3. Randomization

Summary

Allocating a program or intervention randomly across a sample of observations is one solution to avoiding selection bias, provided that program impacts are examined at the level of randomization. Careful selection of control areas (or the counterfactual) is also important in ensuring comparability with participant areas and ultimately calculating the treatment effect (or difference in outcomes) between the two groups. The treatment effect can be distinguished as the *average treatment effect* (ATE) between participants and control units, or the *treatment effect on the treated* (TOT), a narrower measure that compares participant and control units, conditional on participants being in a treated area.

Randomization could be conducted purely randomly (where treated and control units have the same expected outcome in absence of the program); this method requires ensuring external and internal validity of the targeting design. In actuality, however, researchers have worked in partial randomization settings, where treatment and control samples are chosen randomly, conditional on some observable characteristics (for example, landholding or income). If these programs are exogenously placed, conditional on these observed characteristics, an unbiased program estimate can be made.

Despite the clarity of a randomized approach, a number of factors still need to be addressed in practice. They include resolving ethical issues in excluding areas that share similar characteristics with the targeted sample, accounting for spillovers to nontargeted areas as well as for selective attrition, and ensuring heterogeneity in participation and ultimate outcomes, even if the program is randomized.

Learning Objectives

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

- How to construct an appropriate counterfactual
- How to design a randomized experiment, including external and internal validity
- How to distinguish the ATE from the TOT
- How to address practical issues in evaluating randomized interventions, including accounting for spillovers, selective attrition, ethical issues, and selective heterogeneity in program impacts among the treated sample

Setting the Counterfactual

As argued in chapter 2, finding a proper counterfactual to treatment is the main challenge of impact evaluation. The counterfactual indicates what would have happened to participants of a program had they not participated. However, the same person cannot be observed in two distinct situations—being treated and untreated at the same time.

The main conundrum, therefore, is how researchers formulate counterfactual states of the world in practice. In some disciplines, such as medical science, evidence about counterfactuals is generated through randomized trials, which ensure that outcomes in the control group really do capture the counterfactual for a treatment group.

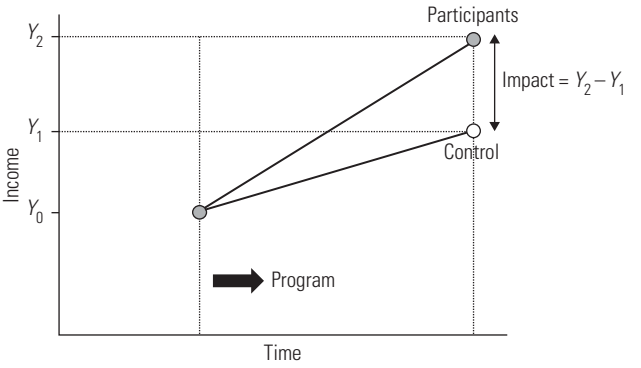
Figure 3.1 illustrates the case of randomization graphically. Consider a random distribution of two "similar" groups of households or individuals—one group is treated and the other group is not treated. They are similar or "equivalent" in that both groups prior to a project intervention are observed to have the same level of income (in this case, Y_0). After the treatment is carried out, the observed income of the treated group is found to be Y_2 while the income level of the control group is Y_1. Therefore, the effect of program intervention can be described as $(Y_2 - Y_1)$, as indicated in figure 3.1. As discussed in chapter 2, extreme care must be taken in selecting the control group to ensure comparability.

Statistical Design of Randomization

In practice, however, it can be very difficult to ensure that a control group is very similar to project areas, that the treatment effects observed in the sample are generalizable, and that the effects themselves are a function of only the program itself.

Statisticians have proposed a two-stage randomization approach outlining these priorities. In the first stage, a sample of potential participants is selected randomly

Figure 3.1 The Ideal Experiment with an Equivalent Control Group



Source: Authors' representation.

from the relevant population. This sample should be representative of the population, within a certain sampling error. This stage ensures *external validity* of the experiment. In the second stage, individuals in this sample are randomly assigned to treatment and comparison groups, ensuring *internal validity* in that subsequent changes in the outcomes measured are due to the program instead of other factors. Conditions to ensure external and internal validity of the randomized design are discussed further later.

Calculating Treatment Effects

Randomization can correct for the selection bias *B,* discussed in chapter 2, by randomly assigning individuals or groups to treatment and control groups. Returning to the setup in chapter 2, consider the classic problem of measuring treatment effects (see Imbens and Angrist 1994): let the treatment, T_i, be equal to 1 if subject i is treated and 0 if not. Let $Y_i(1)$ be the outcome under treatment and $Y_i(0)$ if there is no treatment.

