but who were randomly denied access to the program or treatment. The random assignment does not remove selection bias but instead balances the bias between the participant and non-participant samples. In quasi-experimental designs, statistical models (for example, matching, double differences, instrumental variables) approach this by modeling the selection processes to arrive at an unbiased estimate using nonexperimental data. The general idea is to compare program participants and nonparticipants holding selection processes constant. The validity of this model depends on how well the model is specified.

A good example is the wages of women. The data represent women who choose to work. If this decision were made, we could ignore the fact that not all wages are observed and use ordinary regression to estimate a wage model. Yet the decision by women to work is not made randomly—women who would have low wages may be unlikely to choose to work because their personal reservation wage is greater than the wage offered by employers. Thus the sample of observed wages for women would be biased upward.

This can be corrected for if there are some variables that strongly affect the chances for observation (the reservation wage) but not the outcome under study (the offer wage). Such a variable might be the number of children at home.

*Source:* Greene (1997).

very appealing to evaluators with time constraints and working without the benefit of baseline data given that it can be used with a single cross-section of data. This technique is, however, dependent on having the right data because it relies on oversampling program beneficiaries during the fielding of a larger survey and then “matching” them to a comparison group selected from the larger core sample of the overall effort, often a national household survey. Given the growth in the applications of large surveys in developing countries, such as the multipurpose Living Standards Measurement Studies, this evaluation method seems particularly promising. A good example is the evaluation of a public works program, TRABAJAR, in Argentina (Jalan and Ravallion 1998, Annex 1.1, and chapter 4).

### **Box 1.2 Summary of Quantitative Methods for Evaluating Program Impact**

The main methods for impact evaluation are discussed below. Because no method is perfect, it is always desirable to triangulate.

#### *Experimental or Randomized Control Designs*

- *Randomization*, in which the selection into the treatment and control groups is random within some well-defined set of people. In this case there should be no difference (in expectation) between the two groups besides the fact that the treatment group had access to the program. (There can still be differences due to sampling error; the larger the size of the treatment and control samples the less the error.)

#### *Nonexperimental or Quasi-Experimental Designs*

- *Matching methods or constructed controls*, in which one tries to pick an ideal comparison that matches the treatment group from a larger survey. The most widely used type of matching is *propensity score matching*, in which the comparison group is matched to the treatment group on the basis of a set of observed characteristics or by using the “propensity score” (predicted probability of participation given observed characteristics); the closer the propensity score, the better the match. A good comparison group

comes from the same economic environment and was administered the same questionnaire by similarly trained interviewers as the treatment group.

- *Double difference or difference-in-differences* methods, in which one compares a treatment and comparison group (first difference) before and after a program (second difference). Comparators should be dropped when propensity scores are used and if they have scores outside the range observed for the treatment group.
- *Instrumental variables or statistical control* methods, in which one uses one or more variables that matter to participation but not to outcomes given participation. This identifies the exogenous variation in outcomes attributable to the program, recognizing that its placement is not random but purposive. The “instrumental variables” are first used to predict program participation; then one sees how the outcome indicator varies with the predicted values.
- *Reflexive comparisons*, in which a baseline survey of participants is done before the intervention and a follow-up survey is done after. The baseline provides the comparison group, and impact is measured by the change in outcome indicators before and after the intervention.

## Qualitative Methods

Qualitative techniques are also used for carrying out impact evaluation with the intent to determine impact by the reliance on something other than the counterfactual to make a causal inference (Mohr 1995). The focus instead is on understanding processes, behaviors, and conditions as they are perceived by the individuals or groups being studied (Valadez and Bamberger 1994). For example, qualitative methods and particularly participant observation can provide insight into the ways in which households and local communities perceive a project and how they are affected by it. Because measuring the counterfactual is at the core of impact analysis techniques, qualitative designs have generally been used in conjunction with other evaluation techniques. The qualitative approach uses relatively open-ended methods during design, collection of data, and analysis. Qualitative data can also be quantified. Among the methodologies used in qualitative impact assessments are the techniques developed for rapid rural assessment, which rely on participants’ knowledge of the conditions surrounding the project or program being evaluated, or par-

ticipatory evaluations in which stakeholders are involved in all stages of the evaluation—determining the objectives of the study, identifying and selecting indicators to be used, and participating in data collection and analysis. For a detailed discussion on participatory methods see World Bank (1996), *The World Bank Participation Sourcebook*.

The benefits of qualitative assessments are that they are flexible, can be specifically tailored to the needs of the evaluation using open-ended approaches, can be carried out quickly using rapid techniques, and can greatly enhance the findings of an impact evaluation through providing a better understanding of stakeholders' perceptions and priorities and the conditions and processes that may have affected program impact.

