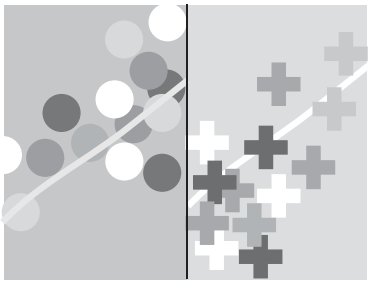
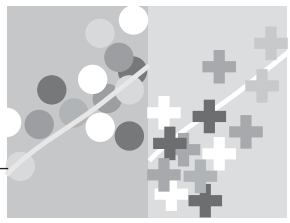
Impact Evaluation in Practice



Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
Christel M. J. Vermeersch

THE WORLD BANK



CHAPTER 3

Causal Inference and Counterfactuals

We begin by examining two concepts that are integral to the process of conducting accurate and reliable evaluations—causal inference and counterfactuals.

Causal Inference

The basic impact evaluation question essentially constitutes a *causal inference* problem. Assessing the impact of a program on a series of outcomes is equivalent to assessing the causal effect of the program on those outcomes. Most policy questions involve cause-and-effect relationships: Does teacher training *improve* students' test scores? Do conditional cash transfer programs *cause* better health outcomes in children? Do vocational training programs *increase* trainees' incomes?

Although cause-and-effect questions are common, it is not a straightforward matter to establish that a relationship is causal. In the context of a vocational training program, for example, simply observing that a trainee's income increases after he or she has completed such a program is not sufficient to establish causality. The trainee's income might have increased even if he had not taken the training course because of his own efforts, because of changing labor market conditions, or because of one of the myriad other factors that can affect income. Impact evaluations help us to overcome the

challenge of establishing causality by empirically establishing to what extent a particular program—*and that program alone*—contributed to the change in an outcome. To establish causality between a program and an outcome, we use impact evaluation methods to rule out the possibility that any factors other than the program of interest explain the observed impact.

The answer to the basic impact evaluation question—*What is the impact or causal effect of a program P on an outcome of interest Y?*—is given by the basic impact evaluation formula:

$$\alpha = (Y \mid P = 1) - (Y \mid P = 0).$$

This formula says that the causal impact (α) of a program (P) on an outcome (Y) is the difference between the outcome (Y) with the program (in other words, when $P = 1$) and the same outcome (Y) without the program (that is, when $P = 0$).

For example, if P denotes a vocational training program and Y denotes income, then the causal impact of the vocational training program (α) is the difference between a person's income (Y) after participating in the vocational training program (in other words, when $P = 1$) and the same person's income (Y) at the same point in time if he or she had not participated in the program (in other words, when $P = 0$). To put it another way, we would like to measure income at the same point in time for the same unit of observation (a person, in this case), but in two different states of the world. If it were possible to do this, we would be observing how much income the same individual would have had at the same point in time both with and without the program, so that the *only* possible explanation for any difference in that person's income would be the program. By comparing the same individual with herself at the same moment, we would have managed to eliminate any outside factors that might also have explained the difference in outcomes. We could then be confident that the relationship between the vocational training program and income is causal.

The basic impact evaluation formula is valid for anything that is being analyzed—a person, a household, a community, a business, a school, a hospital, or any other unit of observation that may receive or be affected by a program. The formula is also valid for any outcome (Y) that is plausibly related to the program at hand. Once we measure the two key components of this formula—the outcome (Y) both with the program and without it—we can answer any question about the program's impact.

The Counterfactual

As discussed above, we can think of the impact (α) of a program as the difference in outcomes (Y) for the same individual with and without partici-

pation in a program. Yet we know that measuring the same person in two different states at the same time is impossible. At any given moment in time, an individual either participated in the program or did not participate. The person cannot be observed simultaneously in two different states (in other words, with and without the program). This is called "the counterfactual problem": How do we measure what would have happened if the other circumstance had prevailed? Although we can observe and measure the outcome (Y) for program participants ($Y \mid P = 1$), there are no data to establish what their outcomes would have been in the absence of the program ($Y \mid P = 0$). In the basic impact evaluation formula, the term ($Y \mid P = 0$) *represents the counterfactual*. We can think of this as *what would have happened* if a participant had not participated in the program. In other words, the counterfactual is what the outcome (Y) would have been in the absence of a program (P).

Key Concept:
The counterfactual is an estimate of what the outcome (Y) would have been for a program participant in the absence of the program (P).

For example, imagine that "Mr. Unfortunate" takes a red pill and then dies five days later. Just because Mr. Unfortunate died after taking the red pill, you cannot conclude that the red pill *caused* his death. Maybe he was very sick when he took the red pill, and it was the illness rather than the red pill that caused his death. Inferring causality will require that you rule out other potential factors that can affect the outcome under consideration. In the simple example of determining whether taking the red pill caused Mr. Unfortunate's death, an evaluator would need to establish what would have happened to Mr. Unfortunate had he *not* taken the pill. Inasmuch as Mr. Unfortunate did in fact take the red pill, it is not possible to observe directly what would have happened if he had not done so. What would have happened to him had he not taken the red pill is the counterfactual, and the evaluator's main challenge is determining what this counterfactual state of the world actually looks like (see box 3.1).

When conducting an impact evaluation, it is relatively easy to obtain the first term of the basic formula ($Y \mid P = 1$)—the outcome under treatment. We simply measure the outcome of interest for the population that participated in the program. However, the second term of the formula ($Y \mid P = 0$) cannot be directly observed for program participants—hence, the need to fill in this missing piece of information by *estimating the counterfactual*. To do this, we typically use *comparison groups* (sometimes called "control groups"). The remainder of part 2 of this book will focus on the different methods or approaches that can be used to identify valid comparison groups that accurately reproduce or mimic the counterfactual. Identifying such comparison groups is the crux of any impact evaluation, regardless of what type of program is being evaluated. Simply put, without a valid estimate of the counterfactual, the impact of a program cannot be established.

Box 3.1: Estimating the Counterfactual
Miss Unique and the Cash Transfer Program

Miss Unique is a newborn baby girl whose mother is offered a monthly cash transfer so long as she ensures that Miss Unique receives regular health checkups at the local health center, that she is immunized, and that her growth is monitored. The government posits that the cash transfer will motivate Miss Unique's mother to seek the health services required by the program and will help Miss Unique grow strong and tall. For its impact evaluation, the government selects height as an outcome indicator for long-term health, and it measures Miss Unique's height 3 years into the cash transfer program.

Assume that you are able to measure Miss Unique's height at the age of 3. Ideally, to evaluate the impact of the program, you would want to measure Miss Unique's height at the age of 3 with her mother having received the cash transfer, and also Miss Unique's height at the age of 3 had her mother not received the cash transfer. You would then compare the two heights. If you were able to compare Miss Unique's height at the age of 3 with the program to Miss Unique's height at the age of 3 without the program, you would know that any difference in height had been caused only by the program. Because everything else about Miss Unique would be the same, there would be no other characteristics that could explain the difference in height.

Unfortunately, however, it is impossible to observe Miss Unique both with and without the cash transfer program: either her family receives the program or it does not. In other words, we do not know what the counterfactual is. Since Miss Unique's mother actually received the cash transfer program, we cannot know how tall she would have been had her mother not received the cash transfer. Finding an appropriate comparison for Miss Unique will be challenging because Miss Unique is, precisely, unique. Her exact socioeconomic background, genetic attributes, and personal characteristics cannot be found in anybody else. If we were simply to compare Miss Unique with a child who is not enrolled in the cash transfer program, say, Mr. Inimitable, the comparison may not be adequate. Miss Unique is not identical to Mr. Inimitable. Miss Unique and Mr. Inimitable may not look the same, they may not live in the same place, they may not have the same parents, and they may not have been the same height when they were born. So if we observe that Mr. Inimitable is shorter than Miss Unique at the age of 3, we cannot know whether the difference is due to the cash transfer program or to one of the many other differences between these two children.

Estimating the Counterfactual

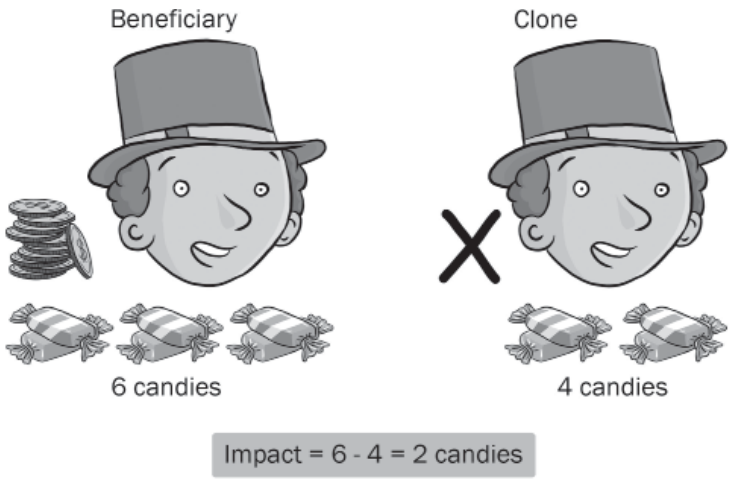
To further illustrate the estimation of the counterfactual, we turn to a hypothetical example that, while not of any policy importance, will help us think through this key concept a bit more fully. On a conceptual level, solving the counterfactual problem requires the evaluator to identify a

"perfect clone" for each program participant (figure 3.1). For example, let us say that Mr. Fulanito receives an additional $12 in his pocket money allowance, and we want to measure the impact of this treatment on his consumptions of candies. If you could identify a perfect clone for Mr. Fulanito, the evaluation would be easy: you could just compare the number of candies eaten by Mr. Fulanito (say, 6) with the number of candies eaten by his clone (say, 4). In this case, the impact of the additional pocket money would be the difference between those two numbers, or 2 candies. In practice, we know that it is impossible to identify perfect clones: even between genetically identical twins there are important differences.

Although no perfect clone exists for a single individual, statistical tools exist that can be used to generate two groups of individuals that, if their numbers are large enough, are statistically indistinguishable from each other. In practice, a key goal of an impact evaluation is to identify a group of program participants (the treatment group) and a group of nonparticipants (the comparison group) that are statistically identical in the absence of the program. If the two groups are identical, excepting only that one group participates in the program and the other does not, then we can be sure that any difference in outcomes must be due to the program.

The key challenge, then, is to identify a valid comparison group that has the same characteristics as the treatment group. Specifically, the treatment and comparison groups must be the same in at least three ways: First, the

Figure 3.1 The Perfect Clone



Source: Authors.

treatment group and the comparison group must be identical in the absence of the program. Although it is not necessary that every unit in the treatment group be identical to every unit in the comparison group, on average the characteristics of treatment and comparison groups should be the same. For example, the average age in the treatment group should be the same as the average age in the comparison group. Second, the treatment and comparison groups should react to the program in the same way. For example, the incomes of units in the treatment group should be as likely to benefit from training as the incomes of the comparison group. Third, the treatment and comparison groups cannot be differentially exposed to other interventions during the evaluation period. For example, if we are to isolate the impact of the additional pocket money on candy consumption, the treatment group could not also have been provided with more trips to the candy store than the controls, as that could confound the effects of the pocket money with the effect of increased access to candy.

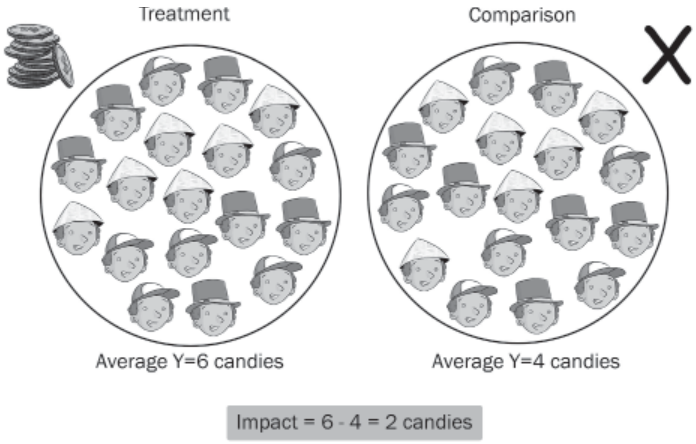
Key Concept:
A valid comparison group will have the same characteristics as the group of participants in the program ("treatment group"), except for the fact that the units in the comparison group do not benefit from the program.