Observe Y_i and T_i, where Yi = $[T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)]$.[1] Strictly speaking, the treatment effect for unit i is $Y_i(1) - Y_i(0)$, and the ATE is ATE = $E[Y_i(1) - Y_i(0)]$, or the difference in outcomes from being in a project relative to control area for a person or unit i randomly drawn from the population. This formulation assumes, for example, that everyone in the population has an equally likely chance of being targeted.

Generally, however, only $E[Y_i(1)|T_i = 1]$, the average outcomes of the treated, conditional on being in a treated area, and $E[Y_i(0)|T_i = 0]$, the average outcomes of the untreated, conditional on not being in a treated area, are observed. With nonrandom targeting and observations on only a subsample of the population, $E[Y_i(1)]$ is not necessarily equal to $E[Y_i(1)|T_i = 1]$, and $E[Y_i(0)]$ is not necessarily equal to $E[Y_i(0)|T_i = 0]$.

Typically, therefore, alternate treatment effects are observed in the form of the TOT: TOT = $E[Y_i(1) - Y_i(0)|T_i = 1]$, or the difference in outcomes from receiving the program as compared with being in a control area for a person or subject i randomly drawn from the treated sample. That is, the TOT reflects the average gains for participants, conditional on these participants receiving the program. Suppose the area of interest is the TOT, $E[Y_i(1) - Y_i(0)|T_i = 1]$. If T_i is nonrandom, a simple difference between treated and control areas, $D = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0]$ (refer to chapter 2), will not be equal to the TOT. The discrepancy between the TOT and this D will be $E[Y_i(0)|T_i = 1] - E[Y_i(0)|T_i = 0]$, which is equal to the bias B in estimating the treatment effect (chapter 2):

$$\text{TOT} = E[Y_i(1) - Y_i(0)|T_i = 1] \tag{3.1}$$

$$= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 1] \tag{3.2}$$

$$= D = E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] \quad \text{if } E[Y_i(0)|T_i = 0] = E[Y_i(0)|T_i = 1] \quad (3.3)$$

$$\Rightarrow \text{TOT} = D \quad \text{if } B = 0. \quad (3.4)$$

Although in principle the counterfactual outcome $E[Y_i(0)|T_i = 1]$ in equation 3.2 cannot be directly observed to understand the extent of the bias, still some intuition about it might exist. Duflo, Glennerster, and Kremer (2008), for example, discuss this problem in the context of a program that introduces textbooks in schools. Suppose one were interested in the effect of this program on students' learning, but the program was nonrandom in that schools that received textbooks were already placing a higher value on education. The targeted sample would then already have higher schooling achievement than the control areas, and $E[Y_i(0)|T_i = 1]$ would be greater than $E[Y_i(0)|T_i = 0]$, so that $B > 0$ and an upward bias exists in the program effect. If groups are randomly targeted, however, $E[Y_i(0)|T_i = 1]$ and $E[Y_i(0)|T_i = 0]$ are equal, and there is no selection bias in participation ($B = 0$).

In an effort to unify the literature on treatment effects, Heckman and Vytlacil (2005) also describe a parameter called the *marginal treatment effect* (MTE), from which the ATE and TOT can be derived. Introduced into the evaluation literature by Björklund and Moffitt (1987), the MTE is the average change in outcomes Y_i for individuals who are at the margin of participating in the program, given a set of observed characteristics X_i and conditioning on a set of unobserved characteristics U_i in the participation equation: MTE = $E(Y_i(1) - Y_i(0)|X_i = x, U_i = u)$. That is, the MTE is the average effect of the program for individuals who are just indifferent between participating and not participating. Chapter 6 discusses the MTE and its advantages in more detail.

Treatment Effect with Pure Randomization

Randomization can be set up in two ways: pure randomization and partial randomization. If treatment were conducted purely randomly following the two-stage procedure outlined previously, then treated and untreated households would have the same expected outcome in the absence of the program. Then, $E[Y_i(0)|T_i = 1]$ is equal to $E[Y_i(0)|T_i = 0]$. Because treatment would be random, and not a function of unobserved characteristics (such as personality or other tastes) across individuals, outcomes would not be expected to have varied for the two groups had the intervention not existed. Thus, selection bias becomes zero under the case of randomization.

Consider the case of pure randomization, where a sample of individuals or households is randomly drawn from the population of interest. The experimental sample is then divided randomly into two groups: (a) the treatment group that is exposed to the program intervention and (b) the control group that does not receive the program. In terms of a regression, this exercise can be expressed as

$$Y_i = \alpha + \beta T_i + \varepsilon_i, \tag{3.5}$$

where T_i is the treatment dummy equal to 1 if unit i is randomly treated and 0 otherwise. As above, Y_i is defined as

$$Y_i \equiv [Y_i(1) \cdot T_i] + [Y_i(0) \cdot (1 - T_i)]. \tag{3.6}$$

If treatment is random (then T and ε are independent), equation 3.5 can be estimated by using ordinary least squares (OLS), and the treatment effect $\hat{\beta}_{OLS}$ estimates the difference in the outcomes of the treated and the control group. If a randomized evaluation is correctly designed and implemented, an unbiased estimate of the impact of a program can be found.

