Summary

A poverty monitoring and evaluation system is required to determine whether a country’s overall poverty reduction strategy, and its main components, are effective.

The first step in poverty monitoring is to define the goals of the strategy. Operationally, this requires the identification of measurable indicators and the establishment of realistic targets that can help policy makers set priorities.

Final indicators measure the outcomes of poverty reduction policies and the impact on dimensions of well-being. Intermediate indicators measure the inputs into a program and the outputs of the program. A good indicator is unambiguous, relevant, reliable, and sensitive to changes, and may be tracked cheaply and frequently. It can also be disaggregated, for instance by geographic area or gender.

Impact evaluation seeks to measure the changes in well-being that can be attributed to a particular project or policy (an “intervention”); the results can help inform decisions about whether the intervention should be expanded or eliminated. The central challenge of impact assessment is constructing a plausible counterfactual.

Several evaluation designs are used in impact evaluation, depending in part on the data that are available. Among the most important are experimental design (randomization); and quasi-experimental designs, including matching comparisons (typically using propensity scores), double differences, instrumental variables (“statistical control”), and reflexive comparisons. Both experimental and quasi-experimental methods try to tackle the problem of sample bias that bedevils impact analysis.

It is difficult to measure the impact of economy-wide shocks, although attempts have been made by measuring deviations from trends, building computable general
equilibrium and other simulation models, applying panel data, and asking households to assess how much they have been affected. While no method of impact evaluation is perfect, such evaluations have had an important influence on policy decisions.

**Learning Objectives**

After completing the chapter on *Poverty Monitoring and Evaluation*, you should be able to

1. Describe the function of a monitoring and evaluation system and explain why it is useful.
2. Summarize the three steps in poverty monitoring.
3. Distinguish between the different categories of poverty indicators, and list the characteristics of a good indicator.
4. Explain why it is necessary to be able to disaggregate indicators.
5. Describe the purposes of impact evaluation.
6. Explain why impact evaluation requires the construction of a counterfactual and why this is difficult to do.
7. Summarize the steps required in an experimental design, and assess the applicability of this method of impact evaluation.
8. For each of the main types of quasi-experimental design—matching comparisons, double differences, instrumental variables, and reflexive comparisons—summarize the steps needed to apply them, and the data requirements, and evaluate their applicability and usefulness.
9. Summarize the principal methods that have been used to measure the impact of economy-wide shocks, and critically assess the value of each method.

**Introduction**

A country has developed a poverty reduction strategy and put in place several specific measures to combat poverty, including a food-for-work program, supplemental nutrition packages for mothers and infants, free school textbooks in poor villages, and accelerated construction of rural roads.
Two questions naturally arise:

- Is the overall strategy effective?
- How large is the impact of each of the main components of the strategy?

The answers to these questions require monitoring and evaluation.

A poverty monitoring system tracks key indicators of poverty over time and space. The resulting data can then be used to evaluate the program. Process evaluation examines how the programs operate, and focuses on problems of service delivery. Cost-benefit analysis and cost-effectiveness analysis weigh program costs against the benefits they deliver. They in turn require thorough impact evaluations, which quantify the effects of programs on individuals and households.

This chapter summarizes the elements required for a good monitoring system and introduces the techniques of impact evaluation at the micro level and at the level of the whole economy. Prennushi, Rubio, and Subbarao (2000) provide an excellent introduction to monitoring and evaluation, Baker (2000) has compiled a useful handbook on impact evaluation, Ravallion’s witty and accessible introduction (1999) to some of the finer points of impact evaluation is well worth reading, and Haughton and Haughton (forthcoming) provide a somewhat more formal treatment. A growing number of impact evaluations are now available on the Web and can serve as templates for new evaluations; for a useful list, see www.worldbank.org/poverty (and follow the links for Impact Evaluation and then Selected Evaluations).

**Poverty Monitoring**

The first challenge in monitoring progress toward poverty reduction is to

- Identify the goals that the strategy is designed to achieve, such as “eradicate hunger” or “halve poverty within a decade.”

- Select the key indicators that measure progress toward the goals, such as the proportion of individuals consuming less 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, halve the number of households living on less than a dollar a day by the year 2015.

