CHAPTER 3

Causal Inference and Counterfactuals

Causal Inference

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

Many 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?

Impact evaluations seek to answer such cause-and-effect questions precisely. Assessing the impact of a program on a set of outcomes is the equivalent of assessing the causal effect of the program on those outcomes.[1]

Although cause-and-effect questions are common, answering them accurately can be challenging. In the context of a vocational training program, for example, simply observing that a trainee's income increases after she has completed such a program is not sufficient to establish causality. The trainee's income might have increased even if she had not taken the training—because of her own efforts, because of changing labor market conditions, or because of many other factors that can affect income. Impact evaluations help us overcome the challenge of establishing causality by empirically establishing to what extent a particular program—*and that program alone*—contributed to

Key Concept

Impact evaluations establish the extent to which a program—and that program alone—caused a change in an outcome.

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:

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

This formula states that the causal impact (Δ) 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 (Δ) is the difference between a person's income (Y) after participation 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 the change in income is causal.

The basic impact evaluation formula is valid for any unit that is being analyzed—a person, a household, a community, a business, a school, a hospital, or other unit of observation that may receive or be affected by a program. The formula is also valid for any outcome (Y) that is 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, we can think of the impact (Δ) of a program as the difference in outcomes (Y) for the same unit (person, household, community, and so on) with and without participation in a program. Yet we know that measuring the same unit in two different states at the same time is impossible. At any given moment in time, a unit either participated in the program or did not participate. The unit 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 a program participant ($Y \mid P = 1$), there are no data to establish what her outcome 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* to the outcome if a person or unit of observation had not participated in the program.

Key Concept

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

When conducting an impact evaluation, it is relatively straightforward to obtain the first term of the basic formula ($Y \mid P = 1$)—the outcome with a program (also known as *under treatment*). We simply measure the outcome of interest for the program participant. However, we cannot directly observe the second term of the formula ($Y \mid P = 0$) for the participant. We need to fill in this missing piece of information by *estimating the counterfactual*.

Key Concept

Since we cannot directly observe the counterfactual, we must estimate it.

To help us think through this key concept of estimating the counterfactual, we turn to another hypothetical example. Solving the counterfactual problem would be possible if the evaluator could find a "perfect clone" for a program participant (figure 3.1). For example, let us say that Mr. Fulanito starts receiving US$12 in pocket money allowance, and we want to measure the impact of this treatment on his consumption of candies. If you could identify a perfect clone for Mr. Fulanito, the evaluation would be easy: you could

Box 3.1: The Counterfactual Problem: "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 of the cash transfer, the government selects height as an outcome indicator for long-term health.

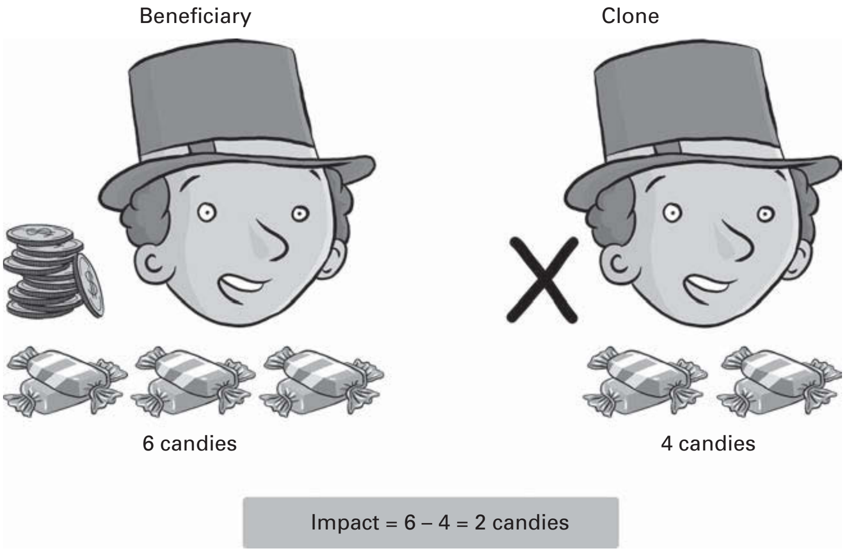
Assume that you are able to measure Miss Unique's height at the age of 3. Ideally, to evaluate the impact of the program, you would want to measure Miss Unique's height at the age of 3 with her mother having received the cash transfer, and also Miss Unique's height at the age of 3 had her mother not received the cash transfer. You would then compare the two heights to establish impact. 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 cash transfer 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 follows the conditions (checkups, immunizations, growth monitoring) and receives the cash transfer or it does not. In other words, we cannot observe what the counterfactual is. Since Miss Unique's mother actually followed the conditions and received the cash transfer, we cannot know how tall Miss Unique would have been had her mother not received the cash transfer.