Treatment Effect with Partial Randomization

A pure randomization is, however, extremely rare to undertake. Rather, *partial randomization* is used, where the treatment and control samples are chosen randomly, conditional on some observable characteristics X (for example, landholding or income). If one can make an assumption called *conditional exogeneity of program placement,* one can find an unbiased estimate of program estimate.

Here, this model follows Ravallion (2008). Denoting for simplicity $Y_i(1)$ as Y_i^T and $Y_i(0)$ as Y_i^C, equation 3.5 could be applied to a subsample of participants and nonparticipants as follows:

$$Y_i^T = \alpha^T + X_i\beta^T + \mu_i^T \quad \text{if } T_i = 1, i = 1,..., n \tag{3.7}$$

$$Y_i^C = \alpha^C + X_i\beta^C + \mu_i^C \quad \text{if } T_i = 0,\ i = 1,..., n. \tag{3.8}$$

It is common practice to estimate the above as a single regression by pooling the data for both control and treatment groups. One can multiply equation 3.7 by T_i and multiply equation 3.8 by $(1 - T_i)$, and use the identity in equation 3.6 to get

$$Y_i = \alpha^C + (\alpha^T - \alpha^C) T_i + X_i\beta^C + X_i(\beta^T - \beta^C) T_i + \varepsilon_i, \tag{3.9}$$

where $\varepsilon_i = T_i(\mu_i^T - \mu_i^C) + \mu_i^C$. The treatment effect from equation 3.9 can be written as $A^{TT} = E(Y_i | T_i = 1, X) = E[\alpha^T - \alpha^C + X_i(\beta^T - \beta^C)]$. Here, A^{TT} is just the treatment effect on the treated, TOT, discussed earlier.

For equation 3.9, one can get a consistent estimate of the program effect with OLS if one can assume $E(\mu_i^T \mid X, T = t) = E(\mu_i^C \mid X, T = t) = 0$, $t = \{0,1\}$. That is, there is no selection bias because of randomization. In practice, a common-impact model is often used that assumes $\beta^T = \beta^C$. The ATE is then simply $\alpha^T - \alpha^C$.

Randomization in Evaluation Design: Different Methods of Randomization

If randomization were possible, a decision would have to be made about what type of randomization (oversubscription, randomized phase-in, within-group randomization, or encouragement design) would be used. These approaches, detailed in Duflo, Glennerster, and Kremer (2008), are discussed in turn below:

- *Oversubscription.* If limited resources burden the program, implementation can be allocated randomly across a subset of eligible participants, and the remaining eligible subjects who do not receive the program can be considered controls. Some examination should be made of the budget, assessing how many subjects could be surveyed versus those actually targeted, to draw a large enough control group for the sample of potential beneficiaries.
- *Randomized phase-in.* This approach gradually phases in the program across a set of eligible areas, so that controls represent eligible areas still waiting to receive the program. This method helps alleviate equity issues and increases the likelihood that program and control areas are similar in observed characteristics.
- *Within-group randomization.* In a randomized phase-in approach, however, if the lag between program genesis and actual receipt of benefits is large, greater controversy may arise about which area or areas should receive the program first. In that case, an element of randomization can still be introduced by providing the program to some subgroups in each targeted area. This approach is therefore similar to phased-in randomization on a smaller scale. One problem is that spillovers may be more likely in this context.
- *Encouragement design.* Instead of randomizing the treatment, researchers randomly assign subjects an announcement or incentive to partake in the program. Some notice of the program is given in advance (either during the time of the baseline to conserve resources or generally before the program is implemented) to a random subset of eligible beneficiaries. This notice can be used as an instrument for take-up in the program. Spillovers might also be measured nicely in this context, if data are also collected on the social networks of households that receive the notice, to see how take-up might differ across households that are connected or not connected to it. Such an experiment would require more intensive data collection, however.

Concerns with Randomization

Several concerns warrant consideration with a randomization design, including ethical issues, external validity, partial or lack of compliance, selective attrition, and spillovers. Withholding a particular treatment from a random group of people and providing

access to another random group of people may be simply unethical. Carrying out randomized design is often politically unfeasible because justifying such a design to people who might benefit from it is hard. Consequently, convincing potential partners to carry out randomized designs is difficult.

External validity is another concern. A project of small-scale job training may not affect overall wage rates, whereas a large-scale project might. That is, impact measured by the pilot project may not be an accurate guide of the project's impact on a national scale. The problem is how to generalize and replicate the results obtained through randomized evaluations.