The Millennium Development Goals consist of a set of goals, indicators, and targets that the countries of the world have agreed to pursue. They are summarized in table 13.1 and give a good sense of the nature and scope of goals, indicators, and targets that individual countries may want to achieve.
<table>
<thead>
<tr>
<th>Goals and targets</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal 1: Eradicate extreme poverty and hunger</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Target 1</strong> Halve, between 1990 and 2015, the proportion of people whose income is less than $1 a day</td>
<td>1. Proportion of population below $1 a day</td>
</tr>
<tr>
<td><strong>Target 2</strong> Halve, between 1990 and 2015, the proportion of people who suffer from hunger</td>
<td>2. Poverty gap ratio (incidence x depth of poverty)</td>
</tr>
<tr>
<td><strong>Goal 2: Achieve universal primary education</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Target 3</strong> Ensure that, by 2015, children everywhere, boys and girls, will be able to complete a full course of primary schooling</td>
<td>3. Share of poorest quintile in national consumption</td>
</tr>
<tr>
<td><strong>Goal 3: Promote gender equality and empower women</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Target 4</strong> Eliminate gender disparity in primary and secondary education, preferably by 2005 and to all levels of education no later than 2015</td>
<td>4. Prevalence of underweight children (under 5 years of age)</td>
</tr>
<tr>
<td><strong>Goal 4: Reduce child mortality</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Target 5</strong> Reduce by two-thirds, between 1990 and 2015, the under-5 mortality rate</td>
<td>5. Proportion of population below minimum level of dietary energy consumption</td>
</tr>
<tr>
<td><strong>Goal 5: Improve maternal health</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Target 6</strong> Reduce by three-quarters, between 1990 and 2015, the maternal mortality ratio</td>
<td>6. Net enrollment ratio in primary education</td>
</tr>
<tr>
<td>****</td>
<td>7. Proportion of pupils starting grade 1 who reach grade 5</td>
</tr>
<tr>
<td>****</td>
<td>8. Literacy rate for 15-to 24- year-olds</td>
</tr>
<tr>
<td>****</td>
<td>9. Ratio of girls to boys in primary, secondary and tertiary education</td>
</tr>
<tr>
<td>****</td>
<td>10. Ratio of literate females to makes of 15- to 24-year-olds</td>
</tr>
<tr>
<td>****</td>
<td>11. Share of women in wage employment in the nonagricultural sector</td>
</tr>
<tr>
<td>****</td>
<td>12. Proportion of seats held by women in national parliament</td>
</tr>
<tr>
<td>****</td>
<td>13. Under-5 mortality rate</td>
</tr>
<tr>
<td>****</td>
<td>14. Infant mortality rate</td>
</tr>
<tr>
<td>****</td>
<td>15. Proportion of 1 year olds immunized against measles</td>
</tr>
<tr>
<td>****</td>
<td>16. Maternal mortality ratio</td>
</tr>
<tr>
<td>****</td>
<td>17. Proportion of births attended by skilled health personnel</td>
</tr>
</tbody>
</table>
Goal 6: Combat HIV/AIDS, malaria, and other diseases

<table>
<thead>
<tr>
<th>Target 7</th>
<th>Have halted by 2015, and begun to reverse, the spread of HIV/AIDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18. HIV prevalence among 15- to 24-year-old pregnant women</td>
</tr>
<tr>
<td></td>
<td>19. Contraceptive prevalence rate</td>
</tr>
<tr>
<td></td>
<td>20. Number of children orphaned by HIV/AIDS</td>
</tr>
<tr>
<td></td>
<td>22. Proportion of population in malaria risk areas using effective malaria prevention and treatment measures</td>
</tr>
<tr>
<td></td>
<td>23. Prevalence and death rates associated with tuberculosis</td>
</tr>
<tr>
<td></td>
<td>24. Proportion of TB cases detected and cured under DOTS (Directly observed treatment short course)</td>
</tr>
</tbody>
</table>

Goal 7: Ensure environmental sustainability

<table>
<thead>
<tr>
<th>Target 9</th>
<th>Integrate the principles of sustainable development into country policies and programs and reverse the loss of environmental resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25. Change in land area covered by forest</td>
</tr>
<tr>
<td></td>
<td>26. Land area protected to maintain biological diversity</td>
</tr>
<tr>
<td></td>
<td>27. GDP per unit of energy use (as proxy for energy efficiency)</td>
</tr>
<tr>
<td></td>
<td>28. Carbon dioxide emissions (per capita)</td>
</tr>
<tr>
<td>Target 10</td>
<td>Halve, by 2015, the proportion of people without sustainable access to safe drinking water</td>
</tr>
<tr>
<td></td>
<td>29. Proportion of population with sustainable access to an improved water source</td>
</tr>
<tr>
<td>Target 11</td>
<td>By 2020, to have achieved a significant improvement in the lives of at least 100 million slum dwellers</td>
</tr>
<tr>
<td></td>
<td>30. Proportion of people with access to improved sanitation</td>
</tr>
<tr>
<td></td>
<td>31. Proportion of people with access to secure tenure</td>
</tr>
</tbody>
</table>

Selecting Indicators

It is helpful to classify indicators into two groups:

- **Final indicators** measure the outcomes of poverty reduction policies (for example, increased school attendance rates) and the impact on dimensions of well-being (for example, higher literacy rates).
- **Intermediate indicators** measure the inputs into a program (for example, spending on health care) and the outputs of the program (for example, clinics built, doctors hired).

Viewed in this way, a poverty monitoring system encompasses both implementation monitoring as well as performance (or results-based) monitoring. Intermediate indicators typically change more quickly than final indicators, respond more rapidly to public interventions, and can be measured more easily and in a more timely fashion.

A good indicator has several qualities, including the following:

- It is a direct unambiguous measure of progress and is easy to understand.
- It is relevant to the objectives, which in the current context concern poverty reduction.
- It varies across areas, groups, and time.
- It is sensitive to changes in policies, programs, and “shocks.”
- It is reliable and hard to manipulate.
- It can be tracked frequently and cheaply.

It is essential to be able to disaggregate the indicators in ways that are useful—for instance, by geography (urban vs. rural, by administrative region, by geoclimatic zone), by gender, by socially defined group (ethnic, linguistic, religious), and by income or expenditure level. The disaggregation is central to the political economy of poverty reduction, but it is also highly sensitive. For instance, Malaysia does not make its household survey data available for public use, in part because of concerns about what the data might reveal about the evolution of the ethnic breakdown of poverty and income.