When these three conditions are met, then only the existence of the program of interest will explain any differences in the outcome (Y) between the two groups once the program has been implemented. The reason is that the only difference between the treatment and comparison groups is that the members of the treatment group will receive the program, while the members of the comparison group will not. When the differences in outcomes can be entirely attributed to the program, the causal impact of the program has been identified. So instead of looking at the impact of additional pocket money only for Mr. Fulanito, you may be looking at the impact for a group of children (figure 3.2). If you could identify another group of children that is totally similar, except that they do not receive additional pocket money, your estimate of the impact of the program would be the difference between the two groups in average consumption of candies. Thus, if the *treated group* consumes an average of 6 candies per person, while the *comparison group* consumes an average of 4, the average impact of the additional pocket money on candy consumption would be 2.

Key Concept:
When the comparison group for an evaluation is invalid, then the estimate of the impact of the program will also be invalid: it will not estimate the true impact of the program. In statistical terms, it will be "biased."

Now that we have defined a *valid comparison group*, it is important to consider what would happen if we decided to go ahead with an evaluation without identifying such a group. Intuitively, this should now be clear: an invalid comparison group is one that differs from the treatment group in some way other than the absence of the treatment. Those additional differences can cause our impact estimate to be invalid or, in statistical terms, *biased*: it will not estimate the true impact of the program. Rather, it will estimate the effect of the program mixed with the effect of those other differences.

Figure 3.2 A Valid Comparison Group



Source: Authors.

Two Types of Impact Estimates

Having estimated the impact of the program, the evaluator needs to know how to interpret the results. An evaluation always estimates the impact of a program by comparing the outcomes for the treatment group with the estimate of the counterfactual obtained from a valid comparison group, using the basic impact evaluation equation. Depending on what the treatment and the counterfactual actually represent, the interpretation of the impact of a program will vary.

The estimated impact α is called the "intention-to-treat" estimate (ITT) when the basic formula is applied to those units to whom the program has been offered, regardless of whether or not they actually enroll in it. The ITT is important for those cases in which we are trying to determine the average impact of a program on the population *targeted* by the program. By contrast, the estimated impact α is called the "treatment-on-the-treated" (TOT) when the basic impact evaluation formula is applied to those units to whom the program has been offered and who have actually enrolled. The ITT and TOT estimates will be the same when there is full compliance, that is, when all units to whom a program has been offered actually decide to enroll in it. We will return to the difference between the ITT and TOT estimates in detail in future sections, but let us begin with an example.

Consider the health insurance subsidy program, or HISP, example described in the introduction to part 2, in which any household in a treatment village can sign up for a health insurance subsidy. Even though all

households in treatment villages are eligible to enroll in the program, some fraction of households, say 10 percent, may decide not to do so (perhaps because they already have insurance through their jobs, because they are healthy and do not anticipate the need for health care, or because of one of many other possible reasons). In this scenario, 90 percent of households in the treatment village decide to enroll in the program and actually receive the services that the program provides. The ITT estimate would be obtained by computing the basic impact evaluation formula for all households who were offered the program, that is, for 100 percent of the households in treatment villages. By contrast, the TOT estimate would be obtained by calculating the basic impact evaluation formula only for the subset of households who actually decided to enroll in the program, that is, for the 90 percent of households in treatment villages that enroll.

Two Counterfeit Estimates of the Counterfactual

In the remainder of part 2 of this book, we will discuss the various methods that can be used to construct valid comparison groups that will allow you to estimate the counterfactual. Before doing so, however, it is useful to discuss two common, but highly risky, methods of constructing comparison groups that can lead to inappropriate estimates of the counterfactual. These two "counterfeit" estimates of the counterfactuals are (1) *before-and-after*, or pre-post, comparisons that compare the outcomes of program participants prior to and subsequent to the introduction of a program and (2) *with-and-without* comparisons between units that choose to enroll and units that choose not to enroll.

Counterfeit Counterfactual 1: Comparing Before and After

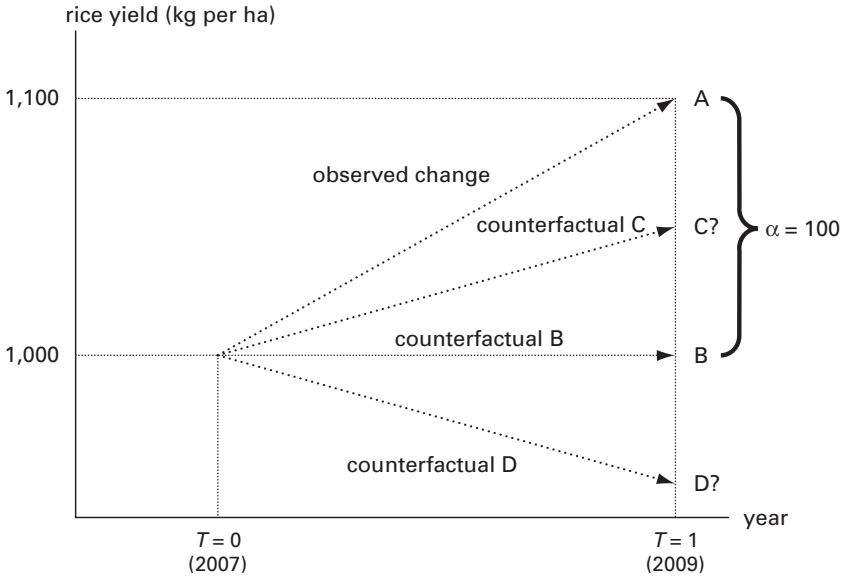
A before-and-after comparison attempts to establish the impact of a program by tracking changes in outcomes for program participants over time. To return to the basic impact evaluation formula, the outcome for the treatment group ($Y \mid P = 1$) is simply the postintervention outcome. However, the counterfactual ($Y \mid P = 0$) is estimated using the preintervention outcome. In essence, this comparison assumes that if the program had never existed, the outcome (Y) for program participants would have been exactly the same as their preprogram situation. Unfortunately, in the vast majority of cases that assumption simply does not hold.

Take the evaluation of a microfinance program for poor, rural farmers. Let us say that the program provides microloans to farmers to enable them

to buy fertilizer to increase their rice production. You observe that in the year before the start of the program, farmers harvested an average of 1,000 kilograms (kg) of rice per hectare. The microfinance scheme is launched, and a year later rice yields have increased to 1,100 kg per hectare. If you were trying to evaluate impact using a before-and-after comparison, you would use the preintervention outcome as a counterfactual. Applying the basic impact evaluation formula, you would conclude that the program had increased rice yields by 100 kg per hectare.

However, imagine that rainfall was normal during the year before the program was launched, but a drought occurred in the year the program started. In this context, the preintervention outcome cannot constitute an appropriate counterfactual. Figure 3.3 illustrates why. Because farmers received the program during a drought year, their average yield without the microloan scheme would have been even lower, at level D, and not level B as the before-and-after comparison assumes. In that case, the true impact of the program is larger than 100 kg. By contrast, if environmental conditions had actually improved over time, the counterfactual rice yield might have been at level C, in which case the true program impact would have been smaller than 100 kg. In other words, unless we can statistically account for

Figure 3.3 Before and After Estimates of a Microfinance Program



Source: Authors, based on the hypothetical example in the text.

rainfall and *every other factor* that can affect rice yields over time, we simply cannot calculate the true impact of the program by making a before-and-after comparison.

Although before-and-after comparisons may be invalid in impact evaluation, that does not mean they are not valuable for other purposes. In fact, administrative data systems for many programs typically record data about participants over time. For example, an education management information system may routinely collect data on student enrollment in the set of schools where a school meal program is operating. Those data allow program managers to observe whether the number of children enrolled in school is increasing over time. This is important and valuable information for managers who are planning and reporting about the education system. However, establishing that the school meal program has *caused* the observed change in enrollment is much more challenging because many different factors affect student enrollment over time. Thus, although monitoring changes in outcomes over time for a group of participants is extremely valuable, it does not usually allow us to determine conclusively whether—or by how much—a particular program of interest contributed to that improvement as long as other time-varying factors exist that are affecting the same outcome.

We saw in the example of the microfinance scheme and rice yields that many factors can affect rice yields over time. Likewise, many factors can affect the majority of outcomes of interest to development programs. For that reason, the preprogram outcome is almost never a good estimate of the counterfactual, and that is why we label it a "counterfeit counterfactual."

Doing a Before-and-After Evaluation of the Health Insurance Subsidy Program

Recall that the HISP is a new program in your country that subsidizes the purchase of health insurance for poor rural households and that this insurance covers expenses related to primary health care and drugs for those enrolled. The objective of the HISP is to reduce the out-of-pocket health expenditures of poor families and ultimately to improve health outcomes. Although many outcome indicators could be considered for the program evaluation, your government is particularly interested in analyzing the effects of the HISP on what poor families spend on primary care and drugs measured as a household's yearly out-of-pocket expenditures per capita (subsequently referred to simply as "health expenditures").

The HISP will represent a hefty proportion of the national budget if scaled up nationally—up to 1.5 percent of gross domestic product (GDP) by

some estimates. Furthermore, substantial administrative and logistical complexities are involved in running a program of this nature. For these reasons, a decision has been made at the highest levels of government to introduce the HISP first as a pilot program and then, depending on the results of the first phase, to scale it up gradually over time. Based on the results of financial and cost-benefit analyses, the president and her cabinet have announced that for the HISP to be viable and to be extended nationally, it must reduce the average yearly per-capita health expenditures of poor rural households by at least $9 below what they would have spent in the absence of the program and it must do so within 2 years.

The HISP will be introduced in 100 rural villages during the initial pilot phase. Just before the start of the program, your government hires a survey firm to conduct a baseline survey of all 4,959 households in these villages. The survey collects detailed information on every household, including their demographic composition, assets, access to health services, and health expenditures in the past year. Shortly after the baseline survey is conducted, the HISP is introduced in the 100 pilot villages with great fanfare, including community events and other promotional campaigns to encourage eligible households to enroll.

Of the 4,959 households in the baseline sample, a total of 2,907 enroll in the HISP during the first 2 years of the program. Over the 2 years, the HISP operates successfully by most measures. Coverage rates are high, and surveys show that most enrolled households are satisfied with the program. At the end of the 2-year pilot period, a second round of evaluation data is collected on the same sample of 4,959 households.[1]

The president and the minister of health have put you in charge of overseeing the impact evaluation for the HISP and recommending whether or not to extend the program nationally. Your impact evaluation question of interest is, *By how much did the HISP lower health expenditures for poor rural households?* Remember that the stakes are high. If the HISP is found to reduce health expenditures by $9 or more, it will be extended nationally. If the program did not reach the $9 target, you will recommend against scaling up the HISP.

The first "expert" evaluation consultant you hire indicates that to estimate the impact of the HISP, you must calculate the change in health expenditures over time for the households that enrolled. The consultant argues that because the HISP covers all health costs related to primary care and medication, any decrease in expenditures over time must be largely attributable to the effect of the HISP. Using only the subset of enrolled households, therefore, you estimate their average health expenditures before the implementation of the program and 2 years later. In

Table 3.1 Case 1—HISP Impact Using Before-After (Comparison of Means)

	After	Before	Difference	*t*-stat
Household health expenditures	7.8	14.4	–6.6	–28.9

Source: Authors' calculations from hypothetical data set.

other words, you perform a before-and-after evaluation. The results are shown in table 3.1.

You observe that the households that enrolled in the HISP reduced their out-of-pocket health expenditures from $14.4 before the introduction of HISP, to $7.8 two years later, a reduction of $6.6 (or 45 percent) over the period. As denoted by the value of the *t*-statistic, the difference between health expenditures before and after the program is *statistically significant,* that is, the probability that the estimated effect is statistically equal to zero is very low.