Finding an appropriate comparison for Miss Unique will be challenging because she is, precisely, unique. Her exact socioeconomic background, genetic attributes, and personal and household 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 cannot be exactly 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.

just compare the number of candies eaten by Mr. Fulanito (say, 6) when he receives the pocket money with the number of candies eaten by his clone (say, 4), who receives no pocket money. In this case, the impact of the pocket money would be 2 candies: the difference between the number of candies consumed under treatment (6) and the number of candies consumed

Figure 3.1 The Perfect Clone



without treatment (4). In reality, we know that it is impossible to identify perfect clones: even between genetically identical twins, there are important differences.

Estimating the Counterfactual

The key to estimating the counterfactual for program participants is to move from the individual or unit level to the group level. Although no perfect clone exists for a single unit, we can rely on statistical properties to generate two *groups* of units that, if their numbers are large enough, are statistically indistinguishable from each other at the group level. The group that participates in the program is known as the *treatment group,* and its outcome is $(Y \mid P = 1)$ after it has participated in the program. The statistically identical *comparison group* (sometimes called the *control group*) is the group that remains unaffected by the program, and allows us to estimate the counterfactual outcome $(Y \mid P = 0)$: that is, the outcome that would have prevailed for the treatment group had it not received the program.

So in practice, the challenge of an impact evaluation is to identify a treatment group and a comparison group that are statistically identical, on average, in the absence of the program. If the two groups are identical, with the sole exception 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. Finding such comparison groups is the crux of any impact evaluation, regardless of what type of program is being evaluated. Simply put, without a comparison group that yields an accurate estimate of the counterfactual, the true impact of a program cannot be established.

Key Concept

Without a comparison group that yields an accurate estimate of the counterfactual, the true impact of a program cannot be established.

The main challenge for identifying impacts, then, is to find *a valid comparison group* that has the same characteristics as the treatment group in the absence of a program. Specifically, the treatment and comparison groups must be the same in at least three ways.

First, the average characteristics of the treatment group and the comparison group must be identical in the absence of the program.[2] Although it is not necessary that individual units in the treatment group have "perfect clones" in the comparison group, *on average* the characteristics of treatment and comparison groups should be the same. For example, the average age of units in the treatment group should be the same as in the comparison group.

Second, the treatment should not affect the comparison group either directly or indirectly. In the pocket money example, the treatment group should not transfer resources to the comparison group (direct effect) or affect the price of candy in the local markets (indirect effect). For example, if we want to isolate the impact of pocket money on candy consumption, the treatment group should not also be offered more trips to the candy store than the comparison group; otherwise, we would be unable to distinguish whether additional candy consumption is due to the pocket money or to the extra trips to the store.

Key Concept

A valid comparison group (1) has the same characteristics, on average, as the treatment group in the absence of the program; (2) remains unaffected by the program; and (3) would react to the program in the same way as the treatment group, if given the program.

Third, the outcomes of units in the control group should change the same way as outcomes in the treatment group, if both groups were given the program (or not). In this sense, the treatment and comparison groups should react to the program in the same way. For example, if incomes of people in the treatment group increased by US$100 thanks to a training program, then incomes of people in the comparison group would have also increased by US$100, had they been given training.