Setting Targets

There are two reasons to set concrete targets for poverty reduction: (1) it forces decision makers to clarify their priorities and adjust the allocation of resources in consequence; and (2) it strengthens accountability.
The standard advice is to set a small number of targets in clear and unambiguous terms. The targets need to be consistent with other national goals, and to be achievable, given the country’s overall level of economic development and its implementation capacity. They also need to be realistic about the resources that will be available.

One of the challenges in setting up a good poverty monitoring system is collating the data and channeling it to policy makers and the public in a timely and coherent manner. The relevant data will come from many sources, including administrative information on public spending or the number of teachers, project information, household survey data, and so on. Thus the development of a poverty monitoring system requires attention to the statistical system as a whole. Some countries have established poverty monitoring units, sometimes attached to the prime minister’s office, to manage the flow of data related to poverty reduction.

Prennushi, Rubio, and Subbarao (2000) make the case that poverty monitoring systems in most less-developed countries need to pay more attention to speeding up the flow of actual expenditure data, because expenditure tracking is often quite slow. They argue for greater use of cost accounting so, for instance, one can more easily determine the cost of educating a child or serving a hospital outpatient. They call for improving the accuracy of data on actual spending, if necessary through the use of expenditure tracking surveys. In an often-cited example, such a survey found that in 1991–95 in Uganda, only 30 percent of the nonsalary funds intended for public schools actually reached the schools; the remaining funds were siphoned off by district administrations. Additionally, Prennushi, Rubio, and Subbarao hope to see a speedier analysis of household survey data, which are sometimes out of date by the time they are published. In this context, simple and rapid surveys based on Core Welfare Indicator Questionnaires (CWIQ) may have a role to play, although as discussed in chapter 2, such methods cannot generate accurate proxies for income or expenditure.\(^1\)

**Review Questions**

1. Which of the following is **not** one of the three main tasks in poverty monitoring?

   - A. Set targets.
   - B. Identify goals.
   - C. Select key indicators.
   - D. Measure the impact of policies.

2. A key Millennium Development Goal is to halve, between 1990 and 2015, the number of people living on less than a dollar a day.

   - True
   - False
Impact Evaluation: Micro Projects

Suppose that a nongovernmental organization or government sets up a microcredit project or builds an irrigation canal. How does one

- Forecast the impact of projects such as these on poverty reduction or other development objectives?
- Find out if the policy or program is cost-effective?
- Improve the design of programs and interventions?

To answer these questions an impact evaluation is required. Generally, an impact evaluation seeks to measure the changes in well-being that can be attributed to a particular project or policy (an “intervention”). The results of an impact evaluation can show which interventions have been effective, and thus inform decisions on whether they should be eliminated, modified, or expanded, as well as what priority they should be accorded.

Impact evaluations are expensive and can be technically complex. It thus makes sense to undertake one only if the policy or program is of strategic importance, or is innovative, and the information from the evaluation is likely to fill gaps in current knowledge. It is also important to master the institutional details of the program or project that is being analyzed before proceeding with the statistical analysis. And the validity of the quantitative part of the evaluation is only as good as the quality of the data that can be brought to bear on the issue.

Not every manager welcomes an impact evaluation, particularly if the results are expected to show a project in a poor light. Ravallion (2008) suggests that, as a consequence, too few evaluations are undertaken, and those that are done are biased toward programs that work well or programs that produce quick, measurable benefits. Such biases in turn weaken the potential usefulness of impact evaluations in general. He also suggests that there is underinvestment in research on the extent to which impact evaluations may be generalized—their external validity—and that all too often evaluators fit their favored techniques to convenient problems, rather than starting with a problem and asking how to measure its impact.

3. A good indicator includes all of the following except:

- A. It can be adjusted to accommodate varying political needs.
- B. It can be tracked frequently and cheaply.
- C. It is sensitive to shocks.
- D. It is an unambiguous measure of progress.
The Challenge of the Counterfactual

To evaluate the impact of an intervention, we need to compare the actual outcome with our evaluation of what would have happened in the absence of the intervention (the counterfactual). The central challenge of impact assessment is constructing a plausible counterfactual.

The challenge is a difficult one. Consider the case of a program that provides additional food—for example maize or milk powder—to poor mothers with infants. Now suppose that the data show that the mothers and infants covered by the program are less well nourished than those who are not covered. Are we to conclude that the project is a failure?

Perhaps; but then again, it is likely that the project targeted poor mothers with malnourished infants, so it is not surprising that households with underweight children are getting additional food. The problem here is one of estimating how malnourished the mothers and infants covered by the program would have been in the absence of the program, in other words, establishing an appropriate counterfactual. A number of methods (“evaluation designs”) have been developed to address questions of this sort, and we now examine these one by one.

Experimental Design

Widely regarded as the best approach to measuring the impact of an intervention, the idea behind experimental design, also known as randomization, is as follows:

1. Before the intervention begins, identify those who are eligible to benefit from it.
2. Then randomly select those who will benefit (the treatment group) and deny the benefit to the others (who will serve as the control group).
3. After the intervention, measure the appropriate outcome variables (for example, degree of malnutrition, poverty gaps).

The difference in the mean values of the outcome variables between the control and treatment groups can be attributed to the effects of the intervention, give or take some sampling error.