Among the main drawbacks are the subjectivity involved in data collection, the lack of a comparison group, and the lack of statistical robustness, given mainly small sample sizes, all of which make it difficult to generalize to a larger, representative population. The validity and reliability of qualitative data are highly dependent on the methodological skill, sensitivity, and training of the evaluator. If field staff are not sensitive to specific social and cultural norms and practices, and nonverbal messages, the data collected may be misinterpreted. And finally, without a comparison group, it is impossible to determine the counterfactual and thus causality of project impact.

### **Integrating Quantitative and Qualitative Methods**

Although there is an extensive literature on quantitative versus qualitative methods in impact evaluation, there is also a growing acceptance of the need for integrating the two approaches. Impact evaluations using quantitative data from statistically representative samples are better suited to assessing causality by using econometric methods or reaching generalizable conclusions. However, qualitative methods allow the in-depth study of selected issues, cases, or events and can provide critical insights into beneficiaries' perspectives, the dynamics of a particular reform, or the reasons behind certain results observed in a quantitative analysis. There are significant tradeoffs in selecting one technique over another.

Integrating quantitative and qualitative evaluations can often be the best vehicle for meeting the project's information needs. In combining the two approaches, qualitative methods can be used to inform the key impact evaluation questions, survey the questionnaire or the stratification of the quantitative sample, and analyze the social, economic, and political context within which a project takes place, whereas quantitative methods can be used to inform qualitative data collection strategies, to design the sample to inform the extent to which the results observed in

the qualitative work can be generalized to a larger population by using a statistically representative sample, and, statistical analysis can be used to control for household characteristics and the socio-economic conditions of different study areas, thereby eliminating alternative explanations of the observed outcomes.

There are several benefits of using integrated approaches in research discussed in Bamberger (2000) that also apply to impact evaluations. Among them:

- Consistency checks can be built in through the use of triangulation procedures that permit two or more independent estimates to be made for key variables (such as income, opinions about projects, reasons for using or not using public services, and specific impact of a project).
- Different perspectives can be obtained. For example, although researchers may consider income or consumption to be the key indicators of household welfare, case studies may reveal that women are more concerned about vulnerability (defined as the lack of access to social support systems in times of crises), powerlessness, or exposure to violence.
- Analysis can be conducted on different levels. Survey methods can provide good estimates of individual, household, and community-level welfare, but they are much less effective for analyzing social processes (social conflict, reasons for using or not using services, and so on) or for institutional analysis (how effectively health, education, credit, and other services operate and how they are perceived by the community). There are many qualitative methods designed to analyze issues such as social process, institutional behavior, social structure, and conflict.
- Opportunities can be provided for feedback to help interpret findings. Survey reports frequently include references to apparent inconsistencies in findings or to interesting differences between communities or groups that cannot be explained by the data. In most quantitative research, once the data collection phase is completed it is not possible to return to the field to check on such questions. The greater flexibility of qualitative research means that it is often possible to return to the field to gather additional data. Survey researchers also use qualitative methods to check on outliers—responses that diverge from the general patterns. In many cases the data analyst has to make an arbitrary decision as to whether a household or community that reports conditions that are significantly above or below the norm should be excluded (on the assumption that it reflects a reporting error) or the figures adjusted. Qualitative methods permit a rapid follow-up in the field to check on these cases.

In practice, the integration of quantitative and qualitative methods should be carried out during each step of the impact evaluation. Chapter 2 mentions many opportunities for doing this. For illustration, the Nicaragua School Autonomy Reform Case provides a good example of integrated methods. Quantitative methods following a quasi-experimental design were used to determine the relationship between decentralized management and learning and to generalize results for different types of schools. In addition, qualitative techniques, including a series of key informant interviews and focus group discussions with different school-based staff and parents, were utilized to analyze the context in which the reform was introduced, examine the decisionmaking dynamics in each school, and assess the perspectives of different school community actors on the autonomy process (see Annex 1.11).

### Other Approaches to Impact Evaluation

Two other topics are particularly relevant to the discussion of evaluating the poverty impact of projects: (a) approaches to measuring the impact of structural adjustment programs, and (b) theory-based evaluations. Both incorporate many of the methodologies discussed above, but each uses a different approach.

**Evaluating Structural Adjustment Programs.** There has been substantial debate on the impact of structural adjustment programs on the poor. Much of the evidence used to support this debate is, however, based on deficient assumptions and methods. As with other projects, the policy changes under structural adjustment projects must be (a) compared with relevant counterfactuals that would respond to the same macroeconomic constraints, and (b) analyzed in the context of the local economic structure and based on empirical information from household surveys. This, however, is very difficult for three reasons. First, policy changes may have economy-wide impact, making it impossible to find comparison groups that are unaffected. Second, because of exogenous factors, lags, feedbacks, and substitutions, any changes in the well-being of the poor must be interpreted with extreme caution. And third, it is difficult to predict what would have happened if adjustment had not taken place—what alternative policies a government might have pursued and what the resulting impact would have been on the poor.