Even though the before-and-after comparison is for the same group of households, you are concerned that some other factors may have changed over time that affected health expenditures. For example, a number of health interventions have been operating simultaneously in the villages in question. Alternatively, some changes in household expenditures may have resulted from the financial crisis that your country recently experienced. To address some of these concerns, your consultant conducts more sophisticated *regression analysis* that will control for the additional external factors. The results appear in table 3.2.

Here, the linear regression is of health expenditures on a binary (0-1) variable for whether the observation is baseline (0) or follow-up (1). The multivariate linear regression additionally *controls for*, or *holds constant*, other characteristics that are observed for the households in your sample, including indicators for wealth (assets), household composition, and so on. You note that the simple linear regression is equivalent to the simple before-and-after difference in health expenditures (a reduction of $6.59). Once you control for other factors available in your data, you find a similar result—a decrease of $6.65.

Table 3.2 Case 1—HISP Impact Using Before-After (Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	–6.59**	–6.65**
	(0.22)	(0.22)

Source: Authors.

Note: Standard errors are in parentheses.

** Significant at the 1 percent level.

QUESTION 1

A. Based on these results from case 1, should the HISP be scaled up nationally?

B. Does this analysis likely control for all the factors that affect health expenditures over time?

Counterfeit Counterfactual 2: Comparing Enrolled and Nonenrolled

Comparing units that receive a program to units that do not receive it ("with-and-without") constitutes another counterfeit counterfactual. Consider, for example, a vocational training program for unemployed youth. Assume that 2 years after the launching of the scheme, an evaluation attempts to estimate the impact of the program on income by comparing the average incomes of a group of youth who chose to enroll in the program versus those of a group who chose not to enroll. Assume that the results show that the youths who enrolled in the program make twice as much as those who did not enroll.

How should these results be interpreted? In this case, the counterfactual is estimated based on the incomes of individuals who decided not to enroll in the program. Yet the two groups of young people are likely to be fundamentally different. Those individuals who chose to participate may be highly motivated to improve their livelihoods and may expect a high return to training. In contrast, those who chose not to enroll may be discouraged youth who do not expect to benefit from this type of program. It is likely that these two types of young people would perform quite differently in the labor market and would have different incomes even without the vocational training program.

Therefore, the group that chose not to enroll does not provide a good estimate of the counterfactual. If a difference in incomes is observed between the two groups, we will not be able to determine whether it comes from the training program or from the underlying differences in motivation and other factors that exist between the two groups. The fact that less-motivated individuals chose not to enroll in the training program therefore leads to a bias in our assessment of the program's impact.[2] This bias is called "selection bias." In this case, if the young people who enrolled would have had higher incomes even in the absence of the program, the selection bias would be positive; in other words, we would overestimate the impact of the vocational training program on incomes.

Key Concept:

Selection bias occurs when the reasons for which an individual participates in a program are correlated with outcomes. This bias commonly occurs when the comparison group is ineligible for the program or decides not to participate in it.

Comparing Units that Chose to Enroll in the Health Insurance Subsidy Program with Those that Chose Not to Enroll

Having thought through the before-after comparison a bit further with your evaluation team, you realize that there are still many time-varying factors

that can explain part of the change in health expenditures over time (in particular, the minister of finance is concerned that the recent financial crisis may have affected households' health expenditures and may explain the observed change). Another consultant suggests that it would be more appropriate to estimate the counterfactual in the postintervention period, that is, 2 years after the program started. The consultant correctly notes that of the 4,959 households in the baseline sample, only 2,907 actually enrolled in the program, and so approximately 41 percent of the households in the sample remain without the HISP coverage. The consultant argues that households within the same locality would be exposed to the same supply-side health interventions and the same local economic conditions, so that the postintervention outcomes of the nonenrolled group would help to control for many of the environmental factors that affect both enrolled and nonenrolled households.

You therefore decide to calculate average health expenditures in the postintervention period for both the households that enrolled in the program and the households that did not, producing the observations shown in table 3.3.

Using the average health expenditures of the nonenrolled households as the estimate of the counterfactual, you find that the program has reduced average health expenditures by approximately $14. When discussing this result further with the consultant, you raise the question of whether the households that chose not to enroll in the program may be systematically different from the ones that did enroll. For example, the households that signed up for the HISP may be ones that expected to have higher health expenditures, or people who were better informed about the program, or people who care more for the health of their families. Alternatively, perhaps the households that enrolled were poorer, on average, than those who did not enroll, given that the HISP is targeted to poor households. Your consultant assures you that regression analysis can control for the potential differences between the two groups. Controlling for all household characteristics that are in the data set, the consultant estimates the impact of the program as shown in table 3.4.

Table 3.3 Case 2—HISP Impact Using Enrolled-Nonenrolled (Comparison of Means)

	Enrolled	Nonenrolled	Difference	*t*-stat
Household health expenditures	7.8	21.8	–13.9	–39.5

Source: Authors.

Table 3.4 Case 2—HISP Impact Using Enrolled-Nonenrolled (Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	−13.9** (0.35)	−9.4** (0.32)

Source: Authors.

Note: Standard errors are in parentheses.

** Significant at the 1 percent level.

With a simple linear regression of health expenditures on an indicator variable for whether or not a household enrolled in the program, you find an estimated impact of minus $13.90; in other words, you estimate that the program has decreased average health expenditures by $13.90. However, when all other characteristics of the sample population are held constant, you estimate that the program has reduced the expenditures of the enrolled households by $9.40 per year.

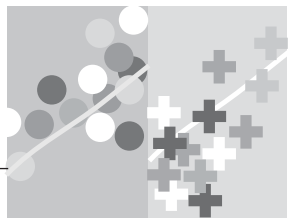
QUESTION 2

A. Based on these results from case 2, should the HISP be scaled up nationally?

B. Does this analysis likely control for all the factors that determine differences in health expenditures between the two groups?

Notes

1. Note that we are assuming zero sample attrition over 2 years, that is, no households will have left the sample. This is not a realistic assumption for most household surveys. In practice, families who move sometimes cannot be tracked to their new location, and some households break up and cease to exist altogether.
2. As another example, if youth who anticipate benefiting considerably from the training scheme are also more likely to enroll (for example, because they anticipate higher wages with training), then we will be comparing a group of individuals who anticipated higher income with a group of individuals who did not anticipate higher income.



CHAPTER 4

Randomized Selection Methods

Having discussed two approaches to constructing counterfactuals that are commonly used but have a high risk of bias—before-and-after comparisons and with-and-without comparisons—we now turn to a set of methods that can be applied to estimate program impacts more accurately. As we will see, however, such estimation is not always as straightforward as it might seem at first glance. Most programs are designed and implemented in a complex and changing environment, in which many factors can influence outcomes both for program participants and for those who do not participate. Droughts, earthquakes, recessions, changes in government, and changes in international and local policies are all part of the real world, and as evaluators, we want to make sure that the estimated impact of our program remains valid despite these myriad factors.

As we will see throughout this part of the book, a program's rules for enrolling participants will be the key parameter for selecting the impact evaluation method. We believe that in most cases the evaluation methods should try to fit within the context of a program's operational rules (with a few tweaks here and there) and not the other way around. However, we also start from the premise that *all social programs should have fair and transparent rules for program assignment*. One of the fairest and most transparent rules for allocating scarce resources among equally deserving populations turns out to be giving everyone who is eligible an equal opportunity to participate in the program. One way to do that is simply to run a lottery. In this chapter, we will examine several *randomized selection methods*; these are

akin to running lotteries that decide who enters a program at a given time and who does not. These randomized selection methods not only provide program administrators with a fair and transparent rule for allocating scarce resources *among equally deserving populations,* but also represent the strongest methods for evaluating the impact of a program.

Randomized selection methods can often be derived from a program's operational rules. For many programs, the population of intended participants—that is, the set of all units that the program would like to serve—is larger than the number of participants that the program can actually accommodate at a given time. For example, in a single year an education program may provide school materials and an upgraded curriculum to 500 schools out of thousands of eligible schools in the country. Or a youth employment program may have a goal of reaching 2,000 unemployed youths within its first year of operation, although there are tens of thousands of unemployed young people that the program ultimately would like to serve. For any of a variety of reasons, programs may be unable to reach the entire population of interest. Budgetary constraints may simply prevent the administrators from offering the program to all eligible units from the beginning. Even if budgets are available to cover an unlimited number of participants, capacity constraints will sometimes prevent a program from rolling out to everyone at the same time. In the youth employment training program example, the number of unemployed youth who want vocational training may be greater than the number of slots available in technical colleges during the first year of the program, and that may limit the number who can enroll.

In reality, most programs have budgetary or operational capacity constraints that prevent reaching every intended participant at the same moment. In this context, where the population of eligible participants is larger than the number of program places available, program administrators must define a rationing mechanism to allocate the program's services. In other words, someone must make a decision about who will enter the program and who will not. The program could be assigned on a first-come-first-served basis, or based on observed characteristics (for example, women and children first, or the poorest municipalities first); or selection could be based on unobserved characteristics (for example, letting individuals sign up based on their own motivation and knowledge), or even on a lottery.

Randomized Assignment of the Treatment

When a program is assigned at random over a large eligible population, we can generate a robust estimate of the counterfactual, considered the gold

standard of impact evaluation. Randomized assignment of treatment essentially uses a lottery to decide who among the equally eligible population receives the program and who does not.[1] Every eligible unit of treatment (for example, an individual, household, community, school, hospital, or other) has an equal probability of selection for treatment.[2]

Before we discuss how to implement randomized assignment in practice and why it generates a strong counterfactual, let us take a few moments to consider why randomized assignment is also a fair and transparent way to assign scarce program services. Once a target population has been defined (say, households below the poverty line, or children under the age of 5, or schools in rural areas), randomized assignment is a fair allocation rule because it allows program managers to ensure that every eligible person or unit has the same chance of receiving the program and that the program is not assigned using arbitrary or subjective criteria, or even through patronage or other unfair practices. When excess demand for a program exists, randomized assignment is a rule that can be easily explained by program managers and easily understood by key constituents. When the selection process is conducted through an open and replicable process, the randomized assignment rule cannot easily be manipulated, and therefore it shields program managers from potential accusations of favoritism or corruption. Randomized assignment thus has its own merits as a rationing mechanism that go well beyond its utility as an impact evaluation tool. In fact, we have come across a number of programs that routinely use lotteries as a way to select participants from the pool of eligible individuals, primarily because of their advantages for administration and governance.[3]

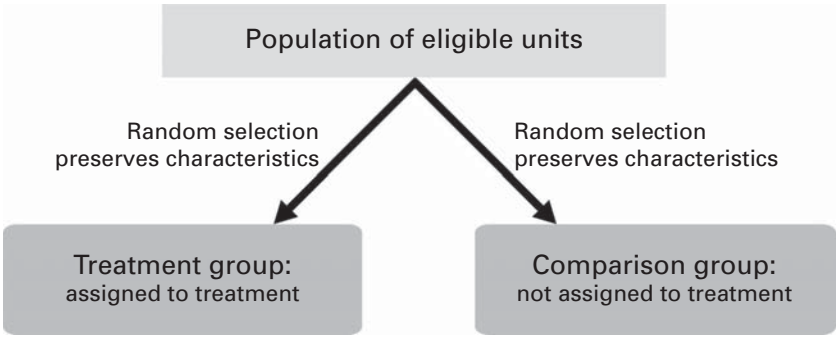
Why Does Randomized Assignment Produce an Excellent Estimate of the Counterfactual?

As discussed previously, the ideal comparison group will be as similar as possible to the treatment group in all respects, except with respect to its enrollment in the program that is being evaluated. The key is that when we randomly select units to assign them to the treatment and comparison groups, that randomized assignment process in itself will produce two groups that have a high probability of being statistically identical, as long as the number of potential participants to which we apply the randomized assignment process is sufficiently large. Specifically, with a large enough number of observations, the randomized assignment process will produce groups that have statistically equivalent *averages for all their characteristics*. In turn, those averages also tend toward the average of the population from which they are drawn.[4]

Figure 4.1 illustrates why randomized assignment produces a comparison group that is statistically equivalent to the treatment group. Suppose the population of eligible units (the potential participants) consists of 1,000 people, of whom half are randomly selected and assigned to the treatment group and the other half to the comparison group. For example, one could imagine writing the names of all 1,000 people on individual pieces of paper, mixing them up in a bowl, and then asking someone to blindly draw out 500 names. If it was determined that the first 500 names would constitute the treatment group, then you would have a randomly assigned treatment group (the first 500 names drawn), and a randomly assigned comparison group (the 500 names left in the bowl).

