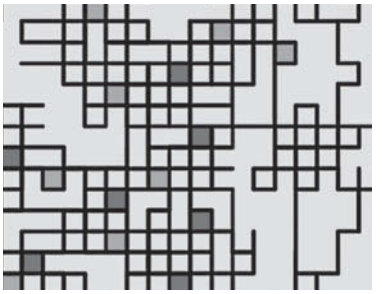
Impact Evaluation in Practice

Second Edition



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

WORLD BANK GROUP

IDB
Inter-American
Development Bank

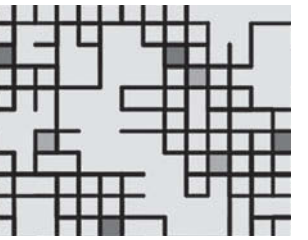
1818 H Street NW, Washington, DC 20433
Telephone: 202-473-1000; Internet: www.worldbank.org

1 2 3 4 19 18 17 16

ISBN (paper): 978-1-4648-0779-4
ISBN (electronic): 978-1-4648-0780-0
DOI: 10.1596/978-1-4648-0779-4

Illustration: C. Andres Gomez-Pena and Michaela Wieser
Cover Design: Critical Stages

Library of Congress Cataloging-in-Publication Data
Names: Gertler, Paul, 1955- author. | World Bank.
Title: Impact evaluation in practice / Paul J. Gertler, Sebastian Martinez, Patrick Premand, Laura B. Rawlings, Christel M. J. Vermeersch.
Description: Second Edition. | Washington, D.C.: World Bank, 2016. | Revised edition of Impact evaluation in practice, 2011.
Identifiers: LCCN 2016029061 (print) | LCCN 2016029464 (ebook) | ISBN 9781464807794 (pdf) | ISBN 9781464807800 | ISBN 9781464807800 ()
Subjects: LCSH: Economic development projects—Evaluation. | Evaluation research (Social action programs)
Classification: LCC HD75.9.G478 2016 (print) | LCC HD75.9 (ebook) | DDC 338.91—dc23
LC record available at https://lccn.loc.gov/2016029061



CHAPTER 5

Instrumental Variables

Evaluating Programs When Not Everyone Complies with Their Assignment

In the discussion of randomized assignment in chapter 4, we 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 are assigned to the treatment and comparison groups comply 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 a given treatment, and second, that none of the subjects in the comparison group take it.[1] More generally in chapter 4, we assumed that programs are able to determine who the potential participants are, excluding some and ensuring that others participate.

However, in real-world social programs, it might be unrealistic to think that the program administrator will be able to ensure full compliance with the group assignment. Yet many programs allow potential participants to choose to enroll and thus are not 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 assigning people to 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.

Key Concept

The instrumental variable method relies on some external source of variation to determine treatment status. An instrumental variable influences the likelihood of participating in a program, but is outside of the participant's control and is unrelated to the participant's characteristics.

A method called *instrumental variables* (IV) can help us evaluate programs with imperfect compliance, voluntary enrollment, or universal coverage. Generally, to estimate impacts, the IV method relies on some external source of variation to determine treatment status. The method has wide-ranging applications beyond impact evaluation. Intuitively, we can think of an IV as something outside the control of the individual that influences her likelihood of participating in a program, but is otherwise not associated with her characteristics.

In this chapter, we discuss how this external variation, or IV, can be generated by the rules of program operation that are under the control of program implementers or evaluation teams. To produce valid impact estimates, this external source of variation must satisfy a number of conditions, which we will discuss in detail in this chapter. It turns out that randomized assignment of treatment, as discussed in chapter 4, is a very good instrument, satisfying the necessary conditions. We will use the IV method in two common applications of impact evaluation. First, we will use it as an extension of the randomized assignment method when not all units comply with their group assignments. Second, we will use it to design randomized promotion of treatment, an evaluation method that can work for some programs that offer voluntary enrollment or universal coverage. Box 5.1 illustrates a creative use of the IV method.

Types of Impact Estimates

An impact evaluation always estimates the impact of a program by comparing the outcomes for a treatment group with the estimate of the counterfactual obtained from a comparison group. In chapter 4, we assumed *full compliance* with treatment: that is, all units to whom a program has been offered actually enroll, and none of the comparison units receive the program. In this scenario, we estimate the *average treatment effect* (ATE) for the population.

In the evaluation of real-world programs where potential participants can decide whether to enroll or not, full compliance is less common than in settings such as laboratory experiments. In practice, programs typically offer the opportunity of treatment to a particular group, and some units participate while others do not. In this case, without full compliance, impact evaluations can estimate the effect of *offering* a program or the effect of *participating* in the program.

Box 5.1: Using Instrumental Variables to Evaluate the Impact of *Sesame Street* on School Readiness

The television show *Sesame Street*, a program aimed at preparing preschool-aged children for primary school, quickly gained critical acclaim and popularity after first airing in 1969. It has since been watched by millions of children. In 2015, Kearney and Levine sought to evaluate the long-term impacts of the program in a retrospective evaluation carried out in the United States. Taking advantage of limitations in television broadcasting technology in the early years of the show, the researchers used an instrumental variables approach.