Compliance may also be a problem with randomization, which arises when a fraction of the individuals who are offered the treatment do not take it. Conversely, some members of the comparison group may receive the treatment. This situation is referred to as partial (or imperfect) compliance. To be valid and to prevent selection bias, an analysis needs to focus on groups created by the initial randomization. The analysis cannot exclude subjects or cut the sample according to behavior that may have been affected by the random assignment. More generally, interest often lies in the effect of a given treatment, but the randomization affects only the *probability* that the individual is exposed to the treatment, rather than the treatment itself.

Also, potential spillover effects arise when treatment helps the control group as well as the sample participants, thereby confounding the estimates of program impact. For example, people outside the sample may move into a village where health clinics have been randomly established, thus contaminating program effects. The chapter now examines how such concerns about randomization have actually been addressed in practice.

Randomized Impact Evaluation in Practice

Randomization has been growing in popularity in some parts of the world, in part because if it can be implemented properly, randomization can give a robust indication of program impact. Also, once the survey has been designed and the data collected, the empirical exercises to infer impacts from randomized experiments are quite straightforward. Typically, justifying or initiating a randomized experiment is easiest at the inception of a program, during the pilot phase. This phase offers a natural opportunity to introduce randomization before the program is scaled up. It presents an occasion for the implementation partner to rigorously assess the effectiveness of the program. It can also provide a chance to improve the program's design. One can also introduce an element of randomization into existing programs in many different ways with minimal disruption. Whereas the earlier sections in this chapter have discussed in theory the concerns with randomization, the following sections discuss various practical issues and case studies in the implementation of randomized studies.

Ethical Issues

Implementing randomized experiments in developing countries often raises ethical issues. For example, convincing government officials to withhold a particular program from a randomly selected contingent that shares the same poverty status and limits on earning opportunities as a randomly targeted group may be difficult. Carrying out randomized designs is often politically unfeasible because of the difficulty in justifying such a design to people who might benefit from it.

One counterargument is that randomization is a scientific way of determining the program's impact. It would therefore ultimately help decide, among a set of different programs or paths available to policy makers, which ones really work and hence deserve investment. Thus, in the long run, randomization can help a greater number of people in addition to those who were initially targeted. A randomly phased-in design such as that used by Mexico's PROGRESA (Programa de Educación, Salud y Alimentación, or Education, Health, and Nutrition Program; see box 3.1) can also allow nontargeted, similarly featured areas ultimately to benefit from the program as well as provide a good comparison sample.

BOX 3.1 Case Study: PROGRESA (Oportunidades)

PROGRESA (now called Oportunidades), described in box 2.1 of chapter 2, combined regional and village-level targeting with household-level targeting within these areas. Only the extreme poor were targeted, using a randomized targeting strategy that phased in the program over time across targeted localities. One-third of the randomly targeted eligible communities were delayed entry into the program by 18 months, and the remaining two-thirds received the program at inception. Within localities, households were chosen on the basis of a discriminant analysis that used their socioeconomic characteristics (obtained from household census data) to classify households as poor or nonpoor. On average, about 78 percent of households in selected localities were considered eligible, and about 93 percent of households that were eligible enrolled in the program.

Regarding potential ethical considerations in targeting the program randomly, the phased-in treatment approach allowed all eligible samples to be targeted eventually, as well as the flexibility to adjust the program if actual implementation was more difficult than initially expected. Monitoring and operational evaluation of the program, as discussed in chapter 2, were also key components of the initiative, as was a detailed cost-benefit analysis.

A number of different evaluations have examined the impact of Oportunidades on health and educational outcomes among the treated sample. They include examinations of the program's benefits to health (Gertler 2004); labor-market outcomes for adults and youth (Behrman, Parker, and Todd 2009; Skoufias and di Maro 2007); schooling (de Janvry and others 2006; Schultz 2004; Todd and Wolpin 2006); and nutrition (Behrman and Hoddinott 2005; Hoddinott and Skoufias 2004). Interest in the design and outcomes of Oportunidades has fostered similar conditional cash-transfer programs in South America and Central America, as well as in Bangladesh and Turkey.

Also, in the presence of limited resources, not all people can be targeted by a program—whether experimental or nonexperimental. In that case, randomized targeting is not unethical. The bottom line is that, in practice, convincing potential partners to carry out randomized designs is often difficult; thus, the first challenge is to find suitable partners to carry out such a design. Governments, nongovernmental organizations, and sometimes private sector firms might be potential partners.

