1. Basic Issues of Evaluation

Summary

Several approaches can be used to evaluate programs. *Monitoring* tracks key indicators of progress over the course of a program as a basis on which to evaluate outcomes of the intervention. *Operational evaluation* examines how effectively programs were implemented and whether there are gaps between planned and realized outcomes. *Impact evaluation* studies whether the changes in well-being are indeed due to the program intervention and not to other factors.

These evaluation approaches can be conducted using quantitative methods (that is, survey data collection or simulations) before or after a program is introduced. *Ex ante evaluation* predicts program impacts using data before the program intervention, whereas *ex post evaluation* examines outcomes after programs have been implemented. Reflexive comparisons are a type of ex post evaluation; they examine program impacts through the difference in participant outcomes before and after program implementation (or across participants and nonparticipants). Subsequent chapters in this handbook provide several examples of these comparisons.

The main challenge across different types of impact evaluation is to find a good counterfactual—namely, the situation a participating subject would have experienced had he or she not been exposed to the program. Variants of impact evaluation discussed in the following chapters include randomized evaluations, propensity score matching, double-difference methods, use of instrumental variables, and regression discontinuity and pipeline approaches. Each of these methods involves a different set of assumptions in accounting for potential selection bias in participation that might affect construction of program treatment effects.

Learning Objectives

After completing this chapter, the reader will be able to discuss and understand

- Different approaches to program evaluation
- Differences between quantitative and qualitative approaches to evaluation, as well as ex ante versus ex post approaches
- Ways selection bias in participation can confound the treatment effect
- Different methodologies in impact evaluation, including randomization, propensity score matching, double differences, instrumental variable methods, and regression discontinuity and pipeline approaches

Introduction: Monitoring versus Evaluation

Setting goals, indicators, and targets for programs is at the heart of a monitoring system. The resulting information and data can be used to evaluate the performance of program interventions. For example, the World Bank Independent Evaluation Group weighs the progress of the World Bank–International Monetary Fund Poverty Reduction Strategy (PRS) initiative against its objectives through monitoring; many countries have also been developing monitoring systems to track implementation of the PRS initiative and its impact on poverty. By comparing program outcomes with specific targets, monitoring can help improve policy design and implementation, as well as promote accountability and dialogue among policy makers and stakeholders.

In contrast, evaluation is a systematic and objective assessment of the results achieved by the program. In other words, evaluation seeks to prove that changes in targets are due only to the specific policies undertaken. Monitoring and evaluation together have been referred to as *M&E*. For example, M&E can include *process evaluation*, which examines how programs operate and focuses on problems of service delivery; *cost-benefit analysis*, which compares program costs against the benefits they deliver; and *impact evaluations*, which quantify the effects of programs on individuals, households, and communities. All of these aspects are part of a good M&E system and are usually carried out by the implementing agency.

Monitoring

The challenges in monitoring progress of an intervention are to

- Identify the *goals* that the program or strategy is designed to achieve, such as reducing poverty or improving schooling enrollment of girls. For example, the Millennium Development Goals initiative sets eight broad goals across themes such as hunger, gender inequalities, schooling, and poverty to monitor the performance of countries and donors in achieving outcomes in those areas.
- Identify key *indicators* that can be used to monitor progress against these goals. In the context of poverty, for example, an indicator could be the proportion of individuals consuming fewer than 2,100 calories per day or the proportion of households living on less than a dollar a day.
- Set *targets*, which quantify the level of the indicators that are to be achieved by a given date. For instance, a target might be to halve the number of households living on less than a dollar a day by 2015.
- Establish a *monitoring system* to track progress toward achieving specific targets and to inform policy makers. Such a system will encourage better management of and accountability for projects and programs.

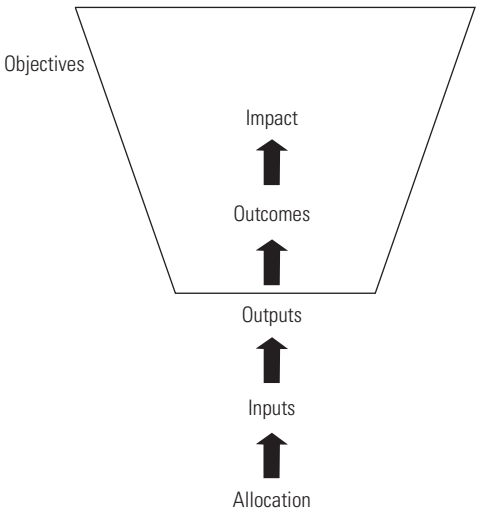
Setting Up Indicators within an M&E Framework

Indicators are typically classified into two major groups. First, *final indicators* measure the outcomes of poverty reduction programs (such as higher consumption per capita) and the impact on dimensions of well-being (such as reduction of consumption poverty). Second, *intermediate indicators* measure inputs into a program (such as a conditional cash-transfer or wage subsidy scheme) and the outputs of the program (such as roads built, unemployed men, and women hired). Target indicators can be represented in four clusters, as presented in figure 2.1. This so-called logic framework spells out the inputs, outputs, outcomes, and impacts in the M&E system. Impact evaluation, which is the focus of this handbook, spans the latter stages of the M&E framework.

Viewed in this framework, monitoring covers both implementation and performance (or results-based) monitoring. Intermediate indicators typically vary more quickly than final indicators, respond more rapidly to public interventions, and can be measured more easily and in a more timely fashion. Selecting indicators for monitoring against goals and targets can be subject to resource constraints facing the project management authority. However, it is advisable to select only a few indicators that can be monitored properly rather than a large number of indicators that cannot be measured well.

One example of a monitoring system comes from PROGRESA (Programa de Educación, Salud y Alimentación, or Education, Health, and Nutrition Program) in Mexico (discussed in more detail in box 2.1). PROGRESA (now called Oportunidades) is one of the largest randomized interventions implemented by a single country. Its aim was

Figure 2.1 Monitoring and Evaluation Framework



Source: Authors' representation.

BOX 2.1 Case Study: PROGRESA (Oportunidades) in Mexico

Monitoring was a key component of the randomized program PROGRESA (now called Oportunidades) in Mexico, to ensure that the cash transfers were directed accurately. Program officials foresaw several potential risks in implementing the program. These risks included the ability to ensure that transfers were targeted accurately; the limited flexibility of funds, which targeted households instead of communities, as well as the nondiscretionary nature of the transfers; and potential intrahousehold conflicts that might result because transfers were made only to women.