In the first few years the show was not accessible to all households. It was only broadcast on ultra-high frequency (UHF) channels. Only about two-thirds of the U.S. population lived in areas where the show was accessible. Thus, Kearney and Levine (2015) used households' distance to the closest television tower that transmitted UHF as an instrument for participation in the program. The researchers argue that since television towers were built in locations chosen by the government—all before *Sesame Street* was ever broadcast—the variable would not be related to household characteristics or changes in the outcome.

The evaluation found positive results on school readiness for preschool-aged children. In areas where there was UHF television reception when the show began, children were more likely to advance through primary school at the appropriate age. This effect was notable for African-American and non-Hispanic children, boys, and children in economically disadvantaged areas.

Source: Kearney and Levine 2015.

In the absence of full compliance in the treatment group, the estimated impact Δ is called the *intention-to-treat* (ITT) when comparing groups to which the program has randomly been *offered* (in the treatment group) or not (in the comparison group)—regardless of whether or not those in the treatment group actually enroll in the program. The ITT is a weighted average of the outcomes of participants and nonparticipants in the treatment group compared with the average outcome of the comparison group. The ITT is important for those cases in which we are trying to determine the average impact of offering a program, and enrollment in the treatment group is voluntary. By contrast, we might also be interested in knowing the impact of a program for the group of individuals who are offered the program and actually participate. This estimated impact is called the *treatment-on-the-treated* (TOT). The ITT and TOT will be the same when there is full compliance. We will return to the difference between the ITT and TOT in future sections, but start with an example to illustrate these concepts.

Key Concept

Intention-to-treat (ITT) estimates the difference in outcomes between the units assigned to the treatment group and the units assigned to the comparison group, irrespective of whether the units assigned to the treatment group actually receive the treatment.

Key Concept

Treatment-on-the-treated (TOT) estimates the difference in outcomes between the units that actually receive the treatment and the comparison group.

Consider the Health Insurance Subsidy Program (HISP), discussed in previous chapters. Because of operational considerations and to minimize spillovers, the unit of treatment assignment chosen by the government is

the village. Households in a treatment village (the villages where the health insurance program is being offered) can sign up for a health insurance subsidy voluntarily, while households in comparison communities cannot. Even though all households in treatment villages are eligible to enroll in the health insurance 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 any other myriad 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 comparing the average outcome for all households that were offered the program—that is, for 100 percent of the households in treatment villages—with the average outcome in the comparison villages (where no households have enrolled). By contrast, the TOT can be thought of as the estimated impact for the 90 percent of households in treatment villages that enrolled in the program. It is important to note that since individuals who participate in a program when offered may differ from individuals who are offered the program but opt out, the TOT impact is not necessarily the same as the impact we would obtain for the 10 percent of households in the treatment villages that did not enroll, should they become enrolled. As such, local treatment effects cannot be extrapolated directly from one group to another.

Imperfect Compliance

As discussed, in real-world social programs, full compliance with a program's selection criteria (and hence adherence to treatment or comparison status) is desirable, and policy makers and evaluation teams alike usually 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 evaluation team. We will now work through the different cases that can occur and discuss implications for the evaluation methods that can be used. We stress up front that the best solution to imperfect compliance is to avoid it in the first place. In this sense, program managers and policy makers should strive to keep compliance as high as possible in the treatment group and as low as possible in the comparison group.

Say you are trying to evaluate a teacher-training program, in which 2,000 teachers are eligible to participate in a pilot training. The teachers have been randomly assigned to one of two groups: 1,000 teachers are assigned to the treatment group and 1,000 teachers are assigned to the comparison group.

When all teachers in the treatment group receive training, and none in the comparison group have, we estimate the ATE by taking the difference in mean outcomes (say student test scores) between the two groups. This ATE is the average impact of the treatment on the 1,000 teachers, given that all teachers assigned to the treatment group actually attend the course, while none of the teachers assigned to the comparison group attend.

The first case of imperfect compliance occurs when some units assigned to the treatment group choose not to enroll or are otherwise left untreated. In the teacher-training example, some teachers assigned to the treatment group do not actually show up on the first day of the course. In this case, we cannot calculate the average treatment for the population of teachers because some teachers never enroll; therefore we can never calculate what their outcomes would have been with treatment. But we can estimate the average impact of the program on those teachers who actually take up or accept the treatment. We want to estimate the impact of the program on those teachers to whom treatment was assigned *and* who actually enrolled. This is the *TOT estimate*. In the teacher-training example, the TOT estimate provides the impact for teachers assigned to the treatment group who actually show up and receive the training.