In the literature, several approaches have been used, each with its own shortcomings. The techniques are in many cases similar to those described in Box 1.2, though, as shown in Box 1.3, estimating the counterfactual requires vast assumptions that may substantially affect the validity of the results. This is most viably handled by isolating specific

policy changes that would affect the population, such as exchange rate policies, trade policies, reductions in public expenditures, and reductions in public sector employment. Yet even with this approach it can be difficult to isolate the impact of specific policies. For examples, see Killick (1995), Poppele, Summarto, and Pritchett (1999), Bourguignon, de Melo, and Suwa (1991), and Sahn, Dorosh, and Younger (1996).

### **Box 1.3 Summary of Methods Used to Evaluate Adjustment Policies**

#### *Approaches with No Counterfactual*

- Qualitative studies that assess conditions of the population (often identifying vulnerable subgroups) before, during, and after adjustment policies are implemented through focus groups, interviews, and other qualitative techniques.
- “Before and After,” which compares the performance of key variables during and after a program with those prior to the program. The approach uses statistical methods to evaluate whether there is a significant change in some essential variables over time. This approach often gives biased results because it assumes that had it not been for the program, the performance indicators would have taken their pre-crisis-period values.

#### *Approaches that Generate a Counterfactual Using Multiple Assumptions*

- Computable general equilibrium models (CGEs) that attempt to contrast outcomes in treatment and comparison groups through simulations. These models seek to trace the operation of the real economy and are generally based on detailed social accounting matrices collected from data on national accounts, household expenditure surveys, and other survey data. CGE models do produce outcomes for the counterfactual, though the strength of the model is entirely dependent on the validity of the assumptions. This can be problematic as databases are often incomplete and many of the parameters have not been estimated by formal econometric methods. CGE models are also very time consuming, cumbersome, and expensive to generate.

*(Box continues on the following page.)*

**Box 1.3** *(continued)*

- With and without comparisons, which compare the behavior in key variables in a sample of program countries with their behavior in nonprogram countries (a comparison group). This is an approach to the counterfactual question, using the experiences of the comparison group as a proxy for what would otherwise have happened in the program countries. It is, however, quite difficult to achieve a true comparison group. The method assumes that only the adoption of an adjustment program distinguishes a program country from the comparison group and that the external environment affects both groups the same.
- Statistical controls consisting of regressions that control for the differences in initial conditions and policies undertaken in program and nonprogram countries. The approach identifies the differences between program and nonprogram countries in the pre-program period and then controls these differences statistically to identify the isolated impacts of the programs in the postreform performance.

**Theory-Based Evaluation.** The premise of theory-based evaluations is that programs and projects are based on explicit or implicit theory about how and why a program will work. The evaluation would then be based on assessing each theory and assumptions about a program during implementation rather than at a midpoint or after the project has been completed. In designing the evaluation, the underlying theory is presented as many microsteps, with the methods then constructed for data collection and analysis to track the unfolding of assumptions. If events do not work out as expected, the evaluation can say with a certain confidence where, why, and how the breakdown occurred.

The approach puts emphasis on the responses of people to program activities. Theories direct the evaluator's attention to likely types of near-term and longer-term effects. Among the advantages are, first, that the evaluation provides early indications of program effectiveness during project implementation. If there are breakdowns during implementation, it is possible to fix them along the way. Second, the approach helps to explain how and why effects occurred. If events work out as expected, the evaluation can say with a certain confidence how the effects were generated. By following the sequence of stages, it is possible to track the microsteps that led from program inputs through to outcomes.

The shortcomings of the approach are similar to many of the other methodologies. In particular, (a) identifying assumptions and theories can be inherently complex; (b) evaluators may have problems in measuring each step unless the right instruments and data are available, (c) problems may be encountered in testing the effort because theory statements may be too general and loosely constructed to allow for clear-cut testing, and (d) there may be problems of interpretation that make it difficult to generalize from results (see Weiss 1998).

An example of theory-based technique is being piloted by the Operations and Evaluation Department of the World Bank to evaluate the impact of social investment funds on community-level decisionmaking processes, traditional power structures and relationships, and community capacity, trust, and well-being. This will be based on the theory that priority groups can effectively implement a project and operate and maintain the investment created by the project. A set of main assumptions and subassumptions has been set out and will be tested using existing household survey data, as well as a specially designed survey instrument for a smaller sample, and focus groups and other PRA techniques. The information from each of these data sources will be triangulated in the analysis.