Effective monitoring therefore required that the main objectives and intermediate indicators be specified clearly. Oportunidades has an institutional information system for the program's operation, known as SIIOP (Sistema Integral de Información para la Operación de Oportunidades, or Complete Information System for the Operation of Oportunidades), as well as an audit system that checks for irregularities at different stages of program implementation. These systems involved several studies and surveys to assess how the program's objectives of improving health, schooling, and nutrition should be evaluated. For example, to determine schooling objectives, the systems ran diagnostic studies on potentially targeted areas to see how large the educational grants should be, what eligibility requirements should be established in terms of grades and gender, and how many secondary schools were available at the local, municipal, and federal levels. For health and nutrition outcomes, documenting behavioral variation in household hygiene and preparation of foods across rural and urban areas helped to determine food supplement formulas best suited for targeted samples.

These systems also evaluated the program's ability to achieve its objectives through a design that included randomized checks of delivery points (because the provision of food supplements, for example, could vary substantially between providers and government authorities); training and regular communication with stakeholders in the program; structuring of fieldwork resources and requirements to enhance productivity in survey administration; and coordinated announcements of families that would be beneficiaries.

The approaches used to address these issues included detailed survey instruments to monitor outcomes, in partnership with local and central government authorities. These instruments helped to assess the impact of the program on households and gave program officials a sense of how effectively the program was being implemented. The surveys included, for example, a pilot study to better understand the needs of households in targeted communities and to help guide program design. Formal surveys were also conducted of participants and nonparticipants over the course of the program, as well as of local leaders and staff members from schools and health centers across the localities. Administrative data on payments to households were also collected.

to target a number of health and educational outcomes including malnutrition, high infant mortality, high fertility, and school attendance. The program, which targeted rural and marginal urban areas, was started in mid-1997 following the macroeconomic crisis of 1994 and 1995. By 2004, around 5 million families were covered, with a budget of about US$2.5 billion, or 0.3 percent of Mexico's gross domestic product.

The main thrust of Oportunidades was to provide conditional cash transfers to households (specifically mothers), contingent on their children attending school

and visiting health centers regularly. Financial support was also provided directly to these institutions. The average benefit received by participating households was about 20 percent of the value of their consumption expenditure before the program, with roughly equal weights on the health and schooling requirements. Partial participation was possible; that is, with respect to the school subsidy initiative, a household could receive a partial benefit if it sent only a proportion of its children to school.

Results-Based Monitoring

The actual execution of a monitoring system is often referred to as *results-based monitoring*. Kusek and Rist (2004) outline 10 steps to results-based monitoring as part of an M&E framework.

First, a readiness assessment should be conducted. The assessment involves understanding the needs and characteristics of the area or region to be targeted, as well as the key players (for example, the national or local government and donors) that will be responsible for program implementation. How the effort will respond to negative pressures and information generated from the M&E process is also important.

Second, as previously mentioned, program evaluators should agree on specific outcomes to monitor and evaluate, as well as key performance indicators to monitor outcomes. Doing so involves collaboration with recipient governments and communities to arrive at a mutually agreed set of goals and objectives for the program. Third, evaluators need to decide how trends in these outcomes will be measured. For example, if children's schooling were an important outcome for a program, would schooling achievement be measured by the proportion of children enrolled in school, test scores, school attendance, or another metric? Qualitative and quantitative assessments can be conducted to address this issue, as will be discussed later in this chapter. The costs of measurement will also guide this process.

Fourth, the instruments to collect information need to be determined. Baseline or preprogram data can be very helpful in assessing the program's impact, either by using the data to predict outcomes that might result from the program (as in ex ante evaluations) or by making before-and-after comparisons (also called *reflexive comparisons*). Program managers can also engage in frequent discussions with staff members and targeted communities.

Fifth, targets need to be established; these targets can also be used to monitor results. This effort includes setting periodic targets over time (for example, annually or every two years). Considering the duration of the likely effects of the program, as well as other factors that might affect program implementation (such as political considerations), is also important. Monitoring these targets, in particular, embodies the sixth step in this results-based framework and involves the collection of good-quality data.

The seventh step relates to the timing of monitoring, recognizing that from a management perspective the timing and organization of evaluations also drive the extent to which evaluations can help guide policy. If actual indicators are found to be diverging rapidly from initial goals, for example, evaluations conducted around that time can help program managers decide quickly whether program implementation or other related factors need to be adjusted.

The eighth step involves careful consideration of the means of reporting, including the audience to whom the results will be presented. The ninth step involves using the results to create avenues for feedback (such as input from independent agencies, local authorities, and targeted and nontargeted communities). Such feedback can help evaluators learn from and update program rules and procedures to improve outcomes.

Finally, successful results-based M&E involves sustaining the M&E system within the organization (the 10th step). Effective M&E systems will endure and are based on, among other things, continued demand (a function of incentives to continue the program, as well as the value for credible information); transparency and accountability in evaluation procedures; effective management of budgets; and well-defined responsibilities among program staff members.

One example of results-based monitoring comes from an ongoing study of microhydropower projects in Nepal under the Rural Electrification Development Program (REDP) administered by the Alternative Energy Promotion Center (AEPC). AEPC is a government institute under the Ministry of Environment, Science, and Technology. The microhydropower projects began in 1996 across five districts with funding from the United Nations Development Programme; the World Bank joined the REDP during the second phase in 2003. The program is currently in its third phase and has expanded to 25 more districts. As of December 2008, there were about 235 microhydropower installations (3.6 megawatt capacity) and 30,000 beneficiary households. Box 2.2 describes the monitoring framework in greater detail.

Challenges in Setting Up a Monitoring System

Primary challenges to effective monitoring include potential variation in program implementation because of shortfalls in capacity among program officials, as well as ambiguity in the ultimate indicators to be assessed. For the microhydropower projects in Nepal, for example, some challenges faced by REDP officials in carrying out the M&E framework included the following:

- Key performance indicators were not well defined and hence not captured comprehensively.
- Limited human resources were available for collecting and recording information.