The second case of imperfect compliance is when individuals assigned to the comparison group manage to participate in the program. Here the impacts cannot be directly estimated for the entire treatment group because some of their counterparts in the comparison group cannot be observed without treatment. The treated units in the comparison group were supposed to generate an estimate of the counterfactual for some units in the treatment group, but they receive the treatment; therefore there is no way of knowing what the program's impact would have been for this subset of individuals. In the teacher-training example, say that the most motivated teachers in the comparison group manage to attend the course somehow. In this case, the most motivated teachers in the treatment group would have no counterparts in the comparison group, and so it would not be possible to estimate the impact of the training for that segment of motivated teachers.

When there is noncompliance on either side, you should consider carefully what type of treatment effect you estimate and how to interpret them. A first option is to compute a straight comparison of the group originally assigned to treatment with the group originally assigned to comparison; this will yield the *ITT estimate*. The ITT compares 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). If the noncompliance is only on the treatment side, this can be an interesting and relevant measure of impact because in any case most policy makers and program managers can only offer a program and cannot force the program on their target population.

In the teacher-training example, the government may want to know the average impact of the program for all assigned teachers, even if some of the teachers do not attend the course. This is because even if the government expands the program, there are likely to be teachers who will never attend. However, if there is noncompliance on the comparison side, the intention-to-treat estimate is not as insightful. In the case of the teacher training, since the comparison group of teachers includes teachers who are trained, the average outcome in the comparison group has been affected by treatment. Let's assume that the effect of teacher training on outcomes is positive. If the noncompliers in the comparison group are the most motivated teachers and they benefit the most from training, the average outcome for the comparison group will be biased upward (because the motivated teachers in the comparison group who got trained will increase the average outcome) and the ITT estimate will be biased downward (since it is the difference between the average outcomes in the treatment and comparison groups).

Under these circumstances of noncompliance, a second option is to estimate what is known as the *local average treatment effect* (LATE). LATE needs to be interpreted carefully, as it represents program effects for only a specific subgroup of the population. In particular, when there is noncompliance in both the treatment group and in the comparison group, the LATE is the impact on the subgroup of compliers. In the teacher-training example, if there is noncompliance in both the treatment and comparison group, then the LATE estimate is valid only for teachers in the treatment group who enrolled in the program and who would have not enrolled had they been assigned to the comparison group.

In the remainder of this section, we will explain how to estimate the LATE, and equally importantly, how to interpret the results. The LATE estimation principles apply when there is noncompliance in the treatment group, comparison group, or both simultaneously. The TOT is simply a LATE in the more specific case when there is noncompliance only in the treatment group. Therefore, the rest of this chapter focuses on how to estimate LATE.

Randomized Assignment 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. The treatment group is assigned to the program, while the comparison group is not. Most likely, you will find three types of individuals in the population:

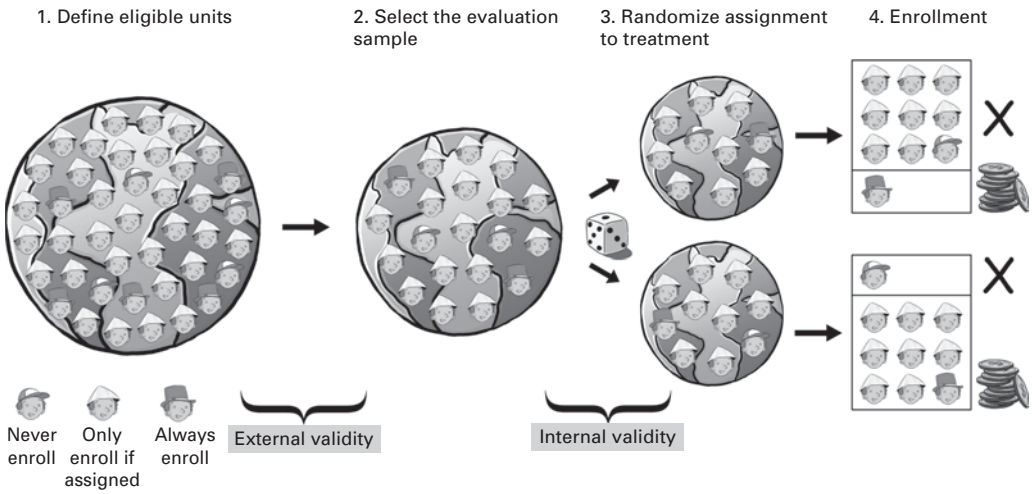
- *Enroll-if-assigned*. These are the individuals who comply with their assignment. If they are assigned to the treatment group (assigned to the program), they take it up, or enroll. If they are assigned to the comparison group (not assigned to the program), they do not enroll.

- *Never*. These are the individuals who never enroll in or take up the program, even if they are assigned to the treatment group. If assigned to the treatment group, these individuals will be *noncompliers*.
- *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. If assigned to the comparison group, these individuals will be *noncompliers*.

