NOTE 6: Identifying an Appropriate Impact Evaluation Method

The objective of this note is to provide practitioners with an overview of the different tools available for an impact evaluation and to provide guidance on which tool may be the most appropriate for a particular program. We present a toolbox of six methods commonly used in impact evaluation, organized by their ability to construct a counterfactual with minimal bias. Each technique has advantages and disadvantages. The choice of an impact evaluation method will depend not only on the theoretical quality of the method, but also on the operational context of the program. Program managers therefore need to be involved during the evaluation design to make sure the evaluation responds to the needs and context of the intervention.

Choosing Among Impact Evaluation Methods

Every impact evaluation technique differs in terms of the circumstances in which it is best applied; every evaluation does not fit every program context. The characteristics and circumstances of our program will thus guide our selection of the impact evaluation method to be used. In particular, as Gertler and colleagues (2011, pp. 143–149) illustrate, we need to consider timing, coverage, targeting, and resources.

Timing

Has the program already started? The key issue here is whether the impact evaluation can be incorporated into the program design. As will be explained in more detail below, when an impact evaluation is planned from the outset of the program, the quality of the evaluation will be greatly increased and a much larger scope of methodologies can be used.

Coverage

Can the program serve all eligible people? Ideally, we would like to serve every young person in need. This is easier for some types of programs than for others. If the program offering is not resource intensive (such as opening savings accounts for minors) or if it is provided via mass media channels (financial literacy campaign via radio or TV) then we may not want to—or even be able to—exclude anyone from benefiting from the intervention. In most cases, however, we do not have enough resources to provide our youth livelihood programs to everyone who is eligible, forcing us to decide which of the eligible youth will receive the program and which will not. Although not being able to reach every youth may be frustrating from a programming perspective, excess demand offers opportunities to identify a comparison group and conduct quality assessments on the impact of our program.

Targeting

How does our program select beneficiaries? Unless we are able to provide the program to all eligible youth, the selection of individuals or groups occurs by the following means:

1. **Random assignment** is the process of giving each individual or group an equal chance to receive benefits. Drawing names out of a hat to decide who will receive job training now and who will be waitlisted is one example.
2. **Eligibility ranking** determines eligibility according to clear criteria using a cutoff point or threshold. Providing scholarships based on test scores, or providing training based on income levels are examples of eligibility ranking.
3. **Selective targeting decision.** Sometimes there are no clear criteria for why one individual or group is selected over another, which, rather than ensuring fairness in selection, leads to a biased selection of participants. Cases such as first come, first served practices; political factors; and reasons of practicality are examples of inherently subjective selection methods.