In practice, researchers more commonly gather information on income, individual and household characteristics \((X)\), and village and community characteristics \((V)\), for the individuals that do, and do not, participate in the scheme. Then they estimate

\[
Y_{ij} = X_{ij}\beta + V_{ij}\gamma + C_{ij}\delta + \epsilon_{ij},
\]  

(13.1)
where $C_{ij}$ is a dummy variable that is equal to 1 if the individual $i$ in village $j$ participates in the scheme, and 0 otherwise. In this case, the individual, household, and community characteristics control for other differences, and it is reasonable to expect that estimated coefficient $\delta$ would measure the impact of the intervention.

In the context of poverty interventions, it is rarely possible to use randomization, although the study by Glewwe, Kremer, and Moulin (2000) of the effects of textbooks on learning in Kenya was able to do so by randomly selecting which schools would receive textbooks. In this case, there was no evidence that the provision of additional textbooks raised average test scores or reduced dropout rates.

Another good example is the study by Angrist et al. (2002) of school vouchers in Colombia. In 1991, the government of Colombia established the PACES (Programa de Ampliacion de Cobertura de la Educacion Secundaria, or Program for the Expansion of Educational Coverage) program, which provided vouchers (that is, scholarships) to students who had applied to and been accepted into private secondary schools. The vouchers were awarded based on a lottery; this provided the randomization that allowed the authors to compare the outcomes for applicants who received vouchers with the outcomes for those who did not. Among the more interesting findings are the following:

- Voucher winners were 15–16 percentage points more likely to be in private school when they were surveyed in 1998.

- The program had a positive and significant effect on the number of years of schooling completed. Those who received vouchers in 1995 in the capital (Bogotá) had completed 0.12–0.16 more years than those who did not.

- Repetition rates fell significantly as a result of the project. In the 1995 Bogotá sample, the probability of repetition was reduced by 5–6 percentage points for lottery winners.

The most serious problem with randomized experiments is that the withholding of treatment may be unethical. For instance, if we are trying to determine the effects of providing vitamin A supplementation, which helps prevent blindness, it is likely to be unethical to withhold this inexpensive treatment from significant numbers of young children. It also may be politically difficult to provide a treatment to one group and not to another. In some cases—if a program is applied universally, for instance—there may be no control group. True random assignment is often difficult in practice. And people may respond to a program, by moving into or out of a treatment area, for instance, thereby contaminating the results. Randomization may be subject to unobserved heterogeneity bias that affects the outcomes of treatment.

So, in practice, randomization can either produce inconsistent results or cannot be implemented, in which case most impact assessments have to rely on quasi-experimental methods.
**Quasi-Experimental Methods**

If households are not assigned randomly to an intervention—such as food stamps, vaccinations, or irrigation water—then those who benefit are unlikely to be typical of the eligible population. There are two reasons for this. First, there may be nonrandom program placement, of which the researcher may or may not be aware; for instance, an antipoverty program may be more likely to be set up in poor villages. This is the problem of *unobserved area heterogeneity*. Second, there may be self-selection into program participation; for instance, more dynamic individuals may be the first to sign up, or program benefits may flow to those who are politically well connected, or sick people may move to villages that have been equipped with clinics. Such effects are often hard to detect, and give rise to the problem of *unobserved individual and household heterogeneity*.

The presence of these unobservables immediately creates the problem of *selection bias*. The basic idea behind quasi-experimental designs is to construct statistical models of selection—matching, double differences, instrumental variables, reflexive comparisons—that permit one to compare program participants and nonparticipants (the *comparison* group) holding the selection processes constant.

To see why these problems arise, suppose we are interested in determining whether a microcredit scheme, initiated in time period 0, raises the income of individual $i$ in time period 1. An appealing approach would be to collect data on the outcome indicator (income, given by $Y_{i1}$), and on individual and household characteristics ($X_{i1}$), for a sample of individuals that do, and do not, participate in the scheme. Then we could estimate an impact equation of the form—

$$Y_{i1} = X_{i1}\beta + P_{i1}\delta + \varepsilon^{Y}_{i1},$$  \hspace{1cm} (13.2)

where $P_{i1}$ is a dummy variable that is set equal to 1 if the individual $i$ participates in the scheme and to 0 otherwise. At first sight, it would appear that the value of the estimated coefficient $\delta$ would measure the impact of the microcredit scheme on income.

Unfortunately, this is unlikely to be the case, because program participation is often related to the other individual, household, and village variables, some of which may not be observable. For instance, those who borrow money may be better educated, or younger, or live in villages with a loan office, or be more motivated. With enough information it may be possible to control for many of these variables, including education and age, but it is never possible to control for all the relevant effects. For example, the degree of individual motivation is unobservable; but a more motivated individual is more likely to participate in the program (a higher $P_{i1}$) and to benefit more from it (a higher $P^{Y}_{i1}$). This creates a correlation between $P_{i1}$ and $\varepsilon^{Y}_{i1}$ and so leads to a biased estimate of $\delta$. As a practical matter, unobservables are always present in
such circumstances, and thus this selection bias (which may also be thought of as a form of omitted variable bias) is present as well.

The path to a solution requires us to envisage a separate program participation equation of the form—

\[ P_y = Z_{it} \gamma + \epsilon_{it}^p. \]

(13.3)

where the \( Z \) variables may be the same as the \( X \) variables, or include additional variables. If one can identify a set of variables that affect only participation, equation (13.3), and not the household outcome, equation (13.2)—generally a difficult task—then it may be possible to arrive at a satisfactory estimate of \( \delta \), the impact of program participation on the outcome of interest. Many quasi-experimental evaluations have been informative; they can often be done fairly quickly and cheaply, and do not necessarily require the collection of data before the project begins. We now consider some specific solutions in more detail.