Internal versus External Validity

Different approaches in implementing randomized studies reflect the need to adapt the program intervention and survey appropriately within the targeted sample. These concerns are embedded in a broader two-stage process guiding the quality of experimental design. In the first stage, policy makers should define clearly not only the random sample that will be selected for analysis but also the population from which that sample will be drawn. Specifically, the experiment would have external validity, meaning that the results obtained could be generalized to other groups or settings (perhaps through other program interventions, for example). Using the notation discussed earlier, this approach would correspond to the conditions $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$ and $E[Y_i(1)|T_i = 1] = E[Y_i(1)|T_i = 0]$.

Second, steps should be taken when randomly allocating this sample across treatment and control conditions to ensure that the treatment effect is a function of the intervention only and not caused by other confounding elements. This criterion is known as *internal validity* and reflects the ability to control for issues that would affect the causal interpretation of the treatment impact. Systematic bias (associated with selection of groups that are not equivalent, selective sample attrition, contamination of targeted areas by the control sample, and changes in the instruments used to measure progress and outcomes over the course of the experiment), as well as the effect of targeting itself on related choices and outcomes of participants within the targeted sample, provides an example of such issues. Random variation in other events occurring while the experiment is in progress, although not posing a direct threat to internal validity, also needs to be monitored within data collection because very large random variation can pose a threat to the predictability of data measurement. The following section discusses some approaches that, along with a randomized methodology, can help account for these potentially confounding factors.

Although following the two-stage approach will lead to a consistent measure of the ATE (Kish 1987), researchers in the behavioral and social sciences have almost never implemented this approach in practice. More specifically, the only assumption that can be made, given randomization, is that $E[Y_i(0)|T_i = 1] = E[Y_i(0)|T_i = 0]$. Even maintaining the criterion for internal validity in an economic setting is very difficult, as will be described. At best, therefore, policy makers examining the effect of randomized program interventions can consistently estimate the TOT or effect on a given subpopulation: TOT = $E[Y_i(1) - Y_i(0)|T_i = 1]$, as opposed to ATE = $E[Y_i(1) - Y_i(0)]$.

Intent-to-Treat Estimates and Measuring Spillovers

Ensuring that control areas and treatment areas do not mix is crucial in measuring an unbiased program impact. In the experimental design, a number of approaches can help reduce the likelihood of contamination of project areas. Project and control areas that are located sufficiently far apart, for example, can be selected so that migration across the two areas is unlikely. As a result, contamination of treatment areas is more likely with projects conducted on a larger scale.

Despite efforts to randomize the program intervention ex ante, however, actual program participation may not be entirely random. Individuals or households in control areas may move to project areas, ultimately affecting their outcomes from exposure to the program. Likewise, targeted individuals in project areas may not ultimately participate but may be indirectly affected by the program as well. If a program to target the treated helps the control group too, it would confound the estimates of program impact. In some cases, projects cannot be scaled up without creating general equilibrium effects. For example, a project of small-scale job training may not affect overall wage rates, whereas a large-scale project might. In the latter case, impact measured by the pilot project would be an inaccurate guide of the project's impact on a national scale. Often the Hawthorne effect might plague results of a randomized experiment, where the simple fact of being included in an experiment may alter behavior nonrandomly.[2]

These partial treatment effects may be of separate interest to the researcher, particularly because they are likely to be significant if the policy will be implemented on a large scale. They can be addressed through measuring *intention-to-treat* (ITT) impacts (box 3.2) or by instrumenting actual program participation by the randomized assignment strategy (box 3.3).

Specifically, in cases where the actual treatment is distinct from the variable that is randomly manipulated, call Z the variable that is randomly assigned (for example, the letter inviting university employees to a fair and offering them US$20 to attend), while T remains the treatment of interest (for example, attending the fair). Using the same notation as previously, one knows because of random assignment that $E[Y_i(0)|Z_i = 1] - E[Y_i(0)|Z_i = 0]$ is equal to zero and that the difference $E[Y_i(1)|Z_i = 1] - E[Y_i(0)|Z_i = 0]$ is equal to the causal effect of Z. However, it is not equal to the effect of the treatment, T, because Z is not equal to T. Because Z has been chosen to at least influence the treatment, this difference is the ITT impact.

Because the ITT is in principle random, it can also act as a valid instrumental variable to identify the treatment impact, given that people who were initially assigned for treatment are in general more likely to have ultimately participated in the program. The ITT estimate would then be the estimated coefficient on the variable describing initial

BOX 3.2 Case Study: Using Lotteries to Measure Intent-to-Treat Impact

The PACES (Plan de Ampliación de Cobertura de la Educación Secundaria, or Plan for Increasing Secondary Education Coverage) school voucher program, established by the Colombian government in late 1991, granted private secondary school vouchers to 125,000 children from poor neighborhoods who were enrolled in public primary schools. These vouchers covered about half of entering students' schooling expenses and were renewable depending on student performance. However, the program faced oversubscription because the number of eligible households (living in neighborhoods falling in the lowest two of six socioeconomic strata spanning the population) exceeded the number of available vouchers. Many vouchers were therefore allocated through a randomized lottery.