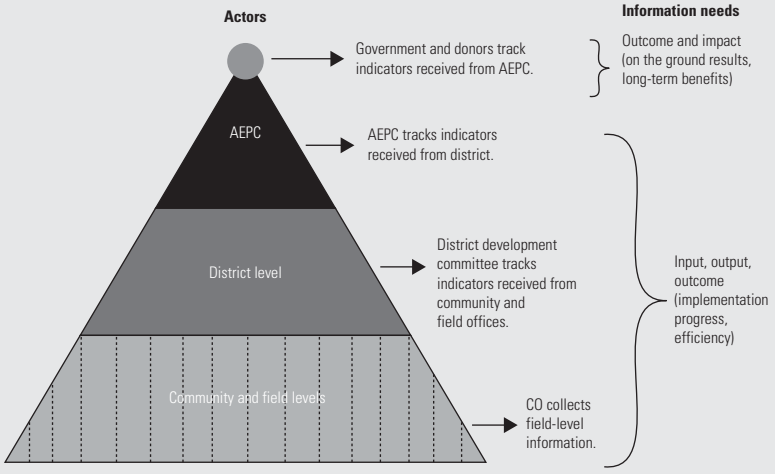
BOX 2.2 **Case Study: Assessing the Social Impact of Rural Energy Services in Nepal**

REDP microhydropower projects include six community development principles: organizational development, skill enhancement, capital formation, technology promotion, empowerment of vulnerable communities, and environment management. Implementation of the REDP microhydropower projects in Nepal begins with community mobilization. Community organizations (COs) are first formed by individual beneficiaries at the local level. Two or more COs form legal entities called *functional groups*. A management committee, represented by all COs, makes decision about electricity distribution, tariffs, operation, management, and maintenance of microhydropower projects.

A study on the social impact of rural energy services in Nepal has recently been funded by Energy Sector Management Assistance Program and is managed by the South Asia Energy Department of the World Bank. In implementing the M&E framework for the microhydropower projects, this study seeks to (a) improve management for the program (better planning and reporting); (b) track progress or systematic measurement of benefits; (c) ensure accountability and results on investments from stakeholders such as the government of Nepal, as well as from donors; and (d) provide opportunities for updating how the program is implemented on the basis of continual feedback on how outcomes overlap with key performance indicators.

Box figure 2.A describes the initial monitoring framework set up to disseminate information about how inputs, outputs, and outcomes were measured and allocated. Information is collected at each of the community, district, and head office (AEPC) levels. Community mobilizers relay field-level information to coordinators at the district level, where additional information is also collected. At the district level, information is verified and sent to AEPC, where reports are prepared and then sent to various stakeholders. Stakeholders, in particular, can include the government of Nepal, as well as donors.

BOX Figure 2.A Levels of Information Collection and Aggregation



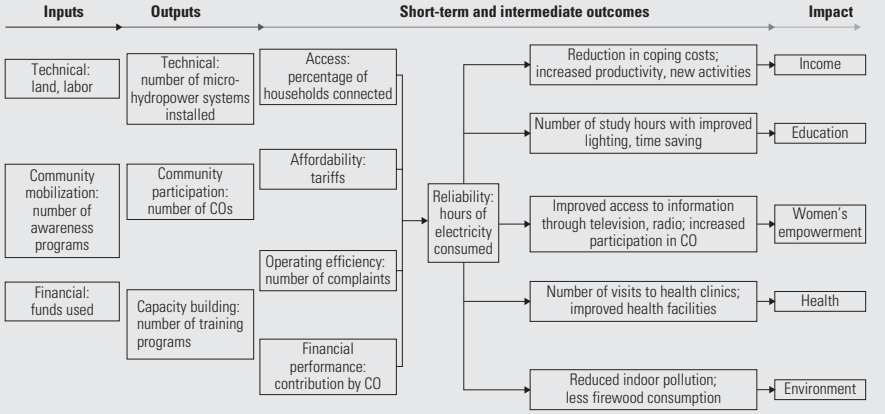
Source: Banerjee, Singh, and Samad 2009.

(*Box continues on the following page.*)

BOX 2.2 Case Study: Assessing the Social Impact of Rural Energy Services in Nepal (continued)

Box figure 2.B outlines how key performance indicators have been set up for the projects. Starting with inputs such as human and physical capital, outputs such as training programs and implementation of systems are generated. Short-term and intermediate outcomes are outlined, including improved productivity and efficiency of household labor stemming from increased access to electricity, leading to broader potential impacts in health, education, women's welfare, and the environment.

BOX Figure 2.B Building up of Key Performance Indicators: Project Stage Details



Source: Banerjee, Singh, and Samad 2009.

- M&E personnel had limited skills and capacity, and their roles and responsibilities were not well defined at the field and head office levels.
- AEPC lacked sophisticated tools and software to analyze collected information.

Weaknesses in these areas have to be addressed through different approaches. Performance indicators, for example, can be defined more precisely by (a) better understanding the inputs and outputs at the project stage, (b) specifying the level and unit of measurement for indicators, (c) frequently collecting community- and beneficiary-level data to provide periodic updates on how intermediate outcomes are evolving and whether indicators need to be revised, and (d) clearly identifying the people and entities responsible for monitoring. For data collection in particular, the survey timing (from a preproject baseline, for example, up to the current period); frequency (monthly or semiannually, for example); instruments (such as interviews or bills); and level of collection (individual, household, community, or a broader administrative unit such as district) need to defined and set up explicitly within the M&E framework. Providing the staff with training and tools for data collection and analysis, as well as

for data verification at different levels of the monitoring structure (see box figure 2.A in box 2.2 for an example), is also crucial.

Policy makers might also need to establish how microlevel program impacts (at the community or regional level) would be affected by country-level trends such as increased trade, inflation, and other macroeconomic policies. A related issue is heterogeneity in program impacts across a targeted group. The effects of a program, for example, may vary over its expected lifetime. Relevant inputs affecting outcomes may also change over this horizon; thus, monitoring long-term as well as short-term outcomes may be of interest to policy makers. Also, although program outcomes are often distinguished simply across targeted and nontargeted areas, monitoring variation in the program's implementation (measures of quality, for example) can be extremely useful in understanding the program's effects. With all of these concerns, careful monitoring of targeted and nontargeted areas (whether at the regional, household, or individual level) will help greatly in measuring program effects. Presenting an example from Indonesia, box 2.3 describes some techniques used to address M&E challenges.

BOX 2.3 Case Study: The Indonesian Kecamatan Development Project

The Kecamatan Development Program (KDP) in Indonesia, a US$1.3 billion program run by the Community Development Office of the Ministry of Home Affairs, aims to alleviate poverty by strengthening local government and community institutions as well as by improving local governance. The program began in 1998 after the financial crisis that plagued the region, and it works with villages to define their local development needs. Projects were focused on credit and infrastructural expansion. This program was not ultimately allocated randomly.

A portion of the KDP funds were set aside for monitoring activities. Such activities included, for example, training and capacity development proposed by the communities and local project monitoring groups. Technical support was also provided by consultants, who were assigned to sets of villages. They ranged from technical consultants with engineering backgrounds to empowerment consultants to support communication within villages.

Governments and nongovernmental organizations assisted in monitoring as well, and villages were encouraged to engage in self-monitoring through piloted village-district parliament councils and cross-village visits. Contracts with private banks to provide village-level banking services were also considered. As part of this endeavor, financial supervision and training were provided to communities, and a simple financial handbook and checklist were developed for use in the field as part of the monitoring initiative. District-level procurement reforms were also introduced to help villages and local areas buy technical services for projects too large to be handled by village management.