**Review Questions**

4. The central challenge of impact assessment is constructing a plausible counterfactual.

   - True
   - False

5. Ideally, with randomization,

   - A. We randomly pick a sample of treated and nontreated individuals.
   - B. We randomly assign the treatment to individuals.
   - C. We randomly select those who have been treated and compare them with a nontreated group.
   - D. All of the above.

6. Selection bias may result from all of the following except:

   - A. Nonrandom program placement.
   - B. Unobserved area heterogeneity.
   - C. Unobserved household heterogeneity
   - D. Random assignment.

**Solution 1. Matching Comparisons.** This approach is widely used and is often feasible even if experimental design is not possible. To undertake matching, one needs survey data for a substantial number of nonparticipants as well as for the participants. The basic idea is to match each participant with a similar nonparticipant (or a small “comparison group”) and then to compare the outcomes between them.
Given survey information, the most common procedure starts by pooling the two samples (that is, the participants and nonparticipants, or in the jargon of matching, the treatment and comparison groups) and estimating a logit model of program participation as a function of all the variables that might influence participation—equation (13.3). Ironically, one does not want the equation to fit too well, because that would make it difficult to identify nonparticipants who are otherwise similar to participants.

The next step is to generate the propensity score, which is the predicted probability of participation, given the observed characteristics $Z$. Some of the individuals in the comparison group may have propensity scores that are far outside the range of the treatment group—they are said to have a “lack of common support”—and these cases may need to be dropped and the logit model reestimated.

Next, for each person in the treatment group, find the member of the comparison group with the closest propensity score (the “nearest neighbor”), or a small group of, say, five nearest neighbors. Compute the difference between the outcome indicator of the person in the treatment group and the mean of the outcome indicators for the nearest neighbors. The mean of these differences, over all the members of the treatment group, gives a measure of the overall impact of the program.

When the correlates of participation in the project are observable, this approach works well, but it is not satisfactory if unobservable differences are important—for instance, if those who sign up for microcredit are the more dynamic individuals. The procedure fails in this case because the differences between the treatment and comparison groups cannot be entirely attributed to whether or not they participated in the program; some, or even most, of the difference may be due to (possibly unobserved) differences in the inherent characteristics of individuals in the two groups.

Box 13.1 Case Study: Workfare and Water in Argentina

The Trabajar II program in Argentina was introduced in 1997 in response to a sharp rise in unemployment. The program provided low-wage work on community projects, and was intended to raise the incomes of the poor.

To analyze the impact of this “workfare” program, Jalan and Ravallion (1999) used the results of the 1997 Encuesta de Desarrollo Social (Social Development Survey), coupled with a similar survey of participants in the Trabajar program, to estimate a logit model of program participation. They used variables such as gender, schooling, housing, and subjective perceptions of welfare, and used the data to derive propensity scores for participants and nonparticipants (after taking care to limit the sample of nonparticipants to those with “common support”).
Solution 2. Double Differences. Also known as the difference-in-difference method, this approach requires information both from a baseline survey before the intervention occurs and a follow-up survey after the program is operating. Both surveys should be comparable in the questions used and the survey methods applied, and they must be administered both to participants and nonparticipants.

In the simplest version, compute the difference between the outcome variable after \((Y_1)\) and before \((Y_0)\) the intervention, both for the treatment and comparison samples. The difference between these two gives an estimate of the impact of the program.

Example: Suppose that the literacy rate rose from 25 percent to 35 percent for the treatment sample, between the beginning and end of an adult literacy project, and that the literacy rate rose from 28 percent to 34 percent over the same period for the comparison group. Then the project may be considered to have raised the literacy rate, for those treated, by 4 percentage points. The logic is that one might have expected literacy to rise by 6 percentage points for everyone, judging by the experience of the comparison group; however, for the treatment group, it rose by 10 percentage points, of which 4 percentage points may thus be attributable to the project.

The double difference method may be refined in a number of ways. By using propensity score matching with data from the baseline survey, one can ensure that the comparison group is similar to the treatment group. And one could use
a differenced form of the regression in equation (13.2) to get a better estimate of the impact of the project (see Ravallion 1999, 23–24, for details).

Done right, this is a relatively satisfactory approach, but it will give biased results if there is selective attrition of the treatment group—for example, if some of the treatment group cannot be resurveyed a second time, or if those who drop out are not a random sample of the treatment group (for instance, if they are older or richer than their peers in the treatment group). The double difference method is also relatively expensive to implement, as it requires at least two rounds of survey data.