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. This is because the only difference between the treatment and comparison groups is that the members of the treatment group receive the program, while the members of the comparison group do not. When the difference in outcome can be entirely attributed to the program, the causal impact of the program has been identified.

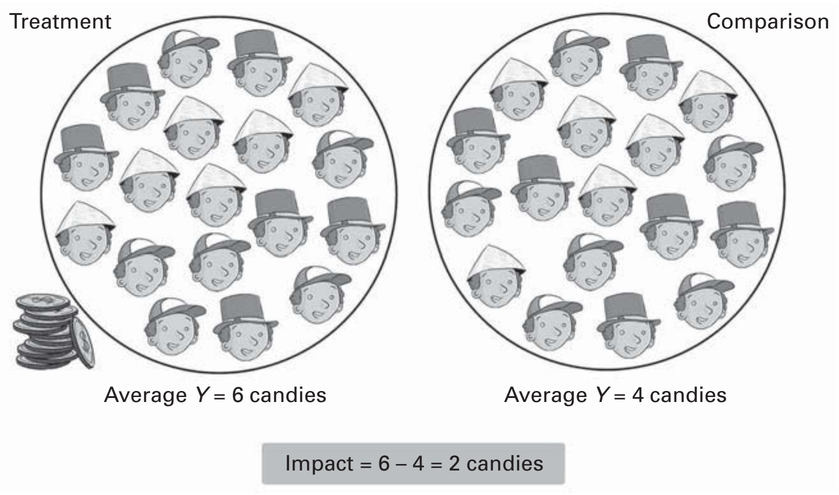
Returning to the case of Mr. Fulanito, we saw that in order to estimate the impact of pocket money on his consumption of candies would require the implausible task of finding Mr. Fulanito's perfect clone. Instead of looking at the impact solely for one individual, it is more realistic to look at the average impact for a group of individuals (figure 3.2). If you could identify another group of individuals that shares the same average age, gender composition, education, preference for candy, and so on, except that it does not receive additional pocket money, then you could estimate the pocket money's impact. This would simply be the difference between the average consumption of candies in the two groups. Thus if the *treatment 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 candies.

Having defined a *valid comparison group*, it is important to consider what would happen if we decided to go ahead with an evaluation without finding such a group. Intuitively, 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 the estimate of impact to be invalid or, in statistical terms, *biased*: the impact evaluation will not estimate the true impact of the program. Rather, it will estimate the effect of the program mixed with those other differences.

Key Concept

When the comparison group does not accurately estimate the true counterfactual, then the estimated impact of the program will be invalid. In statistical terms, it will be *biased*.

Figure 3.2 A Valid Comparison Group



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 many times lead to inappropriate ("counterfeit") estimates of the counterfactual:

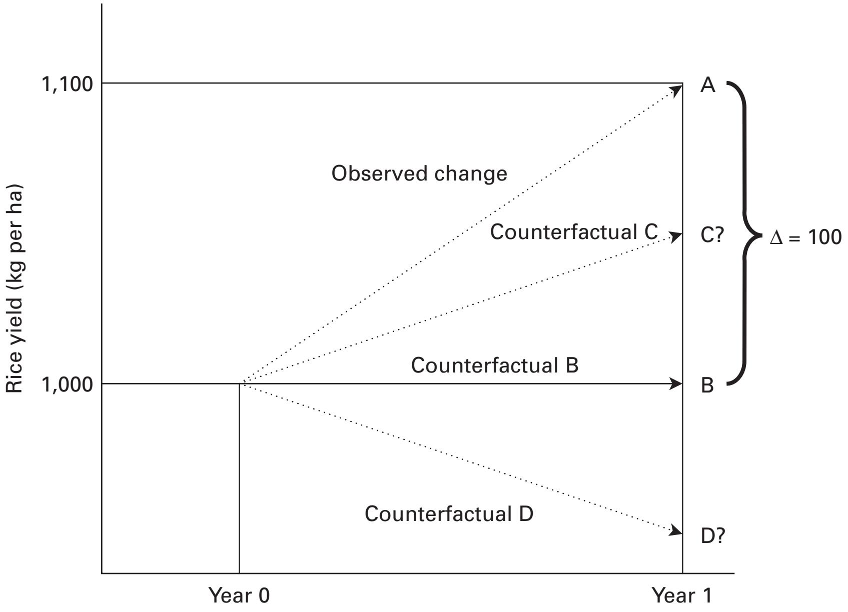
- *Before-and-after comparisons* (also known as *pre-post* or *reflexive comparisons*) compare the outcomes of the same group before and after participating in a program.
- *Enrolled-and-nonenrolled* (or *self-selected*) *comparisons* compare the outcomes of a group that chooses to participate in a program with those of a group that chooses not to participate.