To measure the impact of this school voucher program, Angrist and others (2002) surveyed lottery winners and losers from three groups of applicants. They administered an academic test to both groups, initially finding limited differences in performance for voucher recipients. One reason for this outcome, they suggest, is that about 10 percent of lottery winners did not end up using the voucher or other scholarship, whereas about 25 percent of nonrecipients obtained other scholarships or funding. Angrist and others (2002) therefore used the lottery receipt as an instrument for participation, calculating an intention-to-treat estimate that revealed much larger (50 percent greater) program effects on grade completion and reduced repetitions for lottery winners than in a simple comparison of winners and losers.

assignment. The impact on those whose treatment status is changed by the instrument is also known as the *local average treatment effect* (Abadie, Angrist, and Imbens 2002).

Selective attrition is also a potential problem: people drop out of a program. Box 3.4 describes an example from a schooling program in India, where potential attrition of weaker students could bias the program effect upward.

If measuring the extent of spillovers is of interest to policy makers, randomization can allow this phenomenon to be measured more precisely. The accuracy, of course, depends on the level of spillovers. If spillovers occur at the aggregate or global economy, for example, any methodology—be it randomization or a nonexperimental approach—will have difficulties in capturing the program impact. Local spillovers can, however, be measured with a randomized methodology (Miguel and Kremer 2004; see box 3.5).

Selecting the level of randomization on the basis of the level at which spillovers are expected to occur (that is, whether over individuals, communities, or larger units) is therefore crucial in understanding the program impact. A substantive amount of data measuring factors that might lead to contamination and spillovers (migration, for example) would also need to be examined during the course of the evaluation to be able to estimate the program's impact precisely.

BOX 3.3 Case Study: Instrumenting in the Case of Partial Compliance

Abadie, Angrist, and Imbens (2002) discussed an approach that introduces instrumental variables to estimate the impact of a program that is randomized in intent but for which actual take-up is voluntary. The program they examined involves training under the U.S. Job Training Partnership Act of 1982. Applicants were randomly assigned to treatment and control groups; those in the treated sample were immediately offered training, whereas training programs for the control sample were delayed by 18 months. Only 60 percent of the treated sample actually received training, and the random treatment assignment was used as an instrumental variable.

The study examined a sample of about 6,100 women and 5,100 men, with earnings data for each individual spanning 30 months. Using the instrumental variables estimates, Abadie, Angrist, and Imbens found that the average rise in earnings for men was about US$1,600 (a 9 percent increase), about half as large as the OLS estimate. For women, the average increase was about US$1,800 (growth of about 15 percent) and was not very different from the corresponding OLS estimate.

BOX 3.4 Case Study: Minimizing Statistical Bias Resulting from Selective Attrition

Banerjee and others (2007) examined the impact of two randomized educational programs (a remedial education program and computer-assisted learning) across a sample of urban schools in India. These programs were targeted toward students who, relative to students in other schools, were not performing well in basic literacy and other skills. Government primary schools were targeted in two urban areas, with 98 schools in the first area (Vadodara) and 77 schools in the second area (Mumbai).

With respect to the remedial program in particular, half the schools in each area sample were randomly selected to have the remedial program introduced in grade 3, and the other half received the program in grade 4. Each treated group of students was therefore compared with untreated students from the same grade within the same urban area sample. Tests were administered to treated and untreated students to evaluate their performance.

In the process of administering the program, however, program officials found that students were dropping out of school. If attrition was systematically greater among students with weaker performance, the program impact would suffer from an upward bias. As a result, the testing team took efforts to visit students in all schools across the sample multiple times, tracking down children who dropped out of school to have them take the test. Although the attrition rate among students remained relatively high, it was ultimately similar across the treated and untreated samples, thereby lowering the chance of bias in direct comparisons of test scores across the two groups.

Ultimately, Banerjee and others (2007) found that the remedial education program raised average test scores of all children in treatment schools by 0.14 standard deviations in the first year and 0.28 standard deviations in the second year, driven primarily from improvements at the lower end of the distribution of test scores (whose gains were about 0.40 standard deviations relative to the control group sample).

BOX 3.5 Case Study: Selecting the Level of Randomization to Account for Spillovers

Miguel and Kremer (2004) provided an evaluation of a deworming program across a sample of 75 schools in western Kenya, accounting for treatment externalities that would have otherwise masked the program impact. The program, called the Primary School Deworming Project, involved randomized phase-in of the health intervention at the school level over the years 1998 to 2000.