Solution 3. Instrumental Variables. Sometimes referred to as the *statistical control* method, the idea behind this widely used method is to identify variables that affect participation in the program, but not the outcomes of interest—that is, that enter into equation (13.3) but not equation (13.2). The estimates of equation (13.3) are used to predict participation, and this *predicted* participation is then used in equation (13.2). By using predicted, rather than actual, participation in equation (13.1), one removes (in principle) the biases that would otherwise contaminate the estimates of the impact coefficient $\delta$.5

To see why this works, consider the case of a dynamic individual who, we assume, might be more likely to participate (so $\epsilon_{Pi1}^{P}>0$) and to perform well (so $\epsilon_{Yi1}^{Y}>0$). As a result, $\epsilon_{Yi1}^{Y}$ and $Pi1$ would be correlated and the estimate of $\delta$ (the impact effect) biased. But by using $\hat{P}_{i1}$ instead of $P_{i1}$, the forces that influence $\epsilon_{Yi1}^{Y}$ and $P_{i1}$ now only affect $\epsilon_{Yi1}^{Y}$, but not $\hat{P}_{i1}$, so the correlation disappears along with the bias. This, however, is true only if there are influences on $P_{i1}$ that do not influence $Y_{i1}$. The idea is to create variation in $\hat{P}_{i1}$ so that we have some people in the sample who, even if they have the same $X_{i1}$, may have different $P_{i1}$; in effect, we now have a source of variation in $Y_{i1}$ that is attributable to the program.

The major practical problem is finding appropriate instruments that influence program participation but not the outcome of the program once one is enrolled. However, it is sometimes possible. A recent study of the effect of famine relief in Ethiopia was able to use past climatic variation as one such instrument (Yamano, Alderman, and Christiaensen 2003); and interventions that are undertaken randomly in some villages but not others clearly provide a suitable instrument, because living in a given village determines whether you will be covered by the intervention, but now how much you will profit from it. Pitt and Khandker (1998) used the exogenous program eligibility condition as an instrument to identify the impact of microcredit on household welfare (see box 13.2).

The instrumental variables method is helpful if there is measurement error. Suppose that, because of measurement errors, observed program participation is more variable than true participation; this will lead to an underestimation of the impact of the program (“attenuation bias”), essentially because the noise of
Box 13.2 Case Study: Microfinance and the Poor in Bangladesh

Microfinance, or the provision of small loans to the poor, is often touted as an important tool for reducing poverty. A widely admired model is the Grameen Bank in Bangladesh, brainchild of Mohammed Yunus, which provides small loans to poor people, mainly women. The Bank also runs related education programs. The Bangladesh Rural Advancement Committee and the Bangladesh Rural Development Board run similar operations.


These questions have been addressed using information collected in three postharvest surveys undertaken in Bangladesh in 1991 and 1992, covering 1,798 households in 87 villages. The villages were chosen from 24 subdistricts where microcredit programs had been implemented for at least three years before the survey and from five subdistricts where they had not been implemented. The programs target households who own less than half an acre of land, but not all of the targeted households borrow.

One may divide up the surveyed households as shown here:

**Surveyed Households**

<table>
<thead>
<tr>
<th>Nontarget households (own more than half an acre of land)</th>
<th>Villages with a program in place</th>
<th>Villages without a program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target/eligible households that ...</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Do not or cannot participate</td>
<td>C</td>
<td>E</td>
</tr>
<tr>
<td>Participate</td>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

One possibility would be to compare the impact of the program for households D with that of households C. The results are shown in the box table below:

**Mean of Individual and Household Outcomes by Program Participation**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Program households (D)</th>
<th>Eligible nonprogram households (C)</th>
<th>Noneligible households (A U B U E)</th>
<th>All households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys’ school enrollment rate</td>
<td>45.4</td>
<td>33.3</td>
<td>52.8</td>
<td>43.9</td>
</tr>
<tr>
<td>(age 5–25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Girls’ school enrollment rate</td>
<td>43.7</td>
<td>35.8</td>
<td>49.2</td>
<td>41.5</td>
</tr>
<tr>
<td>(age 5–25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The problem with this comparison is that there is likely to be selection bias; it appears, for instance, that those who are eligible for microfinance but do not participate say that they are concerned about their ability to repay, suggesting that they are more risk averse, or perhaps less motivated, than borrowers. If landowning is exogenous, then one could compare the outcomes of households C and D on the one hand with households E on the other. This assumes, however, that program and nonprogram villages are the same. An even better approach would be to use a double difference, by comparing $Y_{C U D} - Y_A$ with $Y_E - Y_B$, where $Y$ is the impact of interest (consumption or income per capita, school enrollment rates, and the like).

Morduch (1998) used a similar approach and found that the microfinance programs appeared to improve the outcomes of interest. However, on further examination, he found that the programs were seriously mistargeted, with 20–30 percent of the borrowers owning more than the half-acre maximum. When these borrowers were excluded, the microfinance programs had essentially no discernible effects, except perhaps for lowering the variance of consumption and income for participants. On the other hand, it is possible that the comparison households may not have lacked access to microfinance, in which case Morduch’s conclusion would be unsurprising. A serious problem with Morduch’s simple difference-in-difference analysis is that program participation is exogenous or randomly given.

Because of sample selection bias that is inherently present with such nonrandom distribution of borrowers and nonborrowers, Pitt and Khandker (1998) used an instrumental variable method to estimate the program effect. The instrument is based on the exogenous land-based eligibility conditions. They found that program participation matters a lot when sample selection bias is corrected. However, the results are sensitive to the assumption of the land-based exogenous eligibility condition.