Counterfeit Counterfactual Estimate 1: Comparing Outcomes Before and After a Program

A before-and-after comparison attempts to establish the impact of a program by tracking changes in outcomes for program participants over time. Returning to the basic impact evaluation formula, the outcome for the treatment group ($Y \mid P = 1$) is simply the outcome after participating in the program. However, before-and-after comparisons take the estimated counterfactual ($Y \mid P = 0$) as the outcome for the treatment group *before* the intervention started. 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 situation before the program. Unfortunately, for a majority of programs implemented over a series of months or years, this assumption simply does not hold.

Consider the evaluation of a microfinance program for poor, rural farmers. 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 program starts, farmers harvested an average of 1,000 kilograms (kg) of rice per hectare (point B in figure 3.3). The microfinance scheme is launched, and a year later rice yields have increased to 1,100 kg per hectare (point A in figure 3.3). If you were trying to evaluate impact using a before-and-after comparison, you would use the baseline outcome as an estimate of the counterfactual. Applying the basic impact evaluation formula, you would conclude that the program had increased rice yields by 100 kg per hectare (A−B).

Figure 3.3 Before-and-After Estimates of a Microfinance Program



Note: *Δ* = Change in rice yield (kg); ha = hectares; kg = kilograms.

However, imagine that rainfall was normal in the year before the program was launched, but a drought occurred in the year the program operated. Because of the drought, the farmers' average yield without the microloan scheme is likely to be lower than *B:* say, at level *D*. In that case, the true impact of the program would be *A*–*D*, which is larger than the 100 kg estimated using the before-and-after comparison. By contrast, if rainfall actually improved between the two years, the counterfactual rice yield might have been at level *C*. In that case, the true program impact would have been smaller than 100 kg. In other words, unless our impact analysis can account for 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.

In the previous microfinance example, rainfall was one of myriad outside factors which might affect the program's outcome of interest (rice yields) over time. Likewise, many of the outcomes that development programs aim to improve, such as income, productivity, health, or education, are affected by an array of factors over time. For that reason, the baseline outcome is almost never a good estimate of the counterfactual. That is why we consider it a counterfeit estimate of the counterfactual.

Evaluating the Impact of HISP: Doing a Before-and-After Comparison of Outcomes

Recall that the Health Insurance Subsidy Program (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 health care and medicine for those enrolled. The objective of HISP is to reduce what poor households spend on primary care and medicine 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 HISP on per capita yearly out-of-pocket expenditures (subsequently referred to simply as *health expenditures*).

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 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 HISP to be viable and to be extended nationally, it must reduce yearly per capita health expenditures of poor rural households by at least US$10 on average, compared to what they would have spent in the absence of the program, and it must do so within two years.

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, HISP is introduced in the 100 pilot villages with great fanfare, including community events and other promotional campaigns to encourage households to enroll.

Of the 4,959 households in the baseline sample, a total of 2,907 enroll in HISP, and the program operates successfully over the next two years. All health clinics and pharmacies serving the 100 villages accept patients with the insurance scheme, and surveys show that most enrolled households are satisfied with the program. At the end of the two-year pilot period, a second round of evaluation data is collected on the same sample of 4,959 households.[3]

The president and the minister of health have put you in charge of overseeing the impact evaluation for HISP and recommending whether or not to extend the program nationally. Your impact evaluation question of interest is, what is the impact of HISP on poor households' out-of-pocket health expenditures? Remember that the stakes are high. If HISP is found to reduce health expenditures by US$10 or more, it will be extended nationally. If the program did not reach the US$10 target, you will recommend against scaling it up.