### **Cost-Benefit or Cost-Effectiveness Analysis**

While this type of analysis is not strictly concerned with measuring impact, it enables policymakers to measure program efficiency by comparing alternative interventions on the basis of the cost of producing a given output. It can greatly enhance the policy implications of the impact evaluation and therefore should also be included in the design of any impact evaluation. (For a more complete discussion of cost-benefit and cost-effectiveness analysis, see *Handbook on Economic Analysis of Investment Operations*, World Bank 1996.)

Cost-benefit analysis attempts to measure the economic efficiency of program costs versus program benefits, in monetary terms. For many projects, especially in the social sectors, it is not possible to measure all the benefits in monetary terms. For example, the benefits of a program to provide school inputs (textbooks, classroom furniture, preschool programs) would be increased learning. Instead of measuring monetary outcomes, learning achievement scores could be used to quantify the benefits. This would require cost-effectiveness analysis. The concepts for both types of analysis are the same.

The main steps of cost-benefit and cost-effectiveness analysis are to identify all project costs and benefits and then compute a cost-to-effectiveness ratio. In calculating costs, the value of the intervention itself

should be included, as well as all other costs, such as administration, delivery, investment costs (discounted to the net present value), the monetary value of freely provided goods or services, social costs such as environmental deterioration, and health hazards. Benefits can be monetary, such as gain in income, or the number of units delivered, test scores, or health improvements. When benefits cannot be quantified, it is possible to use subjective indicators such as ranking or weighting systems. This approach, however, can be tricky in interpreting subjective scores.

Once the costs and benefits have been determined, the cost-effectiveness ratio (R) is then  $R = \text{cost}/\text{unit}$  (or benefit). This ratio can then be compared across interventions to measure efficiency. In theory, this technique is quite straightforward. In practice, however, there are many caveats involved in identifying and quantifying the costs and benefits. It is important to ensure that appropriate indicators are selected, that the methodologies and economic assumptions used are consistent across ratios, and that the ratios are indeed comparable. And as with other techniques used in impact analysis, measuring cost-effectiveness can be best carried out when included in the evaluation design from the earliest stages. This allows for the collection of the necessary cost and benefit information and ensuring consistency.

### Choosing a Methodology

Given the variation in project types, evaluation questions, data availability, cost, time constraints, and country circumstances, each impact evaluation study will be different and will require some combination of appropriate methodologies, both quantitative and qualitative. The evaluator must carefully explore the methodological options in designing the study, with the aim of producing the most robust results possible. Among quantitative methods, experimental designs are considered the optimal approach and matched comparisons a second-best alternative. Other techniques, however, can also produce reliable results, particularly with a good evaluation design and high-quality data.

The evidence from the “best-practice” evaluations reviewed for this handbook highlights that the choice of impact evaluation methodologies is not mutually exclusive. Indeed, stronger evaluations often combine methods to ensure robustness and to provide for contingencies in implementation. Joining a “with and without” approach with a “before and after” approach that uses baseline and follow-up data is one combination strongly recommended from a methodological perspective (Subbarao and others 1999). Having baseline data available will allow evaluators to verify the integrity of treatment and comparison groups, assess targeting, and prepare for a robust impact evaluation. This is true even for ran-

domized control designs. Although randomization ensures equivalent treatment and comparison groups at the time of randomization, this feature should not influence evaluators into thinking that they do not need baseline data. Indeed, baseline data may be crucial to reconstructing why certain events took place and controlling for these events in the impact assessment.

Incorporating cost-benefit or cost-effectiveness analysis is also strongly recommended. This methodology can enable policymakers to compare alternative interventions on the basis of the cost of producing a given output. This is particularly important in the developing-country context in which resources are extremely limited.

Finally, combining quantitative and qualitative methods is the ideal because it will provide the quantifiable impact of a project as well as an explanation of the processes and interventions that yielded these outcomes. Although each impact evaluation will have unique characteristics requiring different methodological approaches, a few general qualities of a best-practice impact evaluation include:

- An estimate of the counterfactual has been made by (a) using random assignment to create a control group (experimental design), and (b) appropriately and carefully using other methods such as matching to create a comparison group (quasi-experimental design).
- To control for pre- and postprogram differences in participants, and to establish program impacts, there are relevant data collected at baseline and follow-up (including sufficient time frame to allow for program impacts).
- The treatment and comparison groups are of sufficient sizes to establish statistical inferences with minimal attrition.
- Cost-benefit or cost-effectiveness analysis is included to measure project efficiency.
- Qualitative techniques are incorporated to allow for the triangulation of findings.