In the context of the job-training program, the *Never* group might consist of unmotivated people who, even if assigned a place in the course, do not show up. Individuals in 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-assigned* group comprises those who enroll in the course if they are assigned to it, but who do not seek to enroll if they are assigned to the comparison group.

Figure 5.1 presents the randomized assignment of the program and the final enrollment, or take-up, when *Enroll-if-assigned, Never,* and *Always* types are present. Say that the population comprises 80 percent *Enroll-if-assigned,* 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-assigned,* 10 percent *Never,* and 10 percent *Always.* Then if we randomly assign the

Figure 5.1 Randomized Assignment with Imperfect Compliance



evaluation sample to a treatment group and a comparison group, we should again have approximately 80 percent *Enroll-if-assigned,* 10 percent *Never,* and 10 percent *Always* in both groups. In the group that is assigned treatment, the *Enroll-if-assigned* and *Always* individuals will enroll, and only the *Never* group will stay away. In the comparison group, the *Always* will enroll, while the *Enroll-if-assigned* and *Never* groups will stay out. It is important to remember that while we know that these three types of individuals exist in the population, we can not necessarily distinguish an individual's type until we observe certain behaviors. In the treatment group, we will be able to identify the *Never* types when they fail to enroll, but we will not be able to distinguish the *Enroll-if-assigned* from the *Always,* since both types will enroll. In the comparison group, we will be able to identify the *Always* when they enroll, but we won't be able to distinguish between the *Enroll-if-assigned* and the *Never,* since both these types remain unenrolled.

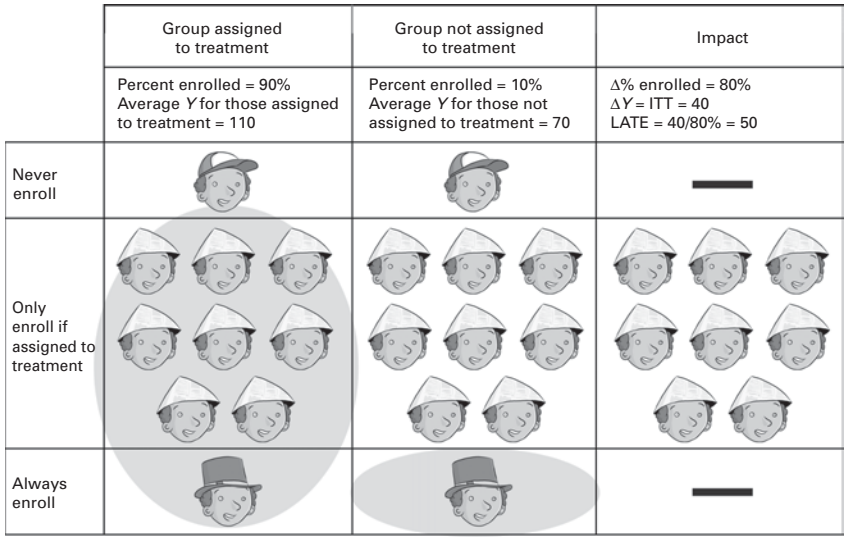
Estimating Impact under Randomized Assignment with Imperfect Compliance

Having established the difference between assigning a program and actual enrollment or take-up, we turn to estimating the LATE of the program. This estimation is done in two steps, which are illustrated in figure 5.2.[2]

To estimate program impacts under randomized assignment with imperfect compliance, we first estimate the ITT impact. Remember that this is just the straight difference in the outcome indicator (Y) for the group that we assigned to treatment and the same indicator for the group that we did not assign to treatment. For example, if the average wage (Y) for the treatment group is US$110, and the average wage for the comparison group is US$70, then the intention-to-treat estimate of the impact would be US$40 (US$110 minus US$70).

Second, we need to recover the LATE estimate for the *Enroll-if-assigned* group from the ITT estimate. To do that, we will need to identify where the US$40 difference came from. Let us proceed by elimination. First, we know that the difference cannot be caused by any differences between the people who never enroll (the *Nevers*) in the treatment and comparison groups. That's because the *Nevers* never enroll in the program, so for them, it makes no difference whether they are in the treatment group or in the comparison group. Second, we know that the US$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

Figure 5.2 Estimating the Local Average Treatment Effect under Randomized Assignment with Imperfect Compliance



Note: Δ = causal impact; Y = outcome. The intention-to-treat (ITT) estimate is obtained by comparing outcomes for those assigned to the treatment group with those assigned to the comparison group, irrespective of actual enrollment. The local average treatment effect (LATE) estimate provides the impact of the program on those who enroll only if assigned to the program (*Enroll-if-assigned*). The LATE estimate does not provide the impact of the program on those who never enroll (the *Nevers*) or on those who always enroll (the *Always*).