Now assume that of the original 1,000 people, 40 percent were women. Because the names were selected at random, of the 500 names drawn from the bowl, approximately 40 percent will also be women. If among the 1,000 people, 20 percent had blue eyes, then approximately 20 percent of both the treatment and the comparison groups should have blue eyes, too. In general, if the population of eligible units is large enough, then any characteristic of the population will flow through to both the treatment group and the comparison groups. We can imagine that if observed characteristics such as sex or the color of a person's eyes flow through to both the treatment and the comparison group, then logically characteristics that are more difficult to observe (unobserved variables), such as motivation, preferences, or other difficult-to-measure personality traits, would also flow through equally to both the treatment and the comparison groups. Thus, treatment and comparison groups that are generated through randomized assignment will be similar not only in their observed characteristics but also in

Figure 4.1 Characteristics of Groups under Randomized Assignment of Treatment



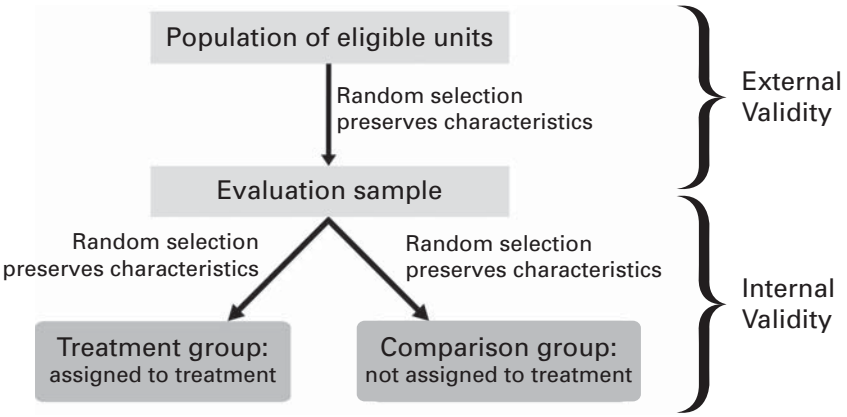
Source: Authors.

their unobserved characteristics. For example, you may not be able to observe or measure how "nice" people are, but you know that if 20 percent of the people in the population of eligible units are nice, then approximately 20 percent of the people in the treatment group will be nice, and the same will be true of the comparison group. Randomized assignment will help guarantee that, on average, the treatment and comparison groups are similar in every way, in both observed and unobserved characteristics.

When an evaluation uses randomized assignment to treatment and comparison groups, we know that theoretically the process should produce two groups that are equivalent. With baseline data on our evaluation sample, we can test this assumption empirically and verify that in fact there are no systematic differences in observed characteristics between the treatment and comparison groups before the program starts. Then, after we launch the program, if we observe differences in outcomes between the treatment and comparison groups, we will know that those differences can be explained only by the introduction of the program, since by construction the two groups were identical at baseline and are exposed to the same external environmental factors over time. In this sense, the comparison group *controls* for all factors that might also explain the outcome of interest. We can be very confident that our estimated average impact, given as the difference between the outcome under treatment (the mean outcome of the randomly assigned treatment group), and our estimate of the counterfactual (the mean outcome of the randomly assigned comparison group) constitute the true impact of the program, since by construction we have eliminated all observed and unobserved factors that might otherwise plausibly explain the difference in outcomes.

In figure 4.1 it is assumed that all units in the eligible population would be assigned to either the treatment or the comparison group. In some cases, however, it is not necessary to include all of them in the evaluation. For example, if the population of eligible units includes a million mothers, and you want to evaluate the effectiveness of cash bonuses on the probability of their vaccinating their children, it may be sufficient to take a representative sample of, say, 1,000 mothers and assign those 1,000 to either the treatment or the comparison group. Figure 4.2 illustrates this process. By the same logic explained above, taking a random sample from the population of eligible units to form the evaluation sample preserves the characteristics of the population of eligible units. The random selection of the treatment and comparison groups from the evaluation sample again preserves the characteristics.

Figure 4.2 Random Sampling and Randomized Assignment of Treatment



Source: Authors.

External and Internal Validity

The steps outlined above for randomized assignment of treatment will ensure both the internal and the external validity of the impact evaluation, as long as the evaluation sample is large enough (figure 4.2).

Key Concept:
An evaluation is internally valid if it uses a valid comparison group.

Internal validity means that the estimated impact of the program is net of all other potential confounding factors, or that the comparison group represents the true counterfactual, so that we are estimating the true impact of the program. Remember that randomized assignment produces a comparison group that is statistically equivalent to the treatment group at baseline, before the program starts. Once the program starts, the comparison group is exposed to the same set of external factors over time, the only exception being the program. Therefore, if any differences in outcomes appear between the treatment and comparison groups, they can only be due to the existence of the program in the comparison group. In other words, the internal validity of an impact evaluation is ensured through the process of *randomized assignment of treatment*.

Key Concept:
An evaluation is externally valid if the evaluation sample accurately represents the population of eligible units. The results are then generalizable to the population of eligible units.

External validity means that the impact estimated in the evaluation sample can be generalized to the population of all eligible units. For this to be possible, the evaluation sample must be representative of the population of eligible units; in practice, it means that the evaluation sample must be selected from the population by using one of several variations of *random sampling*.[5]

Note that we have brought up two different types of randomization: one for the purpose of sampling (for external validity) and one as an impact eval-

uation method (for internal validity). An impact evaluation can produce internally valid estimates of impact through randomized assignment of treatment; however, if the evaluation is performed on a nonrandom sample of the population, the estimated impacts may not be generalizable to the population of eligible units. Conversely, if the evaluation uses a random sample of the population of eligible units, but treatment is not assigned in a randomized way, then the sample would be representative, but the comparison group may not be valid.

When Can Randomized Assignment Be Used?

In practice, randomized assignment should be considered whenever a program is oversubscribed, that is, when the number of potential participants is larger than the number of program spaces available at a given time and the program needs to be phased in. Some circumstances also merit randomized assignment as an evaluation tool even if program resources are not limited. For example, governments may want to use randomized assignment to test new or potentially costly programs whose intended and unintended consequences are unknown. In this context, randomized assignment is justified during a pilot evaluation period to rigorously test the effects of the program before it is rolled out to a larger population.

Two scenarios commonly occur in which randomized assignment is feasible as an impact evaluation method:

1. *When the eligible population is greater than the number of program spaces available.* When the demand for a program exceeds the supply, a simple lottery can be used to select the treatment group within the eligible population. In this context, every unit in the population receives an equal chance of being selected for the program. The group that wins the lottery is the treatment group, and the rest of the population that is not offered the program is the comparison group. As long as a resource constraint exists that prevents scaling the program up to the entire population, the comparison groups can be maintained to measure the short-, medium-, and long-term impacts of the program. In this context, no ethical dilemma arises from holding a comparison group indefinitely, since a subset of the population will necessarily be left out of the program.

   As an example, suppose the ministry of education wants to provide school libraries to public schools throughout the country, but the ministry of finance budgets only enough funds to cover one-third of them. If the ministry of education wants each public school to have an equal chance of receiving a library, it would run a lottery in which each school

has the same chance (1 in 3) of being selected. Schools that win the lottery receive a new library and constitute the treatment group, and the remaining two-thirds of public schools in the country are not offered the library and serve as the comparison group. Unless additional funds are allocated to the library program, a group of schools will remain that do not have funding for libraries through the program, and they can be used as a comparison group to measure the counterfactual.

2. *When a program needs to be gradually phased in until it covers the entire eligible population.* When a program is phased in, randomization of the order in which participants receive the program gives each eligible unit the same chance of receiving treatment in the first phase or in a later phase of the program. As long as the "last" group has not yet been phased into the program, it serves as a valid comparison group from which we can estimate the counterfactual for the groups that have already been phased in.

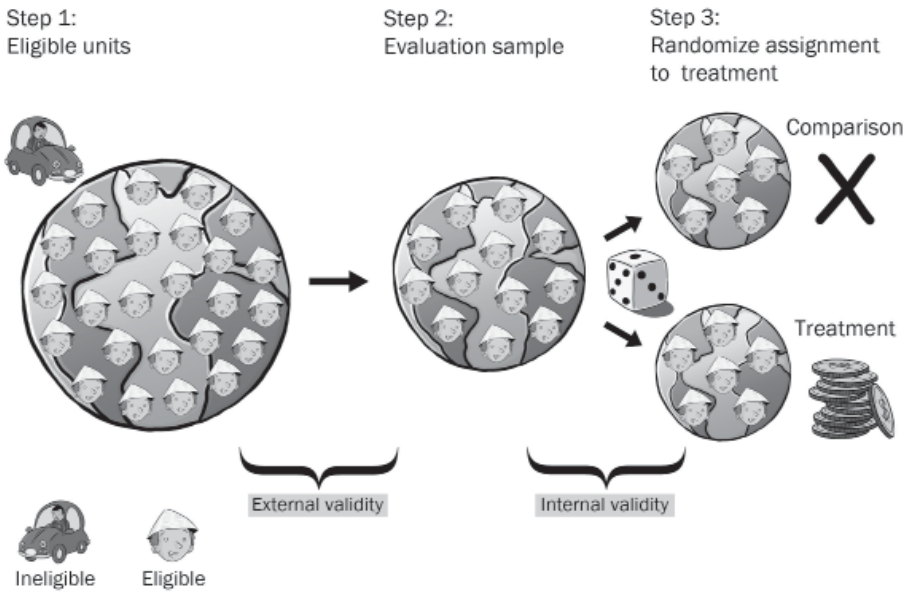
   For example, suppose that the ministry of health wants to train all 15,000 nurses in the country to use a new health protocol but needs three years to train them all. In the context of an impact evaluation, the ministry could randomly select one-third of the nurses to receive training in the first year, one-third to receive training in the second year, and one-third to receive training in the third year. To evaluate the effect of the training program one year after its implementation, the group of nurses trained in year 1 would constitute the treatment group, and the group of nurses randomly assigned to training in year 3 would be the comparison group, since they would not yet have received the training.

How Do You Randomly Assign Treatment?

Now that we have discussed what randomized assignment does and why it produces a good comparison group, we will turn to the steps in successfully assigning treatment in a randomized way. Figure 4.3 illustrates this process.

Step 1 in randomized assignment is to define the units that are eligible for the program. Depending on the particular program, a unit can be a person, a health center, a school, or even an entire village or municipality. The population of eligible units consists of those for which you are interested in knowing the impact of your program. For example, if you are implementing a training program for primary school teachers in rural areas, then secondary school teachers or primary school teachers in urban areas would not belong to your population of eligible units.

Figure 4.3 Steps in Randomized Assignment to Treatment



Source: Authors.