The first "expert" consultant you hire indicates that to estimate the impact of HISP, you must calculate the change in health expenditures over time for the households that enrolled. The consultant argues that because HISP covers all health costs, any decrease in expenditures over time must be attributable to the effect of HISP. Using the subset of enrolled households, you calculate their average health expenditures before the implementation of the program and then again two years later. In other words, you perform a before-and-after comparison. The results are shown in table 3.1. You observe that the treatment group reduced its out-of-pocket health expenditures by US$6.65, from US$14.49 before the introduction of HISP to US$7.84 two years later. As denoted by the value of the t-statistic (*t-stat*), the difference between health expenditures before and after the program is *statistically significant*.[4] This means that you find strong evidence against the claim that the true difference between expenditures before and after the intervention is zero.

Even though the before-and-after comparison is for the same group of households, you are concerned that other circumstances may have also changed for these households over the past two years, affecting their health expenditures. For example, a number of new drugs have recently become available. You are also concerned that the reduction in health expenditures may have resulted in part from the financial crisis that your country recently experienced. To address some of these concerns, your consultant conducts a more sophisticated *regression analysis* that will try to control for some additional factors.

Table 3.1 Evaluating HISP: Before-and-After Comparison

	After	Before	Difference	*t*-stat
Household health expenditures (US$)	7.84	14.49	−6.65**	−39.76

Note: Significance level: ** = 1 percent.

Table 3.2 Evaluating HISP: Before-and-After with Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures (US$)	−6.65** (0.23)	−6.71** (0.23)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

Regression analysis uses statistics to analyze the relationships between a dependent variable (the variable to be explained) and explanatory variables. The results appear in table 3.2. A linear regression is the simplest form: the dependent variable is health expenditures, and there is only one explanatory variable: a binary (0–1) indicator that takes the value 0 if the observation is taken at baseline and 1 if the observation is taken at follow-up.

A multivariate linear regression adds explanatory variables to *control for*, or *hold constant*, other characteristics that are observed for the households in your sample, including indicators for wealth (assets), household composition, and so on.[5]

You note that the result from the linear regression is equivalent to the simple before-and-after difference in average health expenditures from table 3.1 (a reduction of US$6.65 in health expenditures). Once you use multivariate linear regression to control for other factors available in your data, you find a similar result—a decrease of US$6.71 in health expenditures.

HISP Question 1

A. Does the before-and-after comparison control for all the factors that affect health expenditures over time?

B. Based on these results produced by the before-and-after analysis, should HISP be scaled up nationally?

Counterfeit Counterfactual Estimate 2: Comparing Enrolled and Nonenrolled (Self-Selected) Groups

Comparing a group of individuals that voluntarily signs up for a program to a group of individuals that *chooses* not participate is another risky approach to evaluating impact. A comparison group that *self-selects* out of a program will provide another counterfeit counterfactual estimate. *Selection* occurs when program participation is based on the preferences, decisions, or unobserved characteristics of potential participants.

Consider, for example, a vocational training program for unemployed youth. Assume that two years after the program has been launched, an evaluation attempts to estimate its impact on income by comparing the average incomes of a group of youth who chose to enroll in the program versus a group of youth who, despite being eligible, chose not to enroll. Assume that the results show that youth who chose to enroll in the program make twice as much as those who chose not to 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 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 would perform quite differently in the labor market and would have different incomes even without the vocational training program.

The same issue arises when admission to a program is based on unobserved preferences of program administrators. Say, for example, that the program administrators base admission and enrollment on an interview. Those individuals who are admitted to the program might be those who the administrators think have a good chance of benefiting from the program. Those who are not admitted might show less motivation at the interview, have lower qualifications, or just lack good interview skills. Again, it is likely that these two groups of young people would have different incomes in the labor market even in absence of a vocational training program.