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-assigned* group. So if we can identify the *Enroll-if-assigned* 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 separate out unique individuals by whether they are *Enroll-if-assigned, Never,* or *Always*. In the group that was assigned treatment, we can identify the *Nevers* (because they have not enrolled), but we cannot differentiate between the *Always* and the *Enroll-if-assigned* (because both are enrolled). In the group that was not assigned treatment, we can identify the *Always* group (because they enroll in the program), but we cannot differentiate between the *Nevers* and the *Enroll-if-assigned*.

However, once we observe that 90 percent of the units in the group that was assigned treatment do 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 assigned treatment who did not enroll). In addition, if we observe that 10 percent of units in the group not assigned treatment enroll, we know that 10 percent are *Always* (again, the fraction of individuals in our group that was not assigned treatment who did enroll). This leaves 80 percent of the units in the *Enroll-if-assigned* group. We know that the entire impact of US$40 came from a difference in enrollment for the 80 percent of the units in our sample who are *Enroll-if-assigned*. Now if 80 percent of the units are responsible for an average impact of US$40 for the entire group assigned treatment, then the impact on those 80 percent of *Enroll-if-assigned* must be 40/0.8, or US$50. Put another way, the impact of the program for the *Enroll-if-assigned* is US$50, but when this impact is spread across the entire group assigned 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 conduct an evaluation where units are randomly assigned to the program, but actual participation is voluntary or a way exists 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-assigned*. However, provided that the noncompliance is not too large, 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, the estimate should be interpreted as a *local* estimate that applies only to a specific subgroup within our target population, the *Enroll-if-assigned*.

Randomized assignment of a program has two important characteristics that allow us to estimate impact when there is imperfect compliance (see box 5.2):

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

2. Since the two groups (assigned and not assigned treatment) are generated through a randomized 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).

Box 5.2: Using Instrumental Variables to Deal with Noncompliance in a School Voucher Program in Colombia

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

Angrist and others (2002) 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 assignment. Only about 90 percent of the lottery winners actually used the voucher or another form of scholarship, and 24 percent of the lottery losers actually received scholarships. Using our earlier terminology, the population must have contained 10 percent *Never*, 24 percent *Always*, and 66 percent *Enroll-if-assigned*. Angrist and others (2002) therefore also used the original assignment, or a student's lottery win or loss status, as an instrumental variable for the treatment-on-the-treated, or actual receipt of a scholarship. 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 and others 2002.

In statistical terms, the randomized assignment serves as an IV. It is a variable that predicts actual enrollment of units in a program, but is not correlated with other characteristics of the units that may be related to outcomes. While some part of the decision of individuals to enroll in a program cannot be controlled by the program administrators, another part of the decision is under their control. In particular, the part of the decision that can be controlled is the assignment to the treatment and comparison groups. Insofar as assignment to the treatment and comparison groups predicts final enrollment in the program, the randomized assignment can be used as an instrument to predict final enrollment. Having this IV allows us to recover the estimates of the local average treatment effect from the estimates of the intention-to-treat effect for the *Enroll-if-assigned* type of units.

A valid IV must satisfy two basic conditions:

1. The IV should not be correlated with the characteristics of the treatment and comparison groups. This is achieved by randomly assigning treatment among the units in the evaluation sample. This is known as *exogeneity*. It is important that the IV not directly affect the outcome of interest. Impacts must be caused only through the program we are interested in evaluating.
2. The IV must affect participation rates in the treatment and comparison groups differently. We typically think of increasing participation in the treatment group. This can be verified by checking that participation is higher in the treatment group compared with the comparison group. This condition is known as *relevance*.

Interpreting the Estimate of the Local Average Treatment Effect

The difference between an estimate of an ATE and an estimate of a LATE is especially important when it comes to interpreting the results of an evaluation. Let's think systematically about how to interpret a LATE estimate. First, we must recognize that individuals who comply in a program (the *Enroll-if-assigned* type) are different from individuals who do not comply (the *Never* and *Always* types). In particular, in the treatment group, noncompliers/nonparticipants (*Never*) may be those who expect to gain little from the intervention. In the comparison group, the noncompliers/participants (*Always*) are likely the group of individuals who expect to benefit the most from participation. In our teacher-training example, teachers who are assigned to the training but decide not to participate (the *Never* type) may be those who feel they don't need training, teachers with a higher opportunity cost of time (for example, because they hold a second job or have children to care for), or teachers with lax supervision who can get away with not attending. On the other hand, teachers who are assigned to the comparison group but enroll anyways (the *Always* type) may be those who feel they absolutely need training, teachers who don't have children of their own to care for, or teachers with a strict principal who insists everyone needs to be trained.