Project monitoring combined quantitative and qualitative approaches. On the quantitative side, representative sample surveys helped assess the poverty impact of the project across different areas. On the qualitative side, consultants prepared case studies to highlight lessons learned

(*Box continues on the following page.*)

BOX 2.3 Case Study: The Indonesian Kecamatan Development Project (continued)

from the program, as well as to continually evaluate KDP's progress. Some issues from these case studies include the relative participation of women and the extreme poor, conflict resolution, and the role of village facilitators in disseminating information and knowledge.

Given the wide scope of the program, some areas of improvement have been suggested for KDP monitoring. Discussions or sessions conducted with all consultants at the end of each evaluation cycle can encourage feedback and dialogue over the course of the program, for example. Focus groups of consultants from different backgrounds (women, for example) might also elicit different perspectives valuable to targeting a diverse population. Suggestions have also been made to develop themes around these meetings, such as technical issues, transparency and governance, and infrastructure. Consultants were also often found to not regularly report problems they found in the field, often fearing that their own performance would be criticized. Incentives to encourage consultants to accurately report developments in their areas have also been discussed as part of needed improvements in monitoring.

Operational Evaluation

An operational evaluation seeks to understand whether implementation of a program unfolded as planned. Specifically, operational evaluation is a retrospective assessment based on initial project objectives, indicators, and targets from the M&E framework. Operation evaluation can be based on interviews with program beneficiaries and with officials responsible for implementation. The aim is to compare what was planned with what was actually delivered, to determine whether there are gaps between planned and realized outputs, and to identify the lessons to be learned for future project design and implementation.

Challenges in Operational Evaluation

Because operational evaluation relates to how programs are ultimately implemented, designing appropriate measures of implementation quality is very important. This effort includes monitoring how project money was ultimately spent or allocated across sectors (as compared to what was targeted), as well as potential spillovers of the program into nontargeted areas. Collecting precise data on these factors can be difficult, but as described in subsequent chapters, it is essential in determining potential biases in measuring program impacts. Box 2.4, which examines FONCODES (Fondo de Cooperación para el Desarrollo Social, or Cooperation Fund for Social Development), a poverty alleviation program in Peru, shows how operational evaluation also often involves direct supervision of different stages of program implementation. FONCODES has both educational and nutritional objectives. The nutritional

BOX 2.4 Case Study: Monitoring the Nutritional Objectives of the FONCODES Project in Peru

Within the FONCODES nutrition initiative in Peru, a number of approaches were taken to ensure the quality of the nutritional supplement and efficient implementation of the program. At the program level, the quality of the food was evaluated periodically through independent audits of samples of communities. This work included obtaining and analyzing random samples of food prepared by targeted households. Every two months, project officials would randomly visit distribution points to monitor the quality of distribution, including storage. These visits also provided an opportunity to verify the number of beneficiaries and to underscore the importance of the program to local communities.

Home visits were also used to evaluate beneficiaries' knowledge of the project and their preparation of food. For example, mothers (who were primarily responsible for cooking) were asked to show the product in its bag, to describe how it was stored, and to detail how much had been consumed since the last distribution. They were also invited to prepare a ration so that the process could be observed, or samples of leftovers were taken for subsequent analysis.

The outcomes from these visits were documented regularly. Regular surveys also documented the outcomes. These data allowed program officials to understand how the project was unfolding and whether any strategies needed to be adjusted or reinforced to ensure program quality. At the economywide level, attempts were made at building incentives within the agrifood industry to ensure sustainable positioning of the supplement in the market; companies were selected from a public bidding process to distribute the product.

The operational efforts aimed at ultimately reducing poverty in these areas, however, did vary from resulting impact estimates. FONCODES was not allocated randomly, for example, and Schady (1999) found that the flexibility of allocation of funds within FONCODES, as well as in the timing and constitution of expenditures, made the program very vulnerable to political interference. Paxson and Schady (2002) also used district-level data on expenditures from the schooling component of the program to find that though the program did reach the poorest districts, it did not necessarily reach the poorest households in those districts. They did find, however, that the program increased school attendance, particularly that of younger children. Successful program implementation therefore requires harnessing efforts over all of the program's objectives, including effective enforcement of program targeting.

component involves distributing precooked, high-nutrition food, which is currently consumed by about 50,000 children in the country. Given the scale of the food distribution initiative, a number of steps were taken to ensure that intermediate inputs and outcomes could be monitored effectively.

Operational Evaluation versus Impact Evaluation

The rationale of a program in drawing public resources is to improve a selected outcome over what it would have been without the program. An evaluator's main problem is to measure the impact or effects of an intervention so that policy makers can decide

whether the program intervention is worth supporting and whether the program should be continued, expanded, or disbanded.

Operational evaluation relates to ensuring effective implementation of a program in accordance with the program's initial objectives. *Impact evaluation* is an effort to understand whether the changes in well-being are indeed due to project or program intervention. Specifically, impact evaluation tries to determine whether it is possible to identify the program effect and to what extent the measured effect can be attributed to the program and not to some other causes. As suggested in figure 2.1, impact evaluation focuses on the latter stages of the log frame of M&E, which focuses on outcomes and impacts.

Operational and impact evaluation are complementary rather than substitutes, however. An operational evaluation should be part of normal procedure within the implementing agency. But the template used for an operational evaluation can be very useful for more rigorous impact assessment. One really needs to know the context within which the data was generated and where policy effort was directed. Also, the information generated through project implementation offices, which is essential to an operational evaluation, is also necessary for interpretation of impact results.

However, although operational evaluation and the general practice of M&E are integral parts of project implementation, impact evaluation is not imperative for each and every project. Impact evaluation is time and resource intensive and should therefore be applied selectively. Policy makers may decide whether to carry out an impact evaluation on the basis of the following criteria:

- The program intervention is innovative and of strategic importance.
- The impact evaluation exercise contributes to the knowledge gap of what works and what does not. (Data availability and quality are fundamental requirements for this exercise.)

Mexico's Oportunidades program is an example in which the government initiated a rigorous impact evaluation at the pilot phase to determine whether to ultimately roll out the program to cover the entire country.

Quantitative versus Qualitative Impact Assessments

Governments, donors, and other practitioners in the development community are keen to determine the effectiveness of programs with far-reaching goals such as lowering poverty or increasing employment. These policy quests are often possible only through impact evaluations based on hard evidence from survey data or through related quantitative approaches.