Examining the impact at the individual (child) level might be of interest, because children were ultimately recipients of the intervention. However, Miguel and Kremer (2004) found that since infections spread easily across children, strong treatment externalities existed across children randomly treated as part of the program and children in the comparison group. Not accounting for such externalities would therefore bias the program impact, and randomizing the program within schools was thus not possible.

Miguel and Kremer (2004) therefore examined impacts at the school level, because the deworming program was randomized across schools, and treatment and comparison schools were located sufficiently far apart that the likelihood of spillovers across schools was much smaller. They measured the size of the externality by comparing untreated students in treated schools with the comparison group. Their study found that treated schools exhibited significantly (about 25 percent) lower absenteeism rates, although academic test scores did not improve relative to comparison schools. Their analysis also found substantial treatment externalities, in that untreated children in treatment schools exhibited significantly improved health and school participation rates compared with children in nontreated schools. Including the externality benefits, Miguel and Kremer found the cost per additional year of school participation was just US$3.50, making deworming more cost-effective than subsidies in reducing absenteeism.

Heterogeneity in Impacts: Estimating Treatment Impacts in the Treated Sample

The level at which the randomized intervention occurs (for example, the national, regional, or community level) therefore affects in multiple ways the treatment effects that can be estimated. Randomization at an aggregate (say, regional) level cannot necessarily account for individual heterogeneity in participation and outcomes resulting from the program.

One implication of this issue is that the ultimate program or treatment impact at the individual level cannot necessarily be measured accurately as a binary variable (that is, $T = 1$ for an individual participant and $T = 0$ for an individual in a control area). Although a certain program may be randomized at a broader level, individual selection may still exist in the response to treatment. A mixture of methods can be used, including instrumental variables, to account for unobserved selection at the individual level. Interactions between the targeting criteria and the treatment indicator can also be introduced in the regression.

Quantile treatment effects can also be estimated to measure distributional impacts of randomized programs on outcomes such as per capita consumption and expenditure (Abadie, Angrist, and Imbens 2002). Chapter 8 discusses this approach in more detail. Dammert (2007), for example, estimates the distributional impacts on expenditures from a conditional cash-transfer program in rural Nicaragua. This program, Red de Protección Social (or Social Protection Network), was a conditional cash-transfer program created in 2000. It was similar to PROGRESA in that eligible households received cash transfers contingent on a few conditions, including that adult household members (often mothers) attended educational workshops and sent their children under 5 years of age for vaccinations and other health appointments and sent their children between the ages of 7 and 13 regularly to school. Some aspects of the evaluation are discussed in box 3.6. Djebbari and Smith (2008) also provide a similar discussion using data from PROGRESA (Oportunidades).

BOX 3.6 Case Study: Measuring Impact Heterogeneity from a Randomized Program

Dammert (2007) examined distributional impacts of the Nicaraguan social safety net program Red de Protección Social, where 50 percent of 42 localities identified as sufficiently poor for the program (according to a marginality index) were randomly selected for targeting. The evaluation survey covered 1,359 project and control households through a baseline, as well as two follow-up surveys conducted one year and two years after program intervention.

Because the cash transfers depended on regular school attendance and health visits, however, whether a household in a targeted locality was already meeting these requirements before the intervention (which correlated heavily with the household's preexisting income and education levels) could result in varying program impacts across households with different socioeconomic backgrounds. For households whose children were already enrolled in school and sent regularly for health checkups, the cash transfer would provide a pure income effect, whereas for households not meeting the criteria, the cash transfer would induce both an income and substitution effect.

As one approach, Dammert (2007) therefore interacted the program variable with household characteristics on which targeting was based, such as education of the household head, household expenditures, and the marginality index used for targeting. Children in poorer localities were found to have greater improvements in schooling, for example. Also, to examine variation in program impacts not driven by observable characteristics, Dammert calculated quantile treatment effects separately for 2001 and 2002. The results show that growth in total per capita expenditures as well as per capita food expenses was lower for households at the bottom of the expenditure distribution. Specifically, in 2001, the program's impact on increased total per capita expenditures ranged from US$54 to US$237; in 2002, this range was US$20 to US$99, with households at the top of the distribution receiving more than five times the impact than households with lower expenditures.

Thus, simply relying on average treatment impacts may not reveal important areas of concern, such as, perhaps, that households at the lower end of the expenditure distribution experience higher costs (and thus reduced benefits) from participating.

A related departure from perfect randomization is when randomization is a function of some set of observables (climate, population density, and the like) affecting the probabilities that certain areas will be selected. Treatment status is therefore randomly conditioned on a set of observed characteristics. Within each treated area, however, treatment is randomized across individuals or communities. Treatment and comparison observations within each area can therefore be made, and a weighted average can be taken over all areas to give the average effect of the program on the treated samples.