A follow-up panel survey was undertaken in 1998–99 to measure the effects of microfinance. Khandker (2005) estimated a number of regressions, using both instrumental variables and fixed effects (at the level of villages and households). A selection of results for

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Program households (D)</th>
<th>Eligible nonprogram households (C)</th>
<th>Noneligible households (A U B U E)</th>
<th>All households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current contraceptive use rate (currently married women, 14–50)</td>
<td>41.9</td>
<td>35.9</td>
<td>35.6</td>
<td>36.9</td>
</tr>
<tr>
<td>Household weekly per capita consumption (taka)</td>
<td>86.4</td>
<td>78.2</td>
<td>124.8</td>
<td>95.7</td>
</tr>
</tbody>
</table>

Source: Khandker 2000.
female participants is shown in the box table below; each row represents a different dependent variable, and each column a different estimation technique. Although the magnitudes of the impacts vary with the technique used, the overall results are more positive for borrowing by women than for men. That is, impacts are sensitive to controlling for household-specific unobservables.

Based on these and other findings, Khandker (2005) concludes that microcredit can reduce poverty in a cost-effective way, and benefits do flow to women. While a subsidy is required to develop the initial institutions, this subsidy dependence can be reduced over time (Khandker 1998).

Khandker also found that the ultrapoor do not join microcredit programs. A similar conclusion was reached by Patten and Rosengard (1991) in their evaluation of the microlending activities of Bank Rakyat Indonesia (BRI) in the 1980s; however BRI, unlike the Grameen Bank, did not require a subsidy for its microcredit.

### Impact of Women's Borrowing from Grameen Bank on Individual and Household Outcomes

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Naïve model</th>
<th>Instrumental variables (IV) method</th>
<th>Village-level fixed effects &amp; IV method</th>
<th>Household-level fixed-effects model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys' school enrollment rate (age 5–25)</td>
<td>0.062&lt;sup&gt;a&lt;/sup&gt; (4.461)</td>
<td>0.131 (4.022)</td>
<td>0.103 (2.364)</td>
<td>0.013 (2.587)</td>
</tr>
<tr>
<td>Girls' school enrollment rate (age 5–25)</td>
<td>0.019&lt;sup&gt;a&lt;/sup&gt; (1.412)</td>
<td>0.085 (2.289)</td>
<td>0.013 (0.334)</td>
<td>0.016 (3.405)</td>
</tr>
<tr>
<td>Current contraceptive use rate (currently married women, age 14–50)</td>
<td>0.026&lt;sup&gt;a&lt;/sup&gt; (1.942)</td>
<td>0.095 (2.580)</td>
<td>−0.091 (−2.011)</td>
<td>−0.001 (−0.260)</td>
</tr>
<tr>
<td>Household weekly per capita consumption (taka)</td>
<td>0.004&lt;sup&gt;b&lt;/sup&gt; (1.765)</td>
<td>0.037 (2.174)</td>
<td>0.043 (4.249)</td>
<td>0.010 (2.697)</td>
</tr>
</tbody>
</table>


*Note:* Figures in parentheses are t-statistics. Household-level fixed effects estimates are preliminary.

- a. probit
- b. OLS.
- c. Results in column 3 are considered to be the best.
measurement error is getting in the way of isolating the effects of program participa-
tion. However, the predicted value of program intervention ($\hat{P}_{it}$) is less likely to reflect measurement error and can reduce the effects of attenuation bias.

**Solution 4. Reflexive Comparisons.** In this approach, one first undertakes a baseline survey of the treatment group before the intervention, with a follow-up survey afterward. The impact of the intervention is measured by comparing the before-and-after data; in effect, the baseline provides the comparison group.

Such comparisons are rarely satisfactory. The problem in this case is that we really want a “with” and “without” comparison, not a “before” and “after.” Put another way, in the reflexive comparison method, there is no proper counterfactual against which the outcomes of the project may be compared. There is also a problem if attrition occurs, so that some of those surveyed before the project drop out in some systematic way. On the other hand, this may be the only option in trying to determine the impact of full-coverage interventions, such as universal vaccinations, where there is no possibility of a comparison or control group.

### Review Questions

**7. Which of the following is not a step in propensity score matching?**

- A. Compare each treated case with its nearest untreated neighbor.
- B. Find the average difference between treated and matched comparators.
- C. Compute the change in the gap between treated and comparators at two points in time.
- D. Estimate a participation equation using logit or probit.

**8. A potable water project raised connections in a project area from 16 percent to 28 percent. The number of connections in a comparator area rose from 14 percent to 25 percent during the same period. Using double differences, the impact of the project on connections was,**

- A. 12 percent.
- B. 3 percent.
- C. 1 percent.
- D. 5 percent.

**9. Reflexive comparisons are especially useful in assessing impact because they compare the “after” results with the “before” results.**

- True
- False
Qualitative Methods

Some evaluations rely largely on qualitative information, from focus groups, unstructured interviews, survey data on perceptions, and a variety of other sources. Such information complements, but does not supplant, the more quantitative impact evaluations, because qualitative methods are based on subjective evaluations, do not generate a control or comparison group, and lack statistical robustness.

Impact Evaluation: Macro Projects

It is much harder to evaluate the impact of an economy-wide shock (for example, a devaluation) or macroeconomic policy change (for example, increase in the money supply) than a project or program change, because the universal nature of the change makes it impossible to construct an appropriate counterfactual. Recognizing that the analysis will always be less than perfect, economists and others have nonetheless used the following methods to try to measure the effect of macro shocks:

Time-Series Data Analysis: Before and After

A time series is a set of data on a variable over time (for example, gross domestic product [GDP] for each of the past 20 years). One could compare the situation of households, using survey data, before and after the shock (that is, in time \( t-1 \) and time \( t \)). This is frequently done, but is quite imperfect because, as in reflexive comparisons, it does not establish an appropriate counterfactual. It implicitly assumes that if there had been no shock, the level of the variables in time \( t-1 \) would have persisted into time \( t \).