This handbook focuses on quantitative impact methods rather than on qualitative impact assessments. Qualitative information such as understanding the local sociocultural and institutional context, as well as program and participant details, is,

however, essential to a sound quantitative assessment. For example, qualitative information can help identify mechanisms through which programs might be having an impact; such surveys can also identify local policy makers or individuals who would be important in determining the course of how programs are implemented, thereby aiding operational evaluation. But a qualitative assessment on its own cannot assess outcomes against relevant alternatives or *counterfactual* outcomes. That is, it cannot really indicate what might happen in the absence of the program. As discussed in the following chapters, quantitative analysis is also important in addressing potential statistical bias in program impacts. A mixture of qualitative and quantitative methods (a *mixed-methods approach*) might therefore be useful in gaining a comprehensive view of the program's effectiveness.

Box 2.5 describes a mixed-methods approach to examining outcomes from the Jamaica Social Investment Fund (JSIF). As with the Kecamatan Development Program in Indonesia (see box 2.3), JSIF involved community-driven initiatives, with communities making cash or in-kind contributions to project development costs (such as construction). The qualitative and quantitative evaluation setups both involved comparisons of outcomes across matched treated and untreated pairs of communities, but with different approaches to matching communities participating and not participating in JSIF.

BOX 2.5 Case Study: Mixed Methods in Quantitative and Qualitative Approaches

Rao and Ibáñez (2005) applied quantitative and qualitative survey instruments to study the impact of Jamaica Social Investment Fund. Program evaluators conducted semistructured in-depth qualitative interviews with JSIF project coordinators, local government and community leaders, and members of the JSIF committee that helped implement the project in each community. This information revealed important details about social norms, motivated by historical and cultural influences that guided communities' decision making and therefore the way the program ultimately played out in targeted areas. These interviews also helped in matching communities, because focus groups were asked to identify nearby communities that were most similar to them.

Qualitative interviews were not conducted randomly, however. As a result, the qualitative interviews could have involved people who were more likely to participate in the program, thereby leading to a bias in understanding the program impact. A quantitative component to the study was therefore also included. Specifically, in the quantitative component, 500 households (and, in turn, nearly 700 individuals) were surveyed, split equally across communities participating and not participating in the fund. Questionnaires covered a range of variables, including socioeconomic characteristics, details of participation in the fund and other local programs, perceived priorities for community development, and social networks, as well as ways a number of their outcomes had changed relative to five years ago (before JSIF began). Propensity score matching, discussed in

(*Box continues on the following page.*)

BOX 2.5 Case Study: Mixed Methods in Quantitative and Qualitative Approaches (continued)

greater detail in chapter 4, was used to compare outcomes for participating and nonparticipating households. Matching was conducted on the basis of a poverty score calculated from national census data. Separate fieldwork was also conducted to draw out additional, unmeasured community characteristics on which to conduct the match; this information included data on local geography, labor markets, and the presence of other community organizations. Matching in this way allowed better comparison of targeted and nontargeted areas, thereby avoiding bias in the treatment impacts based on significant observed and unobserved differences across these groups.

The qualitative data therefore revealed valuable information on the institutional context and norms guiding behavior in the sample, whereas the quantitative data detailed trends in poverty reduction and other related indicators. Overall, when comparing program estimates from the qualitative models (as measured by the difference-in-differences cross-tabulations of survey responses across JSIF and non-JSIF matched pairs—see chapter 5 for a discussion of difference-in-differences methods) with the quantitative impact estimated from nearest-neighbor matching, Rao and Ibáñez found the pattern of effects to be similar. Such effects included an increased level of trust and an improved ability of people from different backgrounds to work together. For the latter outcome, for example, about 21 percent of the JSIF sample said it was "very difficult" or "difficult" for people of different backgrounds to work together in the qualitative module, compared with about 32 percent of the non-JSIF sample. Similarly, the nearest-neighbor estimates revealed a significant positive mean benefit for this outcome to JSIF areas (about 0.33).

The quantitative impacts were also broken down by household socioeconomic characteristics. They tended to show, however, that JSIF may have created better outcomes in terms of increased collective action for wealthier and better-educated participants; qualitative evidence also revealed that these groups tended to dominate the decision-making process.

Quantitative Impact Assessment: Ex Post versus Ex Ante Impact Evaluations

There are two types of quantitative impact evaluations: ex post and ex ante. An ex ante impact evaluation attempts to measure the intended impacts of future programs and policies, given a potentially targeted area's current situation, and may involve simulations based on assumptions about how the economy works (see, for example, Bourguignon and Ferreira 2003; Todd and Wolpin 2006). Many times, ex ante evaluations are based on structural models of the economic environment facing potential participants (see chapter 9 for more discussion on structural modeling). The underlying assumptions of structural models, for example, involve identifying the main economic agents in the development of the program (individuals, communities, local or national governments), as well as the links between the agents and the different markets in determining outcomes from the program. These models predict program impacts.

Ex post evaluations, in contrast, measure actual impacts accrued by the beneficiaries that are attributable to program intervention. One form of this type of evaluation is the treatment effects model (Heckman and Vytlacil, 2005). Ex post evaluations have immediate benefits and reflect reality. These evaluations, however, sometimes miss the mechanisms underlying the program's impact on the population, which structural models aim to capture and which can be very important in understanding program effectiveness (particularly in future settings). Ex post evaluations can also be much more costly than ex ante evaluations because they require collecting data on actual outcomes for participant and nonparticipant groups, as well as on other accompanying social and economic factors that may have determined the course of the intervention. An added cost in the ex post setting is the failure of the intervention, which might have been predicted through ex ante analysis.

One approach is to combine both analyses and compare ex post estimates with ex ante predictions (see Ravallion 2008). This approach can help explain how program benefits emerge, especially if the program is being conducted in different phases and has the flexibility to be refined from added knowledge gained from the comparison. Box 2.6 provides an example of this approach, using a study by Todd and Wolpin (2006) of a school subsidy initiative under PROGRESA.

The case studies discussed in the following chapters primarily focus on ex post evaluations. However, an ex post impact exercise is easier to carry out if the researchers have an ex ante design of impact evaluation. That is, one can plan a design for

BOX 2.6 Case Study: An Example of an Ex Ante Evaluation

Todd and Wolpin (2006) applied an ex ante approach to evaluation, using data from the PROGRESA (now Oportunidades) school subsidy experiment in Mexico. Using an economic model of household behavior, they predicted impacts of the subsidy program on the proportion of children attending school. The predictions were based only on children from the control group and calculated the treatment effect from matching control group children from households with a given wage and income with children from households where wages and income would be affected by the subsidy. See chapter 4 for a detailed discussion on matching methods; chapter 9 also discusses Todd and Wolpin's model in greater detail.