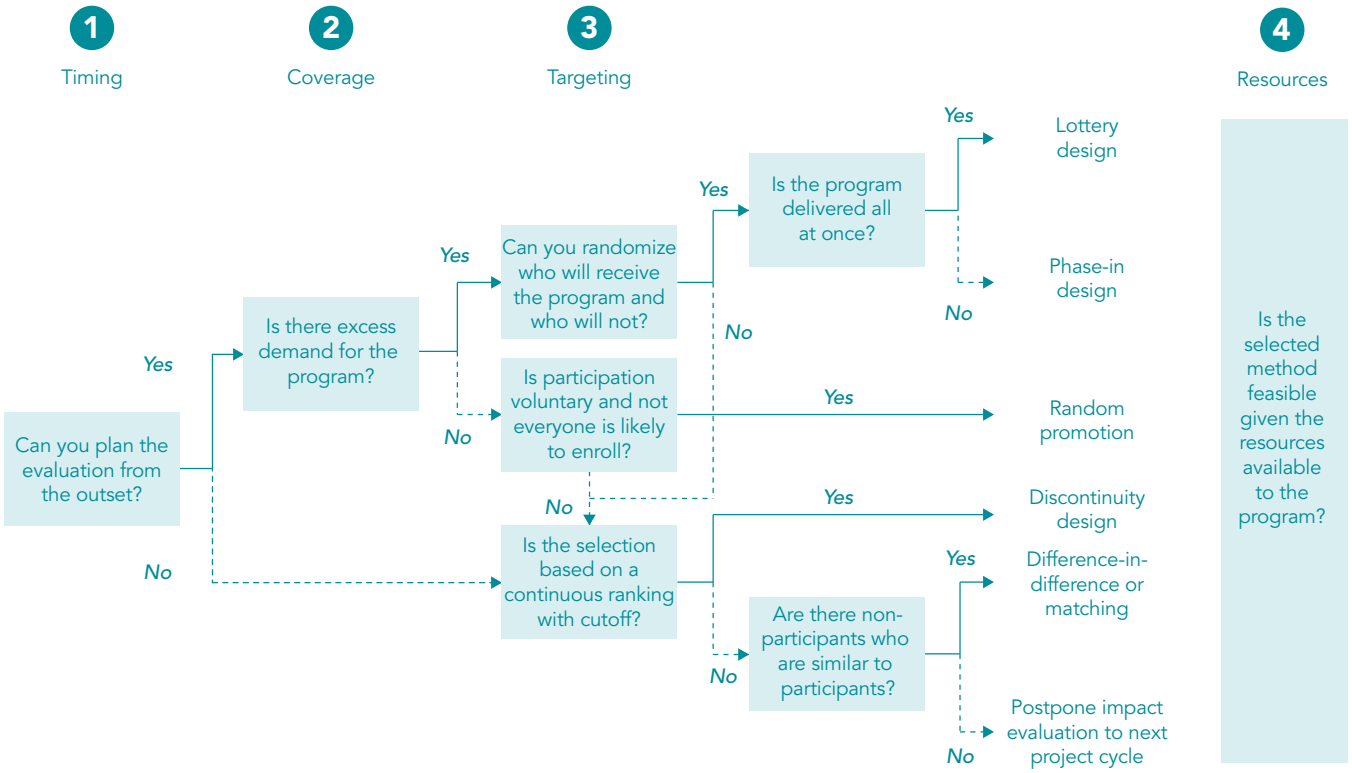
Resources

Does the program have the resources to carry out a specific impact evaluation? Impact evaluation techniques have different requirements in terms of sample size, data collection, complexity of statistical analysis, and cost. Even when we identify a method

that would fit our operational context, it may or may not be feasible given the resources available to us.

The four questions above should be in the back of our minds as we consider various impact evaluation techniques. The answer to these questions will determine which of the six methods is best in our context (see figure 6.1). A discussion follows of the evaluation methods themselves.

FIGURE 6.1 Decision tree for choosing impact evaluation techniques



Sources: Elaborated upon GAO (1991, p. 69); Duflo, Glennerster, and Kremer (2006, pp. 24–27); Gertler et al. (2011, p. 148).

[Definition]

A **randomized controlled trial** is a study in which people are allocated at random (by chance alone) to receive a treatment, such as participating in a specific intervention.

A **sample** is a subset of a population. Since it is usually impossible or impractical to collect information on the entire population of interest, we can instead collect information on a subset of manageable size. If the subset is well chosen, then it is possible to make inferences or extrapolations to the entire population.

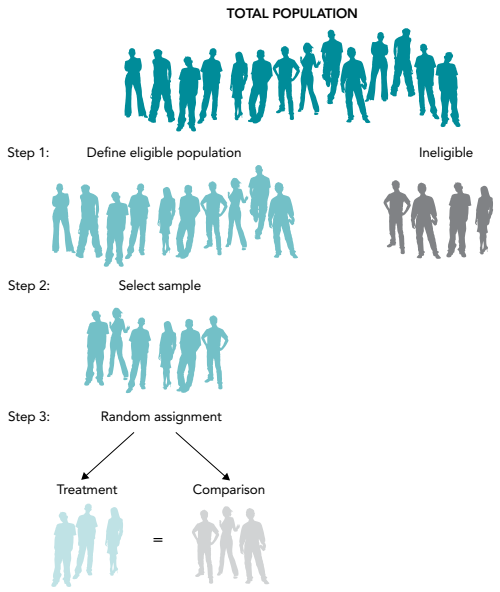
Method 1: Lottery Design

A lottery is a simple and transparent way to assign youth to groups who will receive our services (the treatment group) and those who won't (the comparison group). This is the method used to design randomized controlled trials. It is a statistical regularity that if a large enough sample of people from the same population of interest are randomly assigned to one of two groups, then both groups will, on average, have similar observable characteristics (age, gender, height, level of education, and the like) and unobservable characteristics (such as motivation and state of mind). Through randomization, the difference in outcomes we observe between the two groups at the end of our program can be attributed to the intervention because all other factors that could influence the outcomes are, on average, equal. Lottery designs are considered the most robust type of impact evaluation, so the results are usually the most trusted by donors, stakeholders, and governments.