Time-Series Data Analysis: Deviations from Trend

An improvement over the simple before-and-after comparison is to begin by constructing a counterfactual, usually by predicting what would have happened in the absence of the crisis by projecting past trends into the future. The impact of the crisis is then calculated as the difference between the actual outcome after the crisis, and the predicted one based on the past trend. This is the approach taken by Kakwani (2000) in estimating the effects of the Asian financial crisis of 1997 on poverty and other indicators in Korea and Thailand.

The first difficulty with this method is arriving at a robust counterfactual; for instance, how far back in time should one go when developing an equation that is used for the projections. Second, it is much harder to establish a counterfactual for an
individual household than for the economy as a whole. And, third, this method does not control for the unobserved components of a household’s response to a shock.

**Computable General Equilibrium and Simulation Models**

A computable general equilibrium (CGE) model of an economy is a set of equations that aims to quantify the main interrelationships among households, firms, and government in an economy. CGE models range from just a few to many hundreds of equations. In principle, they may be used to simulate the effects of poverty reduction interventions. Unfortunately, CGE models are technically difficult to build, are typically highly aggregated (which makes it difficult to identify the effects of policies on income distribution and poverty with much precision), require considerable data to construct the underlying social accounting matrix, and produce results that are sensitive to the assumptions made about the parameters. They have been used, however, with some success to evaluate the economic and distributional effects of such interventions as programs to reduce HIV/AIDS, food subsidies, and trade policies. The International Food Policy Research Institute (IFPRI) has developed a standard CGE model that has been applied to a number of problems in developing countries (Loefgren, Harris, and Robinson 2001).

**Household Panel Impact Analysis**

If we have panel data on households—that is, data on households from both before and after the shock—then we can compare the situation of each household before and after the shock. By including household fixed effects in our estimating equation (that is, a separate dummy variable for each household), we can largely eliminate the effects of “time-invariant household and area-specific heterogeneity” (that is, the special or unique features of households, many of which are unobservable, such as whether the head is an alcoholic, or sick, or entrepreneurially inclined).

Again, the main difficulty here is that a before-and-after comparison does not establish an adequate counterfactual. For instance, if the income of a household in the Philippines fell between 1996 and 1998, how do we know that it was due to the 1997 financial crisis? It might have been caused by some other event—a family member falls ill, the village suffers from drought, and so on. No survey is ever complete enough to capture every conceivable relevant explanatory variable.

**Self-Rated Retrospective Evaluation**

Another possibility is to ask the household to assess how much it has been affected by the crisis—as was done, for instance, in the Annual Poverty Indicators Survey in the Philippines in 1998.
By definition, self-rated evaluations are subjective, which makes it difficult to assess whether the reported effects are indeed due to the shocks. In Vietnam, households reported higher levels of illness in 1998 than in 1993, despite being much better off in 1998. This result, which is not uncommon, is hardly plausible, unless one supposes that the definition of “illness” changes over time or with affluence. Whatever the reason, it makes the subjective evaluations untrustworthy.

A variant on this theme is to ask households whether they were hit by a shock. We then compare the situation of households that reported being affected with that of households that did not report being hit by the shock. Because self-reported shocks are highly endogenous—any household that has had a spell of bad luck is likely to report being hit by a shock—researchers often use the shock reported by a cluster (for example, the village or the city ward) as an instrumental variable to help resolve this endogeneity.

Even with this latter adjustment, we are left with the problem of unobserved community-level heterogeneity—for instance, for reasons that may not be apparent, some communities or clusters may report a shock more than others, even if objectively the shock hit all areas equally.

Three final points about impact evaluation are worth mentioning. First, no method of impact evaluation is perfect; the method used will depend on the problem, as well as on the resources and time available. Second, impact evaluation is more difficult with economy-wide policy interventions and crises than with micropolicies. And third, program evaluation is important; it serves as a tool for learning whether and how programs matter, and it has had an important effect on public policy in a number of cases (for some interesting examples, see Bamberger 2005). The usefulness of impact evaluation often requires the creation of adequate feedback mechanisms, however, so that policy makers take the lessons of impact evaluation to heart. The World Bank earmarks about 1 percent of project funds for monitoring and evaluation.

**Review Questions**

10. Real GDP rose by 3 percent in 2005, 2 percent in 2006, 4 percent in 2007, and fell 1 percent in 2008, apparently due to a financial crisis. Which of the following is the most plausible measure of the impact of the crisis on economic growth?

- A. It lowered GDP growth by 6 percentage points.
- B. It lowered GDP growth by 1 percentage point.
- C. It lowered GDP growth by 7 percentage points.
- D. It lowered GDP growth by 4 percentage points.
CHAPTER 13: Poverty Monitoring and Evaluation

Notes

2. Much of this section draws on lecture notes prepared by Shahid Khandker (2000).
3. Sometimes the results of a baseline survey are available, which can add precision to the results, particularly in ensuring that the treatment and control groups are indeed comparable.
4. Equation (13.2) as shown here is linear, but other forms, including logit and probit specifications, are typically used in practice.
5. A variant on this approach is to put the errors from equation (13.2), along with the actual participation rate, into equation (13.1) before estimating it.

References