Predictions from this model were then compared with ex post experimental impacts (over the period 1997–98) measured under the program. Todd and Wolpin (2006) found that the predicted estimates across children 12 to 15 were similar to the experimental estimates in the same age group. For girls between 12 and 15, they found the predicted increase in schooling to be 8.9 percentage points, compared with the actual increase of 11.3 percentage points; for boys, the predicted and experimental estimates were 2.8 and 2.1 percentage points, respectively.

The ex ante evaluation they conducted also allowed them to evaluate how outcomes might change if certain parameters were altered. An ex ante assessment could also describe the potential range of impacts from the program, which could help in ultimate targeting ex post.

an impact evaluation before implementing the intervention. Chapter 9 provides more case studies of ex ante evaluations.

The Problem of the Counterfactual

The main challenge of an impact evaluation is to determine what would have happened to the beneficiaries if the program had not existed. That is, one has to determine the per capita household income of beneficiaries in the absence of the intervention. A beneficiary's outcome in the absence of the intervention would be its *counterfactual*.

A program or policy intervention seeks to alter changes in the well-being of intended beneficiaries. Ex post, one observes outcomes of this intervention on intended beneficiaries, such as employment or expenditure. Does this change relate directly to the intervention? Has this intervention caused expenditure or employment to grow? Not necessarily. In fact, with only a point observation after treatment, it is impossible to reach a conclusion about the impact. At best one can say whether the objective of the intervention was met. But the result after the intervention cannot be attributed to the program itself.

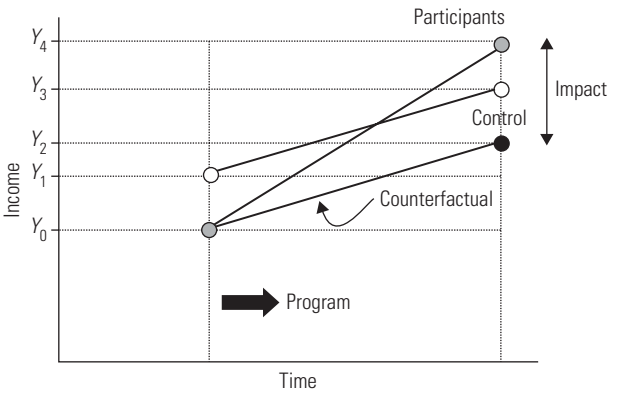
The problem of evaluation is that while the program's impact (independent of other factors) can truly be assessed only by comparing actual and counterfactual outcomes, the counterfactual is not observed. So the challenge of an impact assessment is to create a convincing and reasonable comparison group for beneficiaries in light of this missing data. Ideally, one would like to compare how the same household or individual would have fared with and without an intervention or "treatment." But one cannot do so because at a given point in time a household or an individual cannot have two simultaneous existences—a household or an individual cannot be in the treated and the control groups at the same time. Finding an appropriate counterfactual constitutes the main challenge of an impact evaluation.

How about a comparison between treated and nontreated groups when both are eligible to be treated? How about a comparison of outcomes of treated groups before and after they are treated? These potential comparison groups can be "counterfeit" counterfactuals, as will be discussed in the examples that follow.

Looking for a Counterfactual: With-and-Without Comparisons

Consider the case of Grameen Bank's beneficiaries in Bangladesh. Grameen Bank offers credit to poor women to improve their food consumption. Data, however, show that the per capita consumption among program participants is lower than that of nonparticipants prior to program intervention. Is this a case of failure of Grameen Bank? Not necessarily. Grameen Bank targeted poor families because they had lower per capita food consumption to begin with, so judging the program's impact by comparing the

Figure 2.2 Evaluation Using a With-and-Without Comparison



Source: Authors' representation.

food consumption of program participants with that of nonparticipants is incorrect. What is needed is to compare what would have happened to the food consumption of the participating women had the program not existed. A proper comparison group that is a close counterfactual of program beneficiaries is needed.

Figure 2.2 provides an illustration. Consider the income of Grameen Bank participants after program intervention as Y_4 and the income of nonparticipants or control households as Y_3. This with-and-without group comparison measures the program's effect as $Y_4 - Y_3$. Is this measure a right estimate of program effect? Without knowing why some households participated while others did not when a program such as Grameen Bank made its credit program available in a village, such a comparison could be deceptive. Without such information, one does not know whether Y_3 is the right counterfactual outcome for assessing the program's effect. For example, incomes are different across the participant and control groups before the program; this differential might be due to underlying differences that can bias the comparison across the two groups. If one knew the counterfactual outcomes (Y_0, Y_2), the real estimate of program effect is $Y_4 - Y_2$, as figure 2.2 indicates, and not $Y_4 - Y_3$. In this example, the counterfeit counterfactual yields an underestimate of the program's effect. Note, however, that depending on the preintervention situations of treated and control groups, the counterfeit comparison could yield an over- or underestimation of the program's effect.

Looking for a Counterfactual: Before-and-After Comparisons

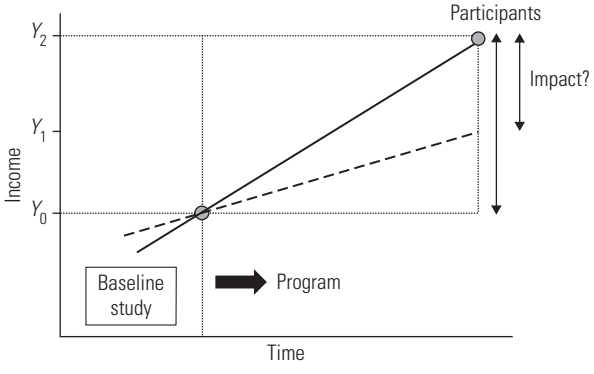
Another counterfeit counterfactual could be a comparison between the pre- and post-program outcomes of participants. One might compare ex post outcomes for beneficiaries with data on their outcomes before the intervention, either with comparable survey

data before the program was introduced or, in the absence of a proper evaluation design, with retrospective data. As shown in figure 2.3, one then has two points of observations for the beneficiaries of an intervention: preintervention income (Y_0) and postintervention income (Y_2). Accordingly, the program's effect might be estimated as ($Y_2 - Y_0$). The literature refers to this approach as the *reflexive method* of impact, where resulting participants' outcomes before the intervention function as comparison or control outcomes. Does this method offer a realistic estimate of the program's effect? Probably not. The time series certainly makes reaching better conclusions easier, but it is in no way conclusive about the impact of a program. Looking at figure 2.3, one sees, for example, that the impact might be ($Y_2 - Y_1$). Indeed, such a simple difference method would not be an accurate assessment because many other factors (outside of the program) may have changed over the period. Not controlling for those other factors means that one would falsely attribute the participant's outcome in absence of the program as Y_0, when it might have been Y_1. For example, participants in a training program may have improved employment prospects after the program. Although this improvement may be due to the program, it may also be because the economy is recovering from a past crisis and employment is growing again. Unless they are carefully done, reflexive comparisons cannot distinguish between the program's effects and other external effects, thus compromising the reliability of results.