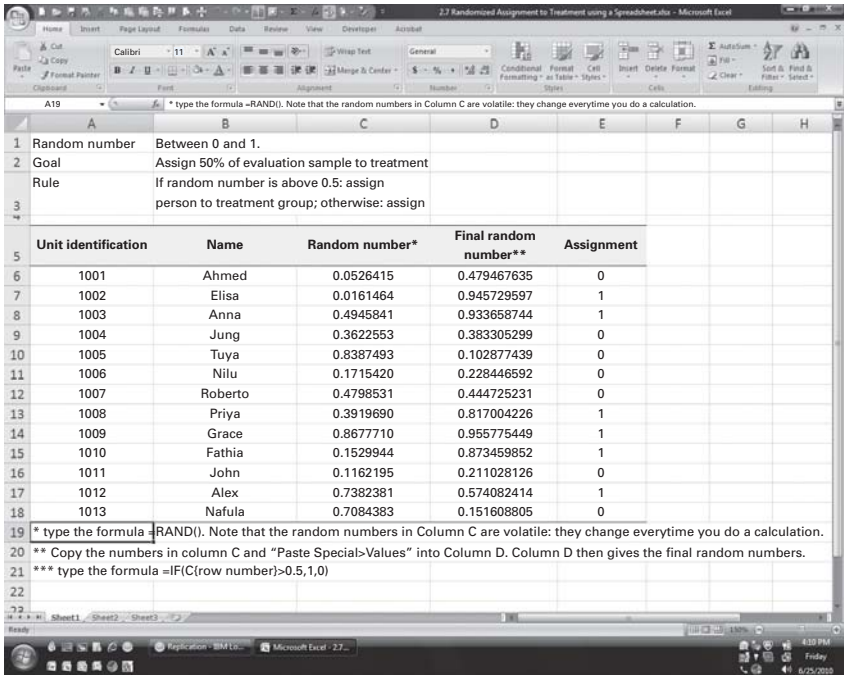
Once you have determined the population of eligible units, it will be necessary to compare the size of the group with the number of observations required for the evaluation. This number is determined through power calculations and is based on the types of questions you would like answered (see chapter 11). If the eligible population is small, all of the eligible units may need to be included in the evaluation. Alternatively, if there are more eligible units than are required for the evaluation, then step 2 is to select a sample of units from the population to be included in the evaluation sample. Note that this second step is done mainly to limit data collection costs. If it is found that data from existing monitoring systems can be used for the evaluation, and that those systems cover the population of eligible units, then you will not need to draw a separate evaluation sample. However, imagine an evaluation in which the population of eligible units includes tens of thousands of teachers in every public school in the country, and you need to collect detailed information on teacher pedagogical knowledge. Interviewing each and every teacher may not be practically feasible, but you may find that it is sufficient to take a sample of 1,000 teachers distributed over 100 schools. As long as the sample of schools and teachers is representative of the whole population of public school teachers, any results found in the evaluation can be generalized to the rest of the teachers and public schools in the country. Collecting data

on this sample of 1,000 teachers will of course be much cheaper than collecting data on every teacher in all public schools in the country.

Finally, step 3 will be forming the treatment and comparison groups among the units in the evaluation sample. This requires that you first decide on a rule for how to assign participants based on random numbers. For example, if you need to assign 40 out of 100 units from the evaluation sample to the treatment group, you may decide to assign those 40 units with the highest random numbers to the treatment group and the rest to the comparison group. You then assign a random number to each unit of observation in the evaluation sample, using a spreadsheet or specialized statistical software (figure 4.4), and use your previously chosen rule to form the treatment and comparison groups. Note that it is important to decide on the rule before you run the software that gives units their random numbers; otherwise, you may be tempted to decide on a rule based on the random numbers you see, and that would invalidate the randomized assignment.

The logic behind the automated process is no different from randomized assignment based on a coin toss or picking names out of a hat: it is a mechanism that determines randomly whether each unit is in the treatment or the

Figure 4.4 Randomized Assignment to Treatment Using a Spreadsheet



	A	B	C	D	E
1	Random number	Between 0 and 1.			
2	Goal	Assign 50% of evaluation sample to treatment			
3	Rule	If random number is above 0.5: assign person to treatment group; otherwise: assign			
5	**Unit identification**	**Name**	**Random number***	**Final random number****	**Assignment**
6	1001	Ahmed	0.0526415	0.479467635	0
7	1002	Elisa	0.0161464	0.945729597	1
8	1003	Anna	0.4945841	0.933658744	1
9	1004	Jung	0.3622553	0.383305299	0
10	1005	Tuya	0.8387493	0.102877439	0
11	1006	Nilu	0.1715420	0.228446592	0
12	1007	Roberto	0.4798531	0.444725231	0
13	1008	Priya	0.3919690	0.817004226	1
14	1009	Grace	0.8677710	0.955775449	1
15	1010	Fathia	0.1529944	0.873459852	1
16	1011	John	0.1162195	0.211028126	0
17	1012	Alex	0.7382381	0.574082414	1
18	1013	Nafula	0.7084383	0.151608805	0
19	* type the formula =RAND(). Note that the random numbers in Column C are volatile: they change everytime you do a calculation.				
20	** Copy the numbers in column C and "Paste Special>Values" into Column D. Column D then gives the final random numbers.				
21	*** type the formula =IF(C{row number}>0.5,1,0)				

Source: Authors.

comparison group. In cases where randomized assignment needs to be done in a public forum, some more "artisanal" techniques for randomized assignment might be used. The following examples assume that the unit of randomization is an individual person:

1. If you want to assign 50 percent of individuals to the treatment group and 50 percent to the comparison group, flip a coin for each person. You must decide in advance whether heads or tails on the coin will assign a person to the treatment group.
2. If you want to assign one-third of the evaluation sample to the treatment group, you can roll dice for each person. First, you must decide on a rule. For example, a thrown die that shows a 1 or a 2 could mean an assignment to the treatment group, whereas a 3, 4, 5, or 6 would mean an assignment to the comparison group. You would roll the die once for each person in the evaluation sample and assign them based on the number that comes up.
3. Write the names of all of the people on pieces of paper of identical size and shape. Fold the papers so that the names cannot be seen, and mix them thoroughly in a hat or some other container. Before you start drawing, decide on your rule, that is, how many pieces of paper you will draw and that one's name being drawn means being assigned to the treatment group. Once the rule is clear, ask someone in the crowd (someone unbiased, such as a child) to draw out as many pieces of paper as you need participants in the treatment group.

Whether you use a public lottery, a roll of dice, or computer-generated random numbers, it is important to document the process to ensure that it is transparent. That means, first, that the assignment rule has to be decided in advance and communicated to any members of the public. Second, you must stick to the rule once you draw the random numbers; and third, you must be able to show that the process was really random. In the cases of lotteries and throwing dice, you could videotape the process; computer-based assignment of random numbers requires that you provide a log of your computations, so that the process can be replicated by auditors.[6]

At What Level Do You Perform Randomized Assignment?

Randomized assignment can be done at the individual, household, community, or regional level. In general, the level at which we randomly assign units to treatment and comparison groups will be greatly affected by where and how the program is being implemented. For example, if a

health program is being implemented at the health clinic level, you would first choose a random sample of health clinics and then randomly assign some of them to the treatment group and others to the comparison group.

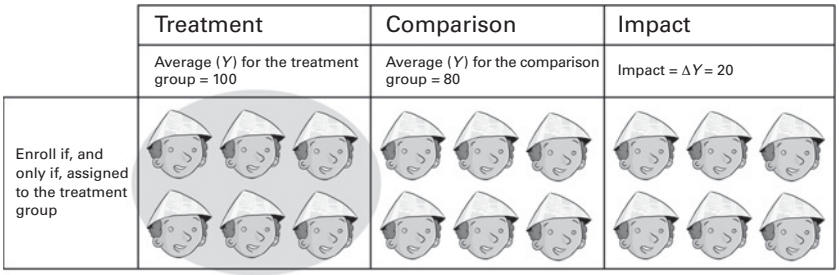
When the level of the randomized assignment is higher, for example, at the level of regions or provinces in a country, it can become very difficult to perform an impact evaluation because the number of regions or provinces in most countries is not sufficiently large to yield balanced treatment and comparison groups. For example, if a country has only six provinces, that would permit only three treatment and three comparison provinces, numbers that are insufficient to ensure that the characteristics of the treatment and comparison groups are balanced.

But as the level of randomized assignment gets lower, for example, down to the individual or household level, the chances of spillovers and contamination increase.[7] For example, if the program consists of providing deworming medicine to households, and a household in the treatment group is located close to a household in the comparison group, then the comparison household may be positively affected by a spillover from the treatment provided to the treatment household because its chances of contracting worms from the neighbors will be reduced. Treatment and comparison households need to be located sufficiently far from each other to avoid such spillovers. Yet, as the distance between the households increases, it will become more costly both to implement the program and to administer surveys. As a rule of thumb, if spillovers can be reasonably ruled out, it is best to perform randomized assignment of the treatment at the lowest possible level of program implementation; that will ensure that the number of units in both the treatment and comparison groups is as large as possible. Spillovers are discussed in chapter 8.

Estimating Impact under Randomized Assignment

Once you have drawn a random evaluation sample and assigned treatment in a randomized fashion, it is quite easy to estimate the impact of the program. After the program has run for some time, outcomes for both the treatment and comparison units will need to be measured. The impact of the program is simply the difference between the average outcome (Y) for the treatment group and the average outcome (Y) for the comparison group. For instance, in figure 4.5, average outcome for the treatment group is 100, and average outcome for the comparison group is 80, so that the impact of the program is 20.

Figure 4.5 Estimating Impact under Randomized Assignment



Source: Authors.

Estimating the Impact of the Health Insurance Subsidy Program under Randomized Assignment

Let us now turn back to the example of the health insurance subsidy program (HISP) and check what "randomized assignment" means in its context. Recall that you are trying to estimate the impact of the program from a pilot that involves 100 treatment villages.

Having conducted two impact assessments using potentially biased counterfactuals (and having reached conflicting policy recommendations; see chapter 3), you decide to go back to the drawing board to rethink how to obtain a more precise counterfactual. After further deliberations with your evaluation team, you are convinced that constructing a valid estimate of the counterfactual will require identifying a group of villages that are identical to the 100 treatment villages in all respects, with the only exception being that one group took part in the HISP and the other did not. Because the HISP was rolled out as a pilot, and the 100 treatment villages were selected randomly from among all of the rural villages in the country, you note that the villages should, on average, have the same characteristics as the general population of rural villages. The counterfactual can therefore be estimated in a valid way by measuring the health expenditures of eligible households in villages that did not take part in the program.

Luckily, at the time of the baseline and follow-up surveys, the survey firm collected data on an additional 100 rural villages that were not offered the program in the first round. Those 100 additional villages were also randomly chosen from the population of eligible villages, which means that they too will, on average, have the same characteristics as the general population of rural villages. Thus, the way that the two groups of villages were chosen ensures that they have identical characteristics, except that the 100 treatment villages received the HISP and the 100 comparison villages did not. Randomized assignment of the treatment has occurred.

Given randomized assignment of the treatment, you are quite confident that no external factors other than the HISP would explain any differences in outcomes between the treatment and comparison villages. To validate this assumption, you test whether eligible households in the treatment and comparison villages have similar characteristics at the baseline as shown in table 4.1.

You observe that the average characteristics of households in the treatment and comparison villages are in fact very similar. The only statistically significant difference is for the number of years of education of the spouse, and that difference is small. Note that even with a randomized experiment on a large sample, a small number of differences can be expected.[8] With the validity of the comparison group established, your estimate of the counterfactual is now the average health expenditures of eligible households in the 100 comparison villages (table 4.2).

Table 4.1 Case 3—Balance between Treatment and Comparison Villages at Baseline

Household characteristics	Treatment villages (N = 2964)	Comparison villages (N = 2664)	Difference	*t*-stat
Health expenditures ($ yearly per capita)	14.48	14.57	−0.09	−0.39
Head of household's age (years)	41.6	42.3	−0.7	−1.2
Spouse's age (years)	36.8	36.8	0.0	0.38
Head of household's education (years)	2.9	2.8	0.1	2.16*
Spouse's education (years)	2.7	2.6	0.1	0.006
Head of household is female = 1	0.07	0.07	−0.0	−0.66
Indigenous = 1	0.42	0.42	0.0	0.21
Number of household members	5.7	5.7	0.0	1.21
Has bathroom = 1	0.57	0.56	0.01	1.04
Hectares of land	1.67	1.71	−0.04	−1.35
Distance to hospital (km)	109	106	3	1.02

Source: Authors' calculation.

* Significant at the 5 percent level.

Table 4.2 Case 3—HISP Impact Using Randomized Assignment (Comparison of Means)

	Treatment	Comparison	Difference	*t*-stat
Household health expenditures baseline	14.48	14.57	−0.09	−0.39
Household health expenditures follow-up	7.8	17.9	−10.1**	−25.6

Source: Authors' calculation.

** Significant at the 1 percent level.

Table 4.3 Case 3—HISP Impact Using Randomized Assignment (Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	−10.1** (0.39)	−10.0** (0.34)

Source: Authors' calculation.

Note: Standard errors are in parentheses.