Second, we know that the LATE estimate provides the impact for a particular subgroup of the population: it takes into account only those subgroups that are not affected by either type of noncompliance. In other words, it takes into account only the *Enroll-if-assigned* type. Since the *Enroll-if-assigned* type is different from *Never* and *Always* types, the impact we find through the LATE estimate does not apply to the *Never* or *Always* types. For example, if the ministry of education were to implement a second round of training and somehow force the *Never* teachers who did not get

trained in the first round to get trained, we don't know if those teachers would have lower, equal, or higher effects compared with the teachers who participated in the first round. Similarly, if the most self-motivated teachers always find a way to take the teacher-training program despite being randomly assigned to the comparison group, then the local average treatment effect for the compliers in both treatment and comparison groups does not give us information about the impact of the program for the highly motivated teachers (the *Always*). The estimate of the local average treatment effect applies only to a specific subset of the population: those types that are not affected by noncompliance—that is, only the complier type—and should not be extrapolated to other subsets of the population.

Randomized Promotion as an Instrumental Variable

In the previous section, we saw how to estimate impact based on randomized assignment of treatment, even if compliance with the originally assigned treatment and comparison groups is imperfect. 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 can otherwise not control who participates and who does not.

This approach, called *randomized promotion* (also known as *encouragement design*), provides an additional encouragement for a random set of units to enroll in the program. This randomized promotion serves as an IV. It serves as an external source of variation that affects the probability of receiving the treatment but is otherwise unrelated to the participants' characteristics.

Voluntary enrollment programs typically allow individuals who are interested in the program to decide on their own to enroll and participate. Again consider 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 different types of people: compliers, a *Never* group, and an *Always* group.

- *Always.* These are the individuals who will always enroll in the program.
- *Never.* These are the individuals who will never enroll.
- *Compliers or Enroll-if-promoted.* In this context, any individual who would like to enroll in the program is free to do so. Yet some individuals may be interested in enrolling but for a variety of reasons, may not have sufficient information or the right incentive to enroll. The compliers here

are those who *enroll-if-promoted*: they are a group of individuals who enroll in the program only if given an additional incentive, stimulus, or promotion that motivates them to enroll. Without this additional stimulus, the *Enroll-if-promoted* would simply remain out of the program.

Returning to the job-training example, if the agency that organizes the training is well funded and has sufficient capacity, it may have an "open-door" policy, treating every unemployed person who wants to participate. It is unlikely, however, that every unemployed person will actually step forward to participate or will even know that the program exists. 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 encourage a randomly selected group of unemployed persons to enroll into the job-training program. Carrying the list of randomly selected 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. The visit is a form of promotion, or encouragement, to participate in the program. 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 randomly assigned a visit by the outreach worker, and those who were randomly 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. We cannot just compare those unemployed people who enroll with those who do not enroll. That's because the unemployed who enroll are probably very different from those who do not enroll in both observed and unobserved ways: they may be more or less 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, there is some additional variation that we can exploit to find a valid comparison group. Consider for a moment whether we can compare the group of people who were randomly assigned to receive a visit from the outreach worker with the group that was not visited. Because the promoted and nonpromoted groups were determined at random, both groups contain identical compositions of 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. Finally, if the outreach worker is

effective at motivating enrollment, 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.

Since the outreach worker visited a group of individuals assigned at random, we can derive a LATE estimate, as discussed earlier. The only difference is that instead of randomly *assigning* the program, we are 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 in sufficient numbers, we have variation between the group *with* the promotion and the group *without* the promotion that allows us to identify the impact of the training on the *Enroll-if-promoted*. Instead of complying with the assignment of the treatment, the *Enroll-if-promoted* are now complying with the promotion.

For this strategy to work, we want the outreach or promotion to be effective in increasing enrollment substantially among the *Enroll-if-promoted* group. At the same time, we do not want the promotion activities themselves to influence the final outcomes of interest (such as earnings), since at the end of the day we are interested primarily in estimating the impact of the training program, and not the impact of the promotion strategy, on final outcomes. 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 activity 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 a program has open enrollment and it is feasible to organize a promotion campaign aimed at a random sample of the population of interest. Randomized promotion is another example of an IV that allows us to estimate impact in an unbiased way. But again, as with randomized assignment with imperfect compliance, impact evaluations relying on randomized promotion provide a LATE estimate: a local estimate of the effect on a specific subgroup of the population, the *Enroll-if-promoted* group. As before, this LATE estimate cannot be directly extrapolated to the whole population, since the *Always* and *Never* groups are likely quite different from the *Enroll-if-promoted* group.

Key Concept

Randomized promotion is an instrumental variable method that allows us to estimate impact in an unbiased way. It randomly assigns a promotion, or encouragement, to participate in the program. It is a useful strategy to evaluate programs that are open to everyone who is eligible.

You Said "Promotion"?

Randomized promotion seeks to increase the take-up of a voluntary program in a randomly selected subsample of the population. The promotion itself 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 available.

As discussed for IV more generally, a number of conditions must be met for the randomized promotion approach to produce valid estimate of program impact:

1. The promoted and nonpromoted groups must be similar. That is, the average characteristics of the two groups must be statistically equivalent. This is achieved by randomly assigning the outreach or promotion activities among the units in the evaluation sample.
2. The promotion itself should not directly affect the outcomes of interest. This is a critical requirement so that we can tell that changes in the outcomes of interest are caused by the program itself and not by the promotion.
3. The promotion campaign must substantially change enrollment rates in the promoted group relative to the nonpromoted group. We typically think of increasing enrollment with promotion. 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.