Reflexive comparisons may be useful in evaluations of full-coverage interventions such as nationwide policies and programs in which the entire population participates and there is no scope for a control group. Even when the program is not as far reaching, if outcomes for participants are observed over several years, then structural changes in outcomes could be tested for (Ravallion 2008).

In this context, therefore, a broad baseline study covering multiple preprogram characteristics of households would be very useful so that one could control for as

Figure 2.3 Evaluation Using a Before-and-After Comparison



Source: Authors' representation.

many other factors as might be changing over time. Detailed data would also be needed on participation in existing programs before the intervention was implemented. The following chapters discuss several examples of before-and-after comparisons, drawing on a reflexive approach or with-and-without approach.

Basic Theory of Impact Evaluation: The Problem of Selection Bias

An impact evaluation is essentially a problem of missing data, because one cannot observe the outcomes of program participants had they not been beneficiaries. Without information on the counterfactual, the next best alternative is to compare outcomes of treated individuals or households with those of a comparison group that has not been treated. In doing so, one attempts to pick a comparison group that is very similar to the treated group, such that those who received treatment would have had outcomes similar to those in the comparison group in absence of treatment.

Successful impact evaluations hinge on finding a good comparison group. There are two broad approaches that researchers resort to in order to mimic the counterfactual of a treated group: (a) create a comparator group through a statistical design, or (b) modify the targeting strategy of the program itself to wipe out differences that would have existed between the treated and nontreated groups before comparing outcomes across the two groups.

Equation 2.1 presents the basic evaluation problem comparing outcomes Y across treated and nontreated individuals i:

$$Y_i = \alpha X_i + \beta T_i + \varepsilon_i. \tag{2.1}$$

Here, T is a dummy equal to 1 for those who participate and 0 for those who do not participate. X is set of other observed characteristics of the individual and perhaps of his or her household and local environment. Finally, ε is an error term reflecting unobserved characteristics that also affect Y. Equation 2.1 reflects an approach commonly used in impact evaluations, which is to measure the direct effect of the program T on outcomes Y. Indirect effects of the program (that is, those not directly related to participation) may also be of interest, such as changes in prices within program areas. Indirect program effects are discussed more extensively in chapter 9.

The problem with estimating equation 2.1 is that treatment assignment is not often random because of the following factors: (a) purposive program placement and (b) self-selection into the program. That is, programs are placed according to the need of the communities and individuals, who in turn self-select given program design and placement. Self-selection could be based on observed characteristics (see chapter 4), unobserved factors, or both. In the case of unobserved factors, the error term in the estimating equation will contain variables that are also correlated with

the treatment dummy T. One cannot measure—and therefore account for—these unobserved characteristics in equation 2.1, which leads to *unobserved selection bias*. That is, cov $(T, \varepsilon) \neq 0$ implies the violation of one of the key assumptions of ordinary least squares in obtaining unbiased estimates: independence of regressors from the disturbance term ε. The correlation between T and ε naturally biases the other estimates in the equation, including the estimate of the program effect β.

This problem can also be represented in a more conceptual framework. Suppose one is evaluating an antipoverty program, such as a credit intervention, aimed at raising household incomes. Let Y_i represent the income per capita for household *i*. For participants, $T_i = 1$, and the value of Y_i under treatment is represented as $Y_i(1)$. For nonparticipants, $T_i = 0$, and Y_i can be represented as $Y_i(0)$. If $Y_i(0)$ is used across nonparticipating households as a comparison outcome for participant outcomes $Y_i(1)$, the average effect of the program might be represented as follows:

$$D = E(Y_i(1) \mid T_i = 1) - E(Y_i(0) \mid T_i = 0). \tag{2.2}$$

The problem is that the treated and nontreated groups may not be the same prior to the intervention, so the expected difference between those groups may not be due entirely to program intervention. If, in equation 2.2, one then adds and subtracts the expected outcome for nonparticipants had they participated in the program—$E(Y_i(0) / T_i = 1)$, or another way to specify the counterfactual—one gets

$$\begin{aligned} D = {} & E(Y_i(1) \mid T_i = 1) - E(Y_i(0) \mid T_i = 0) \\ & + [E(Y_i(0) \mid T_i = 1) - E(Y_i(0) \mid T_i = 1)]. \end{aligned} \tag{2.3}$$

$$\Rightarrow D = ATE + [E(Y_i(0) \mid T_i = 1) - E(Y_i(0) \mid T_i = 0)]. \tag{2.4}$$

$$\Rightarrow D = ATE + B. \tag{2.5}$$

In these equations, *ATE* is the average treatment effect $[E(Y_i(1) \mid T_i = 1) - E(Y_i(0) \mid T_i = 1)]$, namely, the average gain in outcomes of participants relative to nonparticipants, as if nonparticipating households were also treated. The *ATE* corresponds to a situation in which a randomly chosen household from the population is assigned to participate in the program, so participating and nonparticipating households have an equal probability of receiving the treatment T.

The term B, $[E(Y_i(0) \mid T_i = 1) - E(Y_i(0) \mid T_i = 0)]$, is the extent of selection bias that crops up in using D as an estimate of the *ATE*. Because one does not know $E(Y_i(0) \mid T_i = 1)$, one cannot calculate the magnitude of selection bias. As a result, if one does not know the extent to which selection bias makes up D, one may never know the exact difference in outcomes between the treated and the control groups.

The basic objective of a sound impact assessment is then to find ways to get rid of selection bias $(B = 0)$ or to find ways to account for it. One approach, discussed in

chapter 3, is to randomly assign the program. It has also been argued that selection bias would disappear if one could assume that whether or not households or individuals receive treatment (conditional on a set of covariates, X) were independent of the outcomes that they have. This assumption is called the *assumption of unconfoundedness*, also referred to as the *conditional independence assumption* (see Lechner 1999; Rosenbaum and Rubin 1983):

$$(Y_i(1), Y_i(0)) \perp T_i \mid X_i. \tag{2.6}$$

One can also make a weaker assumption of *conditional exogeneity of program placement*. These different approaches and assumptions will be discussed in the following chapters. The soundness of the impact estimates depends on how justifiable the assumptions are on the comparability of participant and comparison groups, as well as the exogeneity of program targeting across treated and nontreated areas. However, without any approaches or assumptions, one will not be able to assess the extent of bias B.