** Significant at the 1 percent level.

Given that you now have a valid estimate of the counterfactual, you can find the impact of the HISP simply by taking the difference between the out-of-pocket health expenditures of eligible households in the treatment villages and the estimate of the counterfactual. The impact is a reduction of $10.10 over two years. Replicating this result through regression analysis yields the same result, as shown in table 4.3.

With randomized assignment, we can be confident that no factors are present that are systematically different between the treatment and comparison groups that might also explain the difference in health expenditures. Both sets of villages have been exposed to the same set of national policies and programs during the two years of treatment. Thus, the most plausible reason that poor households in treatment communities have lower expenditures than households in comparison villages is that the first group received the health insurance program and the other group did not.

QUESTION 3

A. Why is the impact estimate derived using a multivariate linear regression basically unchanged when controlling for other factors?

B. Based on the impact estimated in case 3, should the HISP be scaled up nationally?

Randomized Assignment at Work

Randomized assignment is often used in rigorous impact evaluation work, both in large-scale evaluations and in smaller ones. The evaluation of the Mexico Progresa program (Schultz 2004) is one of the most well-known, large-scale evaluations using randomized assignment (box 4.1).

Two Variations on Randomized Assignment

We now consider two variations that draw on many of the properties of randomized assignment: randomized offering of treatment and randomized promotion of treatment.

Box 4.1: Conditional Cash Transfers and Education in Mexico

The Progresa program, now called "Oportunidades," began in 1998 and provides cash transfers to poor mothers in rural Mexico conditional on their children's enrollment in school, with their attendance confirmed by the teacher. This large-scale social program was one of the first to be designed with a rigorous evaluation in mind, and randomized assignment was used to help identify the effect of conditional cash transfers on a number of outcomes, in particular school enrollment.

The grants, for children in grades 3 through 9, amount to about 50 percent to 75 percent of the private cost of schooling and are guaranteed for three years. The communities and households eligible for the program were determined based on a poverty index created from census data and baseline data collection. Because of a need to phase in the large-scale social program, about two-thirds of the localities (314 out of 495) were randomly selected to receive the program in the first two years, and the remaining 181 served as a control group before entering the program in the third year.

Based on the randomized assignment, Schultz (2004) found an average increase in enrollment of 3.4 percent for all students in grades 1–8, with the largest increase among girls who had completed grade 6, at 14.8 percent.[a] The likely reason is that girls tend to drop out of school at greater rates as they get older, and so they were given a slightly larger transfer to stay in school past the primary grade levels. These short-run impacts were then extrapolated to predict the longer-term impact of the Progresa program on lifetime schooling and earnings.

Source: Schultz 2004.

a. To be precise, Schultz combined randomized assignment with difference-in-difference methods. Chapter 8 discusses the benefits of combining various impact evaluation methodologies.

Randomized Offering: When Not Everyone Complies with Their Assignment

In the earlier, discussion of randomized assignment, we have assumed that the program administrator has the power to assign units to treatment and comparison groups, with those assigned to the treatment taking the program and those assigned to the comparison group not taking the program. In other words, units that were assigned to the treatment and comparison groups *complied with* their assignment. Full compliance is more frequently attained in laboratory settings or medical trials, where the researcher can carefully make sure, first, that all subjects in the treatment group take the pill, and second, that none of the subjects in the comparison group take it.[9]

In real-life social programs, full compliance with a program's selection criteria (and hence, adherence to treatment or comparison status) is optimal, and policy makers and impact evaluators alike strive to come as close to that ideal as possible. In practice, however, strict 100 percent compliance to treatment and comparison assignments may not occur, despite the best efforts of the program implementer and the impact evaluator. Just because a teacher is assigned to the treatment group and is offered training does not mean that she or he will actually show up on the first day of the course. Similarly, a teacher who is assigned to the comparison group may find a way to attend the course anyway. Under these circumstances, a straight comparison of the group originally assigned to treatment with the group originally assigned to comparison will yield the *"intention-to-treat" estimate (ITT)*. The reason is that we will be comparing those whom we intended to treat (those assigned to the treatment group) with those whom we intended not to treat (those assigned to the comparison group). By itself, this is a very interesting and relevant measure of impact, since most policy makers and program managers can only offer a program and cannot force the program on their target population.

But at the same time, we may also be interested in estimating the impact of the program on those who actually take up or accept the treatment. Doing that requires correcting for the fact that some of the units assigned to the treatment group did not actually receive the treatment, or that some of the units assigned to the comparison group actually did receive it. In other words, we want to estimate the impact of the program on those to whom treatment was offered *and* who actually enrolled. This is the *"treatment-on-the-treated" estimate (TOT)*.

Randomized Offering of a Program and Final Take-Up

Imagine that you are evaluating the impact of a job training program on individuals' wages. The program is randomly assigned at the individual

level, and the treatment group is offered the program while the comparison group is not. Most likely, you will find three types of individuals in the population:

- *Enroll-if-offered*. These are the individuals who comply with their assignment. If they are assigned to the treatment group (offered the program), they take it up, or enroll; if they are assigned to the comparison group (not offered the program), they do not enroll.
- *Never*. These are the individuals that never enroll in or take up the program, even if they are assigned to the treatment group. They are noncompliers in the treatment group.
- *Always*. These are the individuals who will find a way to enroll in the program or take it up, even if they are assigned to the comparison group. They are noncompliers in the comparison group.

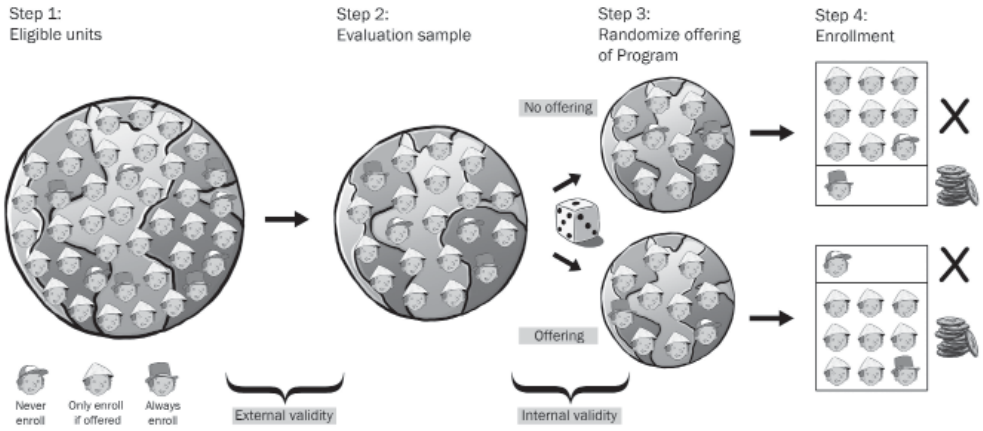
In the context of the job training program, the *Never* group might be unmotivated people who, even if offered a place in the course, do not show up. The *Always* group, in contrast, are so motivated that they find a way to enter the program even if they were originally assigned to the comparison group. The *Enroll-if-offered* group are those who would enroll in the course if it is offered (the treatment group) but do not seek to enroll if they are assigned to the comparison group.

Figure 4.6 presents the randomized offering of the program and the final enrollment, or take-up, when *Enroll-if-offered, Never,* and *Always* groups are present. We assume that the population of units has 80 percent *Enroll-if-offered,* 10 percent *Never,* and 10 percent *Always.* If we take a random sample of the population for the evaluation sample, then the evaluation sample will also have approximately 80 percent *Enroll-if-offered,* 10 percent *Never,* and 10 percent *Always.* Then if we randomly divide the evaluation sample into a treatment group and a comparison group, we should again have approximately 80 percent *Enroll-if-offered,* 10 percent *Never,* and 10 percent *Always* in both groups. In the group that is offered treatment, the *Enroll-if-offered* and *Always* individuals will enroll, and only the *Never* people will stay away. In the group that is not offered treatment, the *Always* will enroll, while the *Enroll-if-offered* and *Never* groups will stay out.

Estimating Impact under Randomized Offering

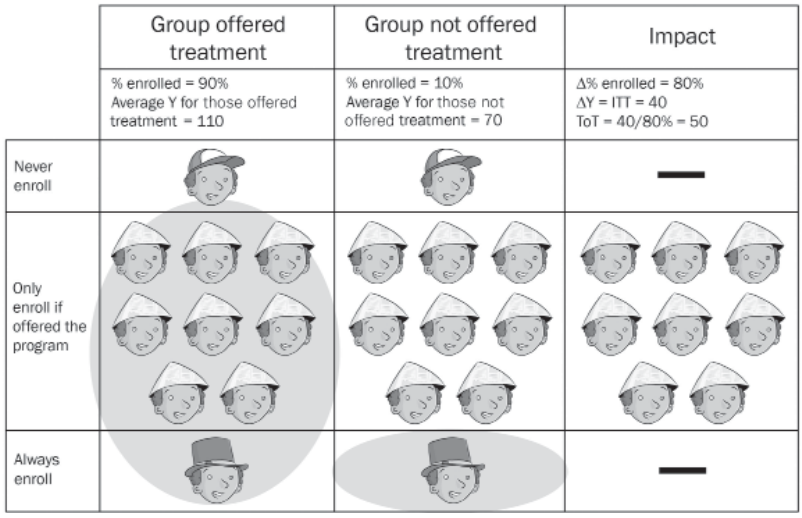
Having established the difference between offering a program and actual enrollment or take-up, we turn to a technique that can be used to estimate the impact of treatment on the treated, that is, the impact of the program on

Figure 4.6 Randomized Offering of a Program



Source: Authors.

Figure 4.7 Estimating the Impact of Treatment on the Treated under Randomized Offering



Source: Authors.

Note: ITT is the "intention-to-treat" estimate obtained by comparing outcomes for those to whom treatment was offered with those to whom treatment was not offered (irrespective of actual enrollment). TOT is the "treatment-on-the-treated" estimate, i.e., the impact of the program estimated on those who were offered treatment and who actually enroll. Characters on shaded background are those that actually enroll.

those who were offered treatment *and* who actually enroll. This estimation is done in two steps, which are illustrated in figure 4.7.[10]

First, we estimate the impact of intention to treat. Remember that this is just the straight difference in the outcome indicator (*Y*) for the group to whom we offered treatment and the same indicator for the group to whom we did not offer treatment. For example, if the average income (*Y*) for the treatment group is $110, and the average income for the comparison group is $70, then the intention-to-treat estimate of the impact (ITT) would be $40.

Second, we need to recover the treatment-on-the-treated estimate (TOT) from the intention-to-treat estimate. To do that, we will need to identify where the $40 difference came from. Let us proceed by elimination. First, we know that the difference cannot be caused by any differences between the *Never*s in the treatment and comparison groups. The reason is that the *Nevers* never enroll in the program, so that for them, it makes no difference whether they are in the treatment group or in the comparison group. Second, we know that the $40 difference cannot be caused by differences between the *Always* people in the treatment and comparison groups because the *Always* people always enroll in the program. For them, too, it makes no difference whether they are in the treatment group or the comparison group. Thus, the difference in outcomes between the two groups must necessarily come from the effect of the program on the only group affected by their assignment to treatment or comparison, that is, the *Enroll-if-offered* group. So if we can identify the *Enroll-if-offered* in both groups, it will be easy to estimate the impact of the program on them.

In reality, although we know that these three types of individuals exist in the population, we cannot uniquely separate out individuals by whether they are *Enroll-if-offered, Never,* or *Always*. In the group that was offered treatment, we can identify the *Nevers* (because they have not enrolled), but we cannot differentiate between the *Always* and the *Enroll-if-offered* (because both are enrolled). In the group that was not offered treatment, we can identify the *Always* group (because they enroll in the program), but we cannot differentiate between the *Nevers* and the *Enroll-if-offered*.