Value of a Baseline Study

Conducting baseline surveys in a randomized setting conveys several advantages. First, baseline surveys make it possible to examine interactions between initial conditions and the impact of the program. In many cases, this comparison will be of considerable importance for assessing external validity. Baseline data are also useful when conducting policy experiments, because treated areas might have had access to similar programs or initiatives before implementation of the new initiative. Comparing participants' uptake of activities, such as credit before and after the randomized intervention, can also be useful in evaluating responses to the experiment.

Other values of a baseline study include the opportunity to check that the randomization was conducted appropriately. Governments participating in randomized schemes may feel the need, for example, to compensate control areas for not receiving the program by introducing other schemes at the same time. Data collected on program interventions in control areas before and during the course of the survey will help in accounting for these additional sources of spillovers. Collecting baseline data also offers an opportunity to test and refine data collection procedures.

Baseline surveys can be costly, however, and should be conducted carefully. One issue with conducting a baseline is that it may lead to bias in program impacts by altering the counterfactual. The decision whether to conduct a baseline survey boils down to comparing the cost of the intervention, the cost of data collection, and the impact that variables for which data can be collected in a baseline survey may have on the final outcome (box 3.7).

Difficulties with Randomization

Because they minimize selection bias in program impacts, randomized evaluations can be very attractive in developing countries. Unfortunately, contextual factors in such settings are rife with situations that can confound randomized implementation and hence the quality of program effects. Detailed data collection on these confounding factors and use of a combination of methods, in addition to examining the ATEs, can therefore help in accounting for resulting individual heterogeneity in treatment impacts (box 3.8).

BOX 3.7 Case Study: Effects of Conducting a Baseline

Giné, Karlan, and Zinman (2008), in a study of a rural hospitalization insurance program offered by the Green Bank in the Philippines, examined the impact of conducting a randomly allocated baseline on a subset of individuals to whom the program was ultimately offered. The baseline (which surveyed a random sample of 80 percent of the roughly 2,000 individual liability borrowers of the Green Bank) elicited indicators such as income, health status, and risky behavior. To avoid revealing information about the upcoming insurance program, the baseline did not cover questions about purchases of insurance, and no connection was discussed between the survey and the bank. However, after the insurance initiative was introduced, take-up was found to be significantly higher (about 3.4 percentage points) among those surveyed than those who were not.

The study therefore points to the benefits of capturing characteristics of surveyed individuals in the baseline that might reveal potential behavioral patterns in subsequent decision making, including their influence in decision making over such issues before program implementation. Randomized variation in the timing of program implementation after the baseline might also be used to test how these effects persist over time.

BOX 3.8 Case Study: Persistence of Unobserved Heterogeneity in a Randomized Program

Behrman and Hoddinott (2005) examined nutritional effects on children from PROGRESA, which also involved the distribution of food supplements to children. Although the program was randomized across localities, a shortage in one nutritional supplement provided to preschool children led local administrators to exercise discretion in how they allocated this supplement, favoring children with poorer nutritional status. As a result, when average outcomes between treatment and control groups were compared, the effect of the program diminished. Behrman and Hoddinott examined a sample of about 320 children in project and control households (for a total sample of about 640). Introducing child-specific fixed-effects regressions revealed a positive program impact on health outcomes for children; height of recipient children increased by about 1.2 percent. Behrman and Hoddinott predicted that this effect alone could potentially increase lifetime earnings for these children by about 3 percent. The fixed-effects estimates controlled for unobserved heterogeneity that were also correlated with access to the nutritional supplement.

Even in the context of industrial countries, Moffitt (2003) discusses how randomized field trials of cash welfare programs in the United States have had limited external validity in terms of being able to shed light on how similar policies might play out at the national level. Although nonexperimental studies also face similar issues with external validity, Moffitt argues for a comprehensive approach comparing experimental with nonexperimental studies of policies and programs; such comparisons may reveal potential mechanisms affecting participation, outcomes, and other participant

behavior, thereby helping evaluators understand potential implications of such programs when applied to different contexts.

In the nonexperimental studies discussed in the following chapters, this book attempts to account for the selection bias issue in different ways. Basically, nonexperimental studies try to replicate a natural experiment or randomization as much as possible. Unlike randomization, where selection bias can be corrected for directly (although problems exist in this area also), in nonexperimental evaluations a different approach is needed, usually involving assumptions about the form of the bias.

One approach is to make the case for assuming unconfoundedness—or of conditional exogeneity of program placement, which is a weaker version of unconfoundedness. The propensity score matching technique and double-difference methods fall under this category. The instrumental variable approach does not need to make this assumption. It attempts to find instruments that are correlated with the participation decision but not correlated with the outcome variable conditional on participation. Finally, other methods, such as regression discontinuity design (also an instrumental variable method), exploit features of program design to assess impact.