The Randomized Promotion Process

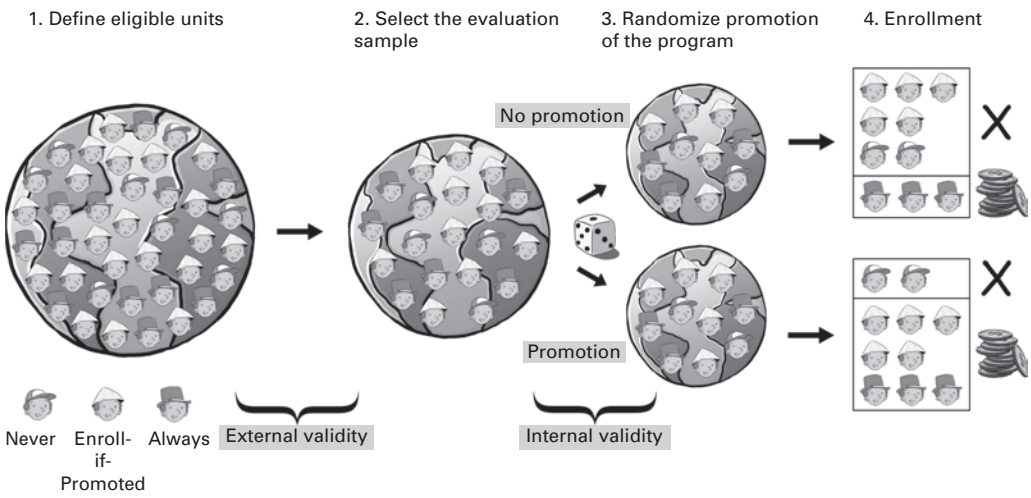
The process of randomized promotion is presented in figure 5.3. As in the previous methods, we begin with the population of eligible units for the program. In contrast with randomized assignment, we can no longer randomly choose who will receive the program and who will not receive the program because 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 easily measured by the program evaluation team because it is related to factors such as motivation, intelligence, and information.

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

Figure 5.3 Randomized Promotion



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.

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 both groups. In the nonpromoted group, only the *Always* will enroll. Although we 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

Imagine that for a group of 10 individuals per group, 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 the average outcome for all individuals the

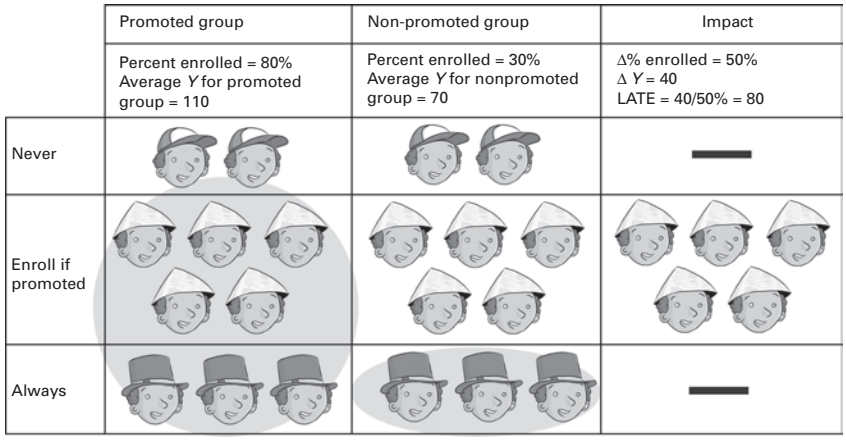
nonpromoted group (10 individuals) is 70, and that average outcome for all individuals in the promoted group (10 individuals) is 110 (figure 5.4). Then what would be the impact of the program?

First, let's compute the straight difference in outcomes between the promoted and the nonpromoted groups, which is 40 (110 minus 70). We 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 should come from the *Always* because they enroll in both groups. So all of this difference of 40 should come from the *Enroll-if-promoted*.

The second step is to obtain the LATE estimate of the program on the *Enroll-if-promoted*. We know that the entire difference between the promoted and nonpromoted groups 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 estimate of the local average treatment effect of the program on the *Enroll-if-promoted* group is 40/0.5=80.

Given that the promotion is assigned randomly, the promoted and nonpromoted groups have equal characteristics. Thus the differences that we observe in average outcomes between the two groups must be caused by

Figure 5.4 Estimating the Local Average Treatment Effect under Randomized Promotion



Note: Δ = causal impact; Y = outcome. Characters that appear against the shaded background are those who enroll.

the fact that in the promoted group, the *Enroll-if-promoted* enroll, while in the nonpromoted group, they do not. Again, we should not directly extrapolate the estimated impacts for the *Enroll-if-promoted* to other groups, since they are likely quite different from the groups that *Never* and *Always* enroll. Box 5.3 presents an example of randomized promotion for a project in Bolivia.