Different Evaluation Approaches to Ex Post Impact Evaluation

As discussed in the following chapters, a number of different methods can be used in impact evaluation theory to address the fundamental question of the missing counterfactual. Each of these methods carries its own assumptions about the nature of potential selection bias in program targeting and participation, and the assumptions are crucial to developing the appropriate model to determine program impacts. These methods, each of which will be discussed in detail throughout the following chapters, include

1. Randomized evaluations
2. Matching methods, specifically propensity score matching (PSM)
3. Double-difference (DD) methods
4. Instrumental variable (IV) methods
5. Regression discontinuity (RD) design and pipeline methods
6. Distributional impacts
7. Structural and other modeling approaches

These methods vary by their underlying assumptions regarding how to resolve selection bias in estimating the program treatment effect. Randomized evaluations involve a randomly allocated initiative across a sample of subjects (communities or individuals, for example); the progress of treatment and control subjects exhibiting similar preprogram characteristics is then tracked over time. Randomized experiments have the advantage of avoiding selection bias at the level of randomization. In the absence of an experiment, PSM methods compare treatment effects across participant and matched nonparticipant units, with the matching conducted on a range of observed characteristics. PSM methods therefore assume that selection bias is based only on observed characteristics; they cannot account for unobserved factors affecting participation.

DD methods assume that unobserved selection is present and that it is time invariant—the treatment effect is determined by taking the difference in outcomes across treatment and control units before and after the program intervention. DD methods can be used in both experimental and nonexperimental settings. IV models can be used with cross-section or panel data and in the latter case allow for selection bias on unobserved characteristics to vary with time. In the IV approach, selection bias on unobserved characteristics is corrected by finding a variable (or instrument) that is correlated with participation but not correlated with unobserved characteristics affecting the outcome; this instrument is used to predict participation. RD and pipeline methods are extensions of IV and experimental methods; they exploit exogenous program rules (such as eligibility requirements) to compare participants and nonparticipants in a close neighborhood around the eligibility cutoff. Pipeline methods, in particular, construct a comparison group from subjects who are eligible for the program but have not yet received it.

Finally, the handbook covers methods to examine the distributional impacts of programs, as well as modeling approaches that can highlight mechanisms (such as intermediate market forces) by which programs have an impact. These approaches cover a mix of different quantitative methods discussed in chapters 3 to 7, as well as ex ante and ex post methods.

The handbook also draws examples and exercises from data on microfinance participation in Bangladesh over two periods (1991/92 and 1998/99) to demonstrate how ex post impact evaluations are conducted.

Overview: Designing and Implementing Impact Evaluations

In sum, several steps should be taken to ensure that impact evaluations are effective and elicit useful feedback. During project identification and preparation, for example, the importance and objectives of the evaluation need to be outlined clearly. Additional concerns include the nature and timing of evaluations. To isolate the effect of the program on outcomes, independent of other factors, one should time and structure impact evaluations beforehand to help program officials assess and update targeting, as well as other guidelines for implementation, during the course of the intervention.

Data availability and quality are also integral to assessing program effects; data requirements will depend on whether evaluators are applying a quantitative or qualitative approach—or both—and on whether the framework is ex ante, ex post, or both. If new data will be collected, a number of additional concerns need to be addressed, including timing, sample design and selection, and selection of appropriate survey instruments. Also, pilot surveys will need to be conducted in the field so that interview questions can be revised and refined. Collecting data on relevant socioeconomic

characteristics at both the beneficiary level and the community level can also help in better understanding the behavior of respondents within their economic and social environments. Ravallion (2003) also suggests a number of guidelines to improving data collection in surveys. These guidelines include understanding different facets and stylized facts of the program and of the economic environments of participants and nonparticipants to improve sampling design and flesh out survey modules to elicit additional information (on the nature of participation or program targeting, for example) for understanding and addressing selection bias later on.

Hiring and training fieldwork personnel, as well as implementing a consistent approach to managing and providing access to the data, are also essential. During project implementation, from a management perspective, the evaluation team needs to be formed carefully to include enough technical and managerial expertise to ensure accurate reporting of data and results, as well as transparency in implementation so that the data can be interpreted precisely. Ongoing data collection is important to keep program officials current about the progress of the program, as well as, for example, any parameters of the program that need to be adapted to changing circumstances or trends accompanying the initiative. The data need to be analyzed carefully and presented to policy makers and other major stakeholders in the program to allow potentially valuable feedback. This input, in addition to findings from the evaluation itself, can help guide future policy design as well.

References

Banerjee, Sudeshna, Avjeet Singh, and Hussain Samad. 2009. "Developing Monitoring and Evaluation Frameworks for Rural Electrification Projects: A Case Study from Nepal." Draft, World Bank, Washington, DC.

Bourguignon, François, and Francisco H. G. Ferreira. 2003. "Ex Ante Evaluation of Policy Reforms Using Behavioral Models." In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, ed. François Bourguignon and Luiz A. Pereira da Silva, 123–41. Washington, DC: World Bank and Oxford University Press.

Heckman, James J., and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3): 669–738.

Kusek, Jody Zall, and Ray C. Rist. 2004. *A Handbook for Development Practitioners: Ten Steps to a Results-Based Monitoring and Evaluation System.* Washington, DC: World Bank.

Lechner, Michael. 1999. "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification." *Journal of Business Economic Statistics* 17 (1): 74–90.

Paxson, Christina, and Norbert Schady. 2002. "The Allocation and Impact of Social Funds: Spending on School Infrastructure in Peru." *World Bank Economic Review* 16 (2): 297–319.

Rao, Vjayendra, and Ana María Ibáñez. 2005. "The Social Impact of Social Funds in Jamaica: A 'Participatory Econometric' Analysis of Targeting, Collective Action, and Participation in Community-Driven Development." *Journal of Development Studies* 41 (5): 788–838.

Ravallion, Martin. 2003. "Assessing the Poverty Impact of an Assigned Program." In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, ed. François Bourguignon and Luiz A. Pereira da Silva, 103–22. Washington, DC: World Bank and Oxford University Press.

———. 2008. "Evaluating Anti-Poverty Programs." In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3787–846. Amsterdam: North-Holland.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Schady, Norbert. 1999. "Seeking Votes: The Political Economy of Expenditures by the Peruvian Social Fund (FONCODES), 1991–95." Policy Research Working Paper 2166, World Bank, Washington, DC.

Todd, Petra, and Kenneth Wolpin. 2006. "Ex Ante Evaluation of Social Programs." PIER Working Paper 06-122, Penn Institute for Economic Research, University of Pennsylvania, Philadelphia.