However, once we observe that 90 percent of units in the group offered treatment enroll, we can deduce that 10 percent of the units in our population must be *Nevers* (that is the fraction of individuals in the group offered treatment that did not enroll). In addition, if we observe that 10 percent of units in the group not offered treatment enroll, we know that 10 percent are *Always* (again, the fraction of individuals in our group that was not offered treatment who did enroll). This leaves 80 percent of the units in the *Enroll-if-offered* group. We know that the entire impact of $40 came from a difference in enrollment for the 80 percent of the units in our sample who are

Enroll-if-offered. Now if 80 percent of the units are responsible for an average impact of $40 for the entire group offered treatment, then the impact on those 80 percent of *Enroll-if-offered* must be 40/0.8, or $50. Put another way, the impact of the program for the *Enroll-if-offered* is $50, but when this impact is spread across the entire group offered treatment, the average effect is watered down by the 20 percent that was noncompliant with the original randomized assignment.

Remember that one of the basic issues with self-selection into programs is that you cannot always know why some people choose to participate and others do not. When we randomly assign units to the program, but actual participation is voluntary or a way may exist for units in the comparison group to get into the program, then we have a similar problem: we will not always understand the behavioral processes that determine whether an individual behaves like a *Never*, an *Always*, or an *Enroll-if-offered* in our example above. However, provided that the noncompliance is not too large, the initial randomized assignment still provides a powerful tool for estimating impact. The downside of randomized assignment with imperfect compliance is that this impact estimate is no longer valid for the entire population. Instead, it applies only to a specific subgroup within our target population, the *Enroll-if-offered*.

Randomized offering of a program has two important characteristics that allow us to estimate impact even without full compliance (see box 4.2):[11]

1. It can serve as a predictor of actual enrollment in the program if most people behave as *Enroll-if-offered*, enrolling in the program when offered treatment and not enrolling when not offered treatment.

2. Since the two groups (offered and not offered treatment) are generated through a random selection process, the characteristics of individuals in the two groups are not correlated with anything else, such as ability or motivation, that may also affect the outcomes (Y).

Randomized Promotion or Encouragement Design

In the previous section, we saw how to estimate impact based on randomized assignment of treatment, even when compliance with the originally assigned treatment and comparison groups is incomplete. Next we propose a very similar approach that can be applied to evaluate programs that have universal eligibility or open enrollment or in which the program administrator cannot control who participates and who does not.

Governments commonly implement programs for which it is difficult either to exclude any potential participants or to force them to participate. Many programs allow potential participants to choose to enroll and are

Box 4.2: Randomized Offering of School Vouchers in Colombia

The Program for Extending the Coverage of Secondary School (Programa de Ampliación de Cobertura de la Educación Secundaria [PACES]), in Colombia, provided more than 125,000 students with vouchers covering slightly over half the cost of attending private secondary school. Because of the limited PACES budget, the vouchers were allocated via a lottery. Angrist et al. (2002) took advantage of this randomly assigned treatment to determine the effect of the voucher program on educational and social outcomes.

They found that lottery winners were 10 percent more likely to complete the 8th grade and scored, on average, 0.2 standard deviations higher on standardized tests three years after the initial lottery. They also found that the educational effects were greater for girls than boys. The researchers then looked at the impact of the program on several noneducational outcomes and found that lottery winners were less likely to be married and worked about 1.2 fewer hours per week.

There was some noncompliance with the randomized design, in that about 90 percent of the lottery winners had actually used the voucher or another form of scholarship and 24 percent of the lottery losers had actually received scholarships. Angrist and colleagues therefore also used intent-to-treat, or a student's lottery win or loss status, as an instrumental variable for the treatment-on-the-treated, or actual scholarship receipt. Finally, the researchers were able to calculate a cost-benefit analysis to better understand the impact of the voucher program on both household and government expenditures. They concluded that the total social costs of the program are small and are outweighed by the expected returns to participants and their families, thus suggesting that demand-side programs such as PACES can be a cost-effective way to increase educational attainment.

Source: Angrist et al. 2002.

not, therefore, able to exclude potential participants who want to enroll. In addition, some programs have a budget that is big enough to supply the program to the entire eligible population immediately, so that randomly choosing treatment and comparison groups and excluding potential participants for the sake of an evaluation would not be ethical. We therefore need an alternative way to evaluate the impact of these kinds of programs—those with voluntary enrollment and those with universal coverage.

Voluntary enrollment programs typically allow individuals who are interested in the program to approach on their own to enroll and participate. Imagine again the job training program discussed earlier, but this time randomized assignment is not possible, and any individual who wishes to enroll in the program is free to do so. Very much in line with our previous example, we will expect to encounter three types of people: compliers, a *Never* group, and an *Always* group. As in the previous case, *Always* people

will always enroll in the program and *Never* people will never enroll. But how about the compliers? In this context, any individual who would like to enroll in the program is free to do so. And what about individuals who may be very interested in enrolling but who, for a variety of reasons, may not have sufficient information or the right incentive to enroll? The compliers in this context will be precisely that group. The compliers here are those who *enroll-if-promoted*: they are a group of individuals who only enroll in the program if given an additional incentive, or promotion, that motivates them to enroll. Without this additional stimulus, the *Enroll-if-promoted* would simply remain out of the program.

Again coming back to the job training example, if the agency that organizes the training is well funded and has sufficient capacity to train everyone who wants to be trained, then the job training program will be open to every unemployed person who wants to participate. It is unlikely, however, that every unemployed person will actually want to participate or will even know of the existence of the program. Some unemployed people may be reluctant to enroll because they know very little about the content of the training and find it hard to obtain additional information. Now assume that the job training agency hires a community outreach worker to go around town to enlist unemployed persons into the job training program. Carrying a list of unemployed people, she knocks on their doors, describes the training program, and offers to help the person to enroll in the program on the spot. Of course, she cannot force anyone to participate. In addition, the unemployed persons whom the outreach worker does not visit can also enroll, although they will have to go to the agency themselves to do so. So we now have two groups of unemployed people—those who were visited by the outreach worker and those who were not visited. If the outreach effort is effective, the enrollment rate among unemployed people who were visited should be higher than the rate among unemployed people who were not visited.

Now let us think about how we can evaluate this job training program. As we know, we cannot just compare those unemployed people who enroll with those who do not enroll. The reason is that the unemployed who enroll are probably very different from those who do not enroll in both observed and nonobserved ways: they may be more educated (this can be observed easily), and they are probably more motivated and eager to find a job (this is hard to observe and measure).

However, we do have some additional variation that we can exploit to find a valid comparison group. Let us consider for a moment whether we can compare the group that was visited by the outreach worker with the group that was not visited. Both groups contain very motivated persons

(Always) who will enroll whether or not the outreach worker knocks on their door. Both groups also contain unmotivated persons *(Never)* who will not enroll in the program despite the efforts of the outreach worker. And finally, some people (*Enroll-if-promoted*) will enroll in the training if the outreach worker visits them but will not enroll if the worker does not come knocking.

If the outreach worker randomly selected the people on her list to visit, we would be able to use the treatment-on-the-treated method discussed earlier. The only difference would be that, instead of randomly *offering* the program, we would be randomly *promoting* it. As long as *Enroll-if-promoted* people (who enroll when we reach out to them but do not enroll when we do not reach out to them) appear, we would have a variation between the group *with* the promotion or outreach and the group *without* the promotion or outreach that would allow us to identify the impact of the training on the *Enroll-if-promoted*. Instead of complying with the offer of the treatment, the *Enroll-if-promoted* are now complying with the promotion.

We want the outreach strategy to be effective and to increase enrollment substantially among the *Enroll-if-promoted* group. At the same time, we do not want the promotion activities to be so widespread and effective that they influence the outcome of interest. For example, if the outreach workers offered large amounts of money to unemployed people to get them to enroll, it would be hard to tell whether any later changes in income were caused by the training or by the outreach or promotion itself.

Randomized promotion is a creative strategy that generates the equivalent of a comparison group for the purposes of impact evaluation. It can be used when it is feasible to organize a promotion campaign aimed at a random sample of the population of interest. Readers with a background in econometrics may again recognize the terminology introduced in the previous section: the randomized promotion is an instrumental variable that allows us to create variation between units and exploit that variation to create a valid comparison group.

You Said "Promotion"?

Randomized promotion seeks to increase the take-up of a voluntary program in a subsample of the population. It can take several forms. For instance, we may choose to initiate an information campaign to reach those individuals who had not enrolled because they did not know or fully understand the content of the program. Alternatively, we may choose to provide incentives to sign up, such as offering small gifts or prizes or making transportation or other help available.

A number of conditions must be met for the randomized promotion methodology to produce a valid impact evaluation.

1. The promoted and nonpromoted groups must be comparable. The characteristics of the two groups must be similar. This is achieved by randomly assigning the outreach or promotion activities among the units in the evaluation sample.
2. The promotion campaign must increase enrollment by those in the promoted group substantially above the rate of the nonpromoted group. This can be verified by checking that enrollment rates are higher in the group that receives the promotion than in the group that does not.
3. It is important that the promotion itself does not directly affect the outcomes of interest, so that we can tell that changes in the outcomes of interest are caused by the program itself and not by the promotion.

Key Concept:
Randomized promotion is a method similar to randomized offering. Instead of randomly selecting units to whom we offer the treatment, we randomly select units to whom we promote the treatment. In this way, we can leave the program open to every unit.

The Randomized Promotion Process

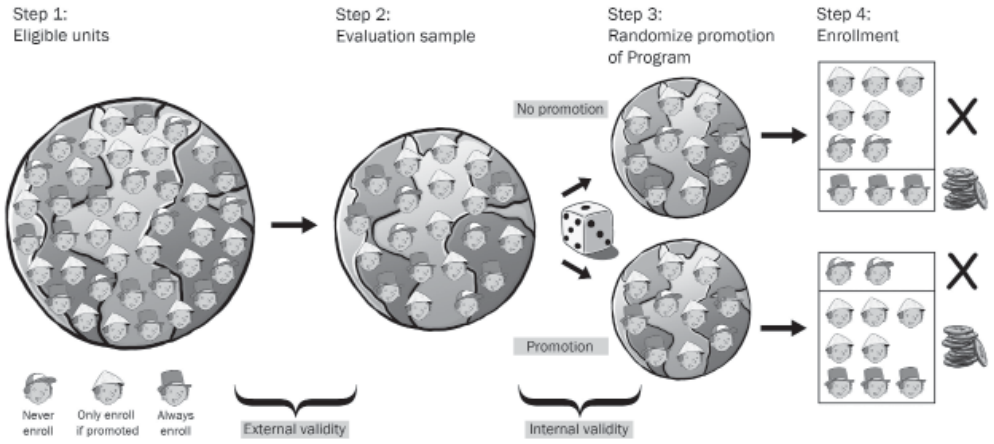
The process of randomized promotion is presented in figure 4.8. As in the previous methods, we begin with the population of eligible units for the program. In contrast with randomized assignment, we can no longer randomly choose who will receive the program and who will not receive the program because the program is fully voluntary. However, within the population of eligible units, there will be three types of units:

- *Always*—those who will always want to enroll in the program
- *Enroll-if-promoted*—those who will sign up for the program only when given additional promotion
- *Never*—those who never want to sign up for the program, whether or not we offer them promotion

Again, note that being an *Always,* an *Enroll-if-promoted,* or a *Never* is an intrinsic characteristic of units that cannot be measured by the program evaluator because it is related to factors such as intrinsic motivation and intelligence.

Once the eligible population is defined, the next step is to randomly select a sample from the population to be part of the evaluation. These are the units on whom we will collect data. In some cases—for example, when we have data for the entire population of eligible units—we may decide to include this entire population in the evaluation sample.

Figure 4.8 Randomized Promotion



Source: Authors.

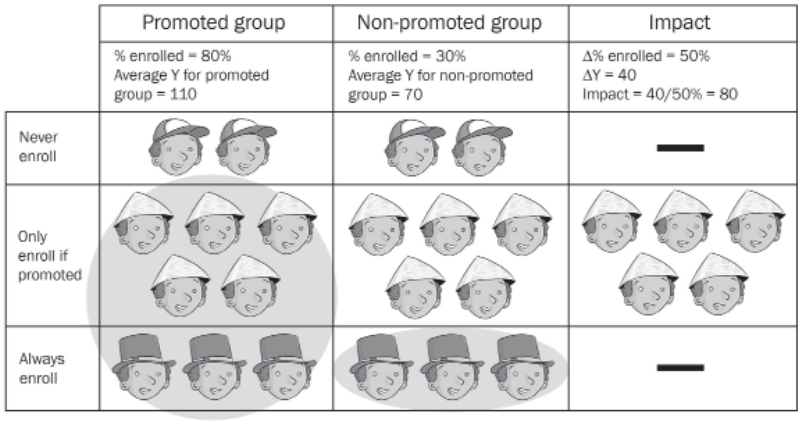
Once the evaluation sample is defined, randomized promotion randomly assigns the evaluation sample into a promoted group and a nonpromoted group. Since we are randomly choosing the members of both the promoted group and the nonpromoted group, both groups will share the characteristics of the overall evaluation sample, and those will be equivalent to the characteristics of the population of eligible units. Therefore, the promoted group and the nonpromoted group will have similar characteristics.