Thus the group that did not enroll does not provide a good estimate of the counterfactual. If you observe a difference in incomes between the two groups, you will not be able to determine whether it comes from the training program or from the underlying differences in motivation, skills, and other factors that exist between the two groups. The fact that less motivated or less qualified individuals did not enroll in the training program therefore leads to a bias in the program's impact.[6] This bias is called *selection bias*. More generally, selection bias will occur when the reasons for which an individual participates in a program are correlated with outcomes, even in absence of the program. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation. In this example, if the young people who enrolled in vocational training would have had higher incomes even in the absence of the program, the selection bias would be positive; in other words, you would overestimate the impact of the vocational training program by attributing to the program the higher incomes that participants would have had anyway.

Key Concept

Selection bias occurs when the reasons for which an individual participates in a program are correlated with outcomes. Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation.

Evaluating the Impact of HISP: Comparing Enrolled and Nonenrolled Households

Having thought through the before-and-after comparison a bit further with your evaluation team, you realize that there are still many other factors that can explain part of the change in health expenditures over time (in particular, the minister of finance is concerned that a recent financial crisis may have affected households' income, and may explain the observed change in health expenditures).

Another consultant suggests that it would be more appropriate to estimate the counterfactual in the post-intervention period: that is, two 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, so approximately 41 percent of the households in the sample remain without HISP coverage. The consultant argues that all households within the 100 pilot villages were eligible to enroll. These households all share the same health clinics and are subject to the same local prices for pharmaceuticals. Moreover, most households are engaged in similar economic activities. The consultant argues that in these circumstances, the outcomes of the nonenrolled group after the intervention could serve to estimate the counterfactual outcome of the group enrolled in HISP. You therefore decide to calculate average health expenditures in the post-intervention period for both the households that enrolled in the program and the households that did not. The results are 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 US$14.46.

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 HISP may be ones that

Table 3.3 Evaluating HISP: Enrolled-Nonenrolled Comparison of Means

	Enrolled	Nonenrolled	Difference	*t*-stat
Household health expenditures (US$)	7.84	22.30	−14.46**	−49.08

Note: Significance level: ** = 1 percent.

Table 3.4 Evaluating HISP: Enrolled-Nonenrolled Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures (US$)	−14.46** (0.33)	−9.98** (0.29)

Note: Standard errors are in parentheses. Significance level: ** = 1 percent.

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 HISP was targeted to poor households. Your consultant argues that regression analysis can control for these potential differences between the two groups. She therefore carries out an additional multivariate regression that controls for all the household characteristics that she can find in the data set, and estimates the impact of the program as shown in table 3.4.

With a simple linear regression of health expenditures on an indicator variable of whether or not a household enrolled in the program, you find an estimated impact of minus US$14.46; in other words, you estimate that the program has decreased average health expenditures by US$14.46. However, when all other characteristics in the data are controlled for, you estimate that the program has reduced health expenditures by US$9.98 per year.

HISP Question 2

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

B. Based on these results produced by the enrolled-nonenrolled method, should HISP be scaled up nationally?

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).

Notes

1. We use the Rubin Causal Model as a framework for causal inference (Imbens and Rubin 2008; Rubin 1974).
2. This condition will be relaxed in some impact evaluation methods, which will require instead that the average *change* in outcomes (trends) is the same in the absence of the program.
3. We are assuming that no households have left the sample over two years (there is zero sample attrition). This is not a realistic assumption for most household surveys. In practice, families that move sometimes cannot be tracked to their new location, and some households break up and cease to exist altogether.
4. Note that a *t*-statistic (*t*-stat) of 1.96 or more (in absolute value) is statistically significant at the 5 percent level.
5. For more on multivariate analysis, see the online technical companion on the Impact Evaluation in Practice website (www.worldbank.org/ieinpractice).
6. 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 comparing them to a group with lower expected returns that does not enroll will yield a biased estimate of impact.

References

Imbens, Guido W., and Donald B. Rubin. 2008. "Rubin Causal Model." In *The New Palgrave Dictionary of Economics*, second edition, edited by Steven N. Durlauf and Lawrence E. Blume. Palgrave.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.