Box 5.3: Randomized Promotion of 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, built 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 and others (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 and others 2002.

Evaluating the Impact of HISP: Randomized Promotion

Let us now try using the randomized promotion method to evaluate the impact of the Health Insurance Subsidy Program (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. You note that this is a different scenario than the randomized assignment case we have considered so far. However, you know that realistically this national scale-up will be incremental over

time, so you reach an agreement to try and accelerate enrollment in a random subset of villages through a promotion campaign. In a random subsample of villages, you undertake an intensive promotion effort that includes communication and social marketing aimed at increasing awareness of HISP. The promotion activities are carefully designed to avoid content that may inadvertently encourage changes in other health-related behaviors, since this would invalidate the promotion as an instrumental variable (IV). Instead, the promotion concentrates exclusively on boosting enrollment in 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 5.1).

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 promotion. 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 US$14.97, compared with US$18.85 in nonpromoted areas (a difference of minus US$3.87). However, because the only difference between the promoted and nonpromoted villages is that enrollment in the program is higher in the promoted villages (thanks to the promotion), this difference of US$3.87 in health expenditures must be due to the additional 40.78 percent of households that enrolled in the promoted villages because of the promotion. Therefore, we need to adjust 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 intention-to-treat estimate—that is, the straight difference between the promoted and nonpromoted groups—by the percentage of *Enroll-if-promoted:* −3.87/0.4078 = −US$9.49.

Table 5.1 Evaluating HISP: Randomized Promotion Comparison of Means

	Promoted villages	Nonpromoted villages	Difference	*t*-stat
Household health expenditures at baseline (US$)	17.19	17.24	−0.05	−0.47
Household health expenditures at follow-up (US$)	14.97	18.85	−3.87	−16.43
Enrollment rate in HISP	49.20%	8.42%	40.78%	49.85

Note: Significance level: ** = 1 percent.

Table 5.2 Evaluating HISP: Randomized Promotion with Regression Analysis

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures (US$)	–9.50** (0.52)	–9.74** (0.46)

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

Your colleague, an econometrician who suggests using the randomized promotion as an IV, then estimates the impact of the program through a two-stage least-squares procedure (see online technical companion at http://www.worldbank.org/ieinpractice for further details on the econometric approach to estimating impacts with IV). She finds the results shown in table 5.2. 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, the *Enroll-if-promoted*.

HISP Question 4

A. What are the key conditions required to accept the results from the randomized promotion evaluation of HISP?
B. Based on these results, should HISP be scaled up nationally?

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 with randomized assignment of treatment.

First, the promotion strategy must be effective. If the promotion campaign does not increase enrollment, then no difference between the promoted and the nonpromoted groups will appear, and there will be nothing to compare. It is thus crucial to carefully design and extensively pilot the promotion campaign 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 after the evaluation period is concluded.

Second, the randomized promotion method estimates the impact of the program for only a subset of the population of eligible units (a LATE).

Specifically, the program's local average impact is estimated 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. 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. A randomized promotion evaluation will not estimate impacts for the group of individuals who enroll in the program without encouragement. In some contexts, this group (the *Always*) may be precisely the group the program is designed to benefit. In this context, the randomized promotion design will shed light on impacts expected for new populations that would enroll from additional promotion, but not on impacts for the population that already enrolls on its own.

Checklist: Randomized Promotion as an Instrumental Variable

Randomized promotion leads to valid estimates of the counterfactual if the promotion campaign substantially increases take-up of the program without directly affecting the outcomes of interest.

- ✓ Are the baseline characteristics balanced between the units that received the promotion campaign and those that did not? Compare the baseline characteristics of the two groups.
- ✓ Does the promotion campaign substantially affect the take-up of the program? It should. Compare the program take-up rates in the promoted and the nonpromoted subsamples.
- ✓ Does the promotion campaign directly affect outcomes? It should not. This cannot usually be directly tested, so you need to rely on theory, common sense, and good knowledge of the setting of the impact evaluation for guidance.

Additional Resources

- For accompanying material to the book and hyperlinks to additional resources, please see the Impact Evaluation in Practice website (http://www.worldbank.org/ieinpractice).
- For additional resources on IV, see the Inter-American Development Bank Evaluation Portal (http://www.iadb.org/evaluationhub).

Notes

1. 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 further control for the *placebo effect*, meaning the potential changes in behavior and outcomes that could occur simply from the act of receiving a treatment, even if the treatment itself is ineffective.
2. These two steps correspond to the econometric technique of two-stage least-squares, which produces an estimate of the local average treatment effect.

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.

Kearney, Melissa S., and Philip B. Levine. 2015. "Early Childhood Education by MOOC: Lessons from *Sesame Street*." NBER Working Paper 21229, National Bureau of Economic Research, Cambridge, MA.

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.