After the promotion campaign is over, we can observe the enrollment rates in the promoted and nonpromoted groups. In the nonpromoted group, only the *Always* will enroll. Although we thus will be able to know which units are *Always* in the nonpromoted group, we will not be able to distinguish between the *Never* and *Enroll-if-promoted* in that group. By contrast, in the promoted group both the *Enroll-if-promoted* and the *Always* will enroll, whereas the *Never* will not enroll. So in the promoted group we will be able to identify the *Never* group, but we will not be able to distinguish between the *Enroll-if-promoted* and the *Always*.

Estimating Impact under Randomized Promotion

Estimating the impact of a program using randomized promotion is a special case of the treatment-on-the-treated method (figure 4.9). Imagine that the promotion campaign raises enrollment from 30 percent in the nonpromoted group (3 *Always*) to 80 percent in the promoted group (3 *Always* and 5 *Enroll-if-promoted*). Assume that average outcome for all individuals in

Figure 4.9 Estimating Impact under Randomized Promotion



Source: Authors.

Note: Characters on shaded background are those that enroll.

the nonpromoted group (10 individuals) is 70, and that average outcome for all individuals in the promoted group (10 individuals) is 110. Then what would the impact of the program be?

First, we can compute the straight difference between the promoted and the nonpromoted groups, which is 40. We also know that none of this difference of 40 comes from the *Nevers* because they do not enroll in either group. We also know that none of this difference of 40 comes from the *Enroll-if-promoted* because they enroll in both groups.

The second step is to recover the impact that the program has had on the *Enroll-if-promoted*. We know the entire average effect of 40 can be attributed to the *Enroll-if-promoted*, who make up only 50 percent of the population. To assess the average effect of the program on a complier, we divide 40 by the percentage of *Enroll-if-promoted* in the population. Although we cannot directly identify the *Enroll-if-promoted*, we are able to deduce what must be their *percentage* of the population: it is the difference in the enrollment rates of the promoted and the nonpromoted groups (50 percent or 0.5). Therefore, the average impact of the program on a complier is 40/0.5 = 80

Given that the promotion is assigned randomly, the promoted and nonpromoted groups have equal characteristics, on average. Thus, the differences that we observe in average outcomes between the two groups must be caused by the fact that in the promoted group the *Enroll-if-promoted* enroll, while in the nonpromoted group they do not.[12]

Using Randomized Promotion to Estimate the Impact of the Health Insurance Subsidy Program

Let us now try using the randomized promotion method to evaluate the impact of the HISP. Assume that the ministry of health makes an executive decision that the health insurance subsidy should be made available immediately to any household that wants to enroll. However, you know that realistically this national scale-up will be incremental over time, and so you reach an agreement to accelerate enrollment in a random subset of villages through a promotion campaign. You undertake an intensive promotion effort in a random subsample of villages, including communication and social marketing campaigns aimed at increasing awareness of the HISP. After two years of promotion and program implementation, you find that 49.2 percent of households in villages that were randomly assigned to the promotion have enrolled in the program, while only 8.4 percent of households in nonpromoted villages have enrolled (table 4.4).

Because the promoted and nonpromoted villages were assigned at random, you know that the average characteristics of the two groups should be the same in the absence of the program. You can verify that assumption by comparing the baseline health expenditures (as well as any other characteristics) of the two populations. After two years of program implementation, you observe that the average health expenditure in the promoted villages is $14.9 compared with $18.8 in nonpromoted areas (a difference of minus $3.9). However, because the only difference between the promoted and nonpromoted villages is that promoted villages have greater enrollment in the program (thanks to the promotion), this difference of $3.9 in health expenditures must be due to the 40.4 percent of households that enrolled in the promoted villages because of the promotion. Therefore, we need to adjust

Table 4.4 Case 4—HISP Impact Using Randomized Promotion (Comparison of Means)

	Promoted villages	Nonpromoted villages	Difference	*t*-stat
Household health expenditures baseline	17.1	17.2	–0.1	–0.47
Household health expenditures follow-up	14.9	18.8	–3.9	–18.3
Enrollment in HISP	49.2%	8.4%	40.4%	

Source: Authors' calculation.

** Significant at the 1 percent level.

Table 4.5 Case 4—HISP Impact Using Randomized Promotion (Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	–9.4** (0.51)	–9.7** (0.45)

Source: Authors' calculation.

Note: Standard errors are in parentheses.

** Significant at the 1 percent level.

the difference in health expenditures to be able to find the impact of the program on the *Enroll-if-promoted*. To do this, we divide the straight difference between the promoted groups by the percentage of *Enroll-if-promoted:* –3.9/0.404 = –$9.65. Your colleague, who took an econometrics class, then estimates the impact of the program through two-stage least squares and finds the results shown in table 4.5. This estimated impact is valid for those households that enrolled in the program because of the promotion but who otherwise would not have done so, in other words, for the *Enroll-if-promoted*. To extrapolate this result for the full population, we must assume that all other households would have reacted in a similar way had they enrolled in the program.

QUESTION 4

A. What are the basic assumptions required to accept the result from case 4?

B. Based on the result from case 4, should the HISP be scaled up nationally?

Randomized Promotion at Work

The randomized promotion method can be used in various settings. Gertler, Martinez, and Vivo (2008) used it to evaluate a maternal and child health insurance program in Argentina. Following the 2001 economic crisis, the government of Argentina observed that the population's health indicators had started deteriorating and, in particular, that infant mortality was increasing. It decided to introduce a national insurance scheme for mothers and their children, which was to be scaled up to the entire country within a year. Still, government officials wanted to evaluate the impact of the program to make sure that it was really improving the health status of the population. How could a comparison group be found if every mother and child in the country was entitled to enroll in the insurance scheme if they so desired? Data for the first provinces implementing the intervention showed that only

40 percent to 50 percent of households were actually enrolling in the program. So the government launched an intensive promotion campaign seeking to inform households about the program. However, the promotion campaign was implemented only in a random sample of villages, not in the entire country.

Other examples include assistance from nongovernmental organizations in a community-based school management evaluation, in Nepal, and the Bolivian Social Investment Fund (detailed in box 4.3).

Limitations of the Randomized Promotion Method

Randomized promotion is a useful strategy for evaluating the impact of voluntary programs and programs with universal eligibility, particularly because it does not require the exclusion of any eligible units. Nevertheless, the approach has some noteworthy limitations compared to randomized assignment of the treatment.

First, the promotion strategy must be effective. If the promotion campaign does not increase enrollment, then no difference between the pro-

Box 4.3: Promoting Education Infrastructure Investments in Bolivia

In 1991, Bolivia institutionalized and scaled up a successful Social Investment Fund (SIF) which provided financing to rural communities to carry out small-scale investments in education, health, and water infrastructure. The World Bank, which was helping to finance SIF, was able to build an impact evaluation into the program design.

As part of the impact evaluation of the education component, communities in the Chaco region were randomly selected for active promotion of the SIF intervention and received additional visits and encouragement to apply from program staff. The program was open to all eligible communities in the region and was demand driven in that communities had to apply for funds for a specific project. Not all communities took up the program, but take-up was higher among promoted communities.

Newman et al. (2002) used the randomized promotion as an instrumental variable. They found that the education investments succeeded in improving measures of school infrastructure quality such as electricity, sanitation facilities, textbooks per student, and student-teacher ratios. However they detected little impact on educational outcomes, except for a decrease of about 2.5 percent in the dropout rate. As a result of these findings, the ministry of education and the SIF now focus more attention and resources on the "software" of education, funding physical infrastructure improvements only when they form part of an integrated intervention.

Source: Newman et al. 2002.

moted and the nonpromoted groups will appear, and there will be nothing to compare. It is thus crucial to pilot the promotion campaign extensively to make sure that it will be effective. On the positive side, the design of the promotion campaign can help program managers by teaching them how to increase enrollment.

Second, the methodology estimates the impact of the program only for a subset of the population of eligible units. Specifically, the program's average impact is computed from the group of individuals who sign up for the program only when encouraged to do so. However, individuals in this group may have very different characteristics than those individuals who always or never enroll, and therefore the average treatment effect for the entire population may be different from the average treatment effect estimated for individuals who participate only when encouraged.

Notes

1. Randomized assignment of treatment is also commonly referred to as "randomized control trials," "randomized evaluations," "experimental evaluations," and "social experiments," among other terms.
2. Note that this probability does not necessarily mean a 50-50 chance of winning the lottery. In fact, most randomized assignment evaluations will give each eligible unit a probability of selection that is determined so that the number of winners (treatments) equals the total available number of benefits. For example, if a program has enough funding to serve only 1,000 communities, out of a population of 10,000 eligible communities, then each community will be given a chance of 1 in 10 of being selected for treatment. Statistical power (a concept discussed in more detail in chapter 11) will be maximized when the evaluation sample is divided equally between the treatment and control groups. In the example here, for a total sample size of 2,000 communities, statistical power will be maximized by sampling all 1,000 treatment communities and a subsample of 1,000 control communities, rather than by taking a simple random sample of 20 percent of the original 10,000 eligible communities (which would produce an evaluation sample of roughly 200 treatment communities and 1,800 control communities).
3. For example, housing programs that provide subsidized homes routinely use lotteries to select program participants.
4. This property comes from the Law of Large Numbers.
5. An evaluation sample can be stratified by population subtypes and can also be clustered by sampling units. The sample size will depend on the particular type of random sampling used (see part 3).
6. Most software programs allow you to set a "seed number" to make the results of the randomized assignment fully transparent and replicable.
7. We will discuss concepts such as spillovers or contamination in more detail in chapter 8.

8. For statistical reasons, not all observed characteristics have to be similar in the treatment and comparison groups for randomization to be successful. As a rule of thumb, randomization will be considered successful if about 95 percent of the observed characteristics are similar. By "similar," we mean that we cannot reject the null hypothesis that the means are different between the two groups when using a 95 percent confidence interval. Even when the characteristics of the two groups are truly equal, one can expect that about 5 percent of the characteristics will show up with a statistically significant difference.
9. Note that in the medical sciences, patients in the comparison group typically receive a placebo, that is, something like a sugar pill that should have no effect on the intended outcome. That is done to additionally control for the "placebo effect," meaning the potential changes in behavior and outcomes from receiving a treatment, even if the treatment itself is ineffective.
10. These two steps correspond to the econometric technique of two-stage-least-squares, which produces a local average treatment effect.
11. Readers with a background in econometrics may recognize the concept: in statistical terms, the randomized offering of the program is used as an instrumental variable for actual enrollment. The two characteristics listed are exactly what would be required from a good instrumental variable:
    - The instrumental variable must be correlated with program participation.
    - The instrumental variable may not be correlated with outcomes (Y) (except through program participation) or with unobserved variables.
12. Again, readers familiar with econometrics may recognize that the impact is estimated by using "randomized assignment to the promoted and nonpromoted groups" as an instrumental variable for actual enrollment in the program.

References

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92 (5): 1535–58.

Gertler, Paul, Sebastian Martinez, and Sigrid Vivo. 2008. "Child-Mother Provincial Investment Project *Plan Nacer*." University of California Berkeley and World Bank, Washington, DC.

Newman, John, Menno Pradhan, Laura B. Rawlings, Geert Ridder, Ramiro Coa, and Jose Luis Evia. 2002. "An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund." *World Bank Economic Review* 16 (2): 241–74.

Schultz, Paul. 2004. "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program." *Journal of Development Economics* 74 (1): 199–250.