How It Works

There are three steps to a lottery design (see figure 6.2).

FIGURE 6.2 Steps in a lottery design



Step 1: Define the Eligible Population

The first step in a randomized controlled trial is to find a group of eligible young people for a program. If a medical scientist is studying the effect of a drug on a childhood disease, she searches for a specific group of children and will not enroll adults or elderly people in the program. Likewise, a youth livelihood program may target urban street youth of a specific age range, and so will not include adults or rural youth. What is important here is to have very clear and transparent criteria (age, gender, income level, employment status, etc.) and to be able to communicate who will be eligible to join the program and who won't.

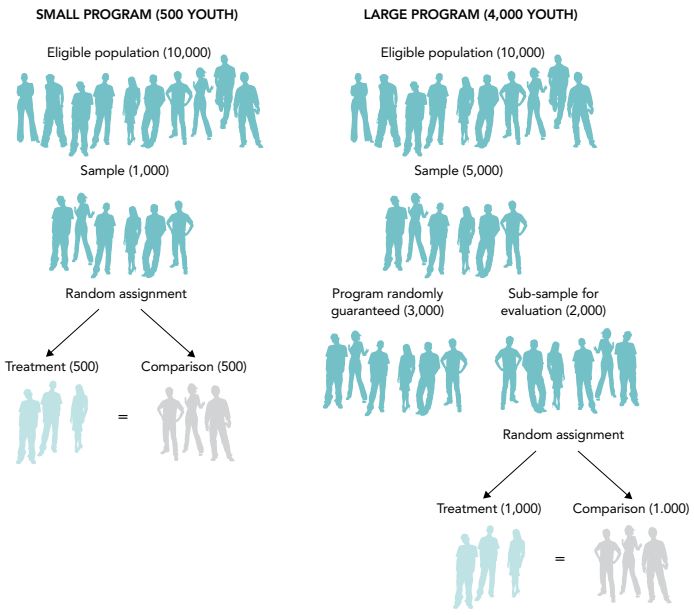
Step 2: Select a Sample for the Evaluation

To evaluate an intervention, we do not need to test everyone who will participate in the intervention. We just need to choose a representative group of people that is numerous enough for the purpose of our evaluation; this is called our *sample* (see note 7 for more details about how to determine the sample and its size). These will be the youth on whom we will collect data.

Choosing the sample for the evaluation can be done in two ways, depending on whether the program is large or small. A small program may find that there are 10,000 eligible beneficiaries, such as urban street youth aged 16–24. The program may have the budget to help 500 of them. Ideally, a comparison group will be equal in size to the treatment group, so 1,000 out of the 10,000 street youth will need to be selected for the program and evaluation (see figure 6.3, left image).

Large programs may be bigger than the sample size needed for an evaluation. If the program is able to serve 4,000 youth, it is not necessary to find an additional 4,000 youth for comparison. Instead, only 1,000 may be needed. The program can then identify a sample of 5,000 youth from the total population of 10,000. Of these, 3,000 youth can be guaranteed admission to the program. The remaining 2,000 will then be randomly split between the program and the comparison group (figure 6.3, right image).

FIGURE 6.3 Choosing samples for small and large programs



In order to make the selection representative of the total eligible population of 10,000 street youth, the sample (whether 1,000 in the first case or 5,000 in the second case) should be selected at random from the eligible population. By selecting randomly, the program participants will, on average, have similar characteristics as the total eligible population. Even though we include only a limited number of youth in the study, the potential impact of the program can be generalized to the entire eligible population, in this case, 10,000 youth.

[Tip]

One way of getting a random sample of youth is to get a list of the total population of street youth from a census, voter registration records, or some other database, and randomly select from that list. If that is not possible, randomly targeting areas where street youth interact, such as an urban center, will produce a random sample. If youth are known to spend time at 50 different centers around a city or country, randomly selecting centers and then selecting a portion of youth at these centers to participate in the study will likely result in a selection of youth with minimal bias. Note 7 will discuss sampling more in detail.

Step 3: Randomize Assignment

The next step is to assign the selected sample of youth to treatment and comparison groups roughly equal in size. In randomized controlled trials, every youth has the same chance of receiving the program. Randomization can be via traditional techniques such as flipping a coin, rolling dice, or drawing names out of a hat. Randomization can be done publicly, if desired, if the sample is relatively small (drawing 2,000 names out of a hat, for example, would not be very practical). Alternatively—and more appropriately if the number of people is large—we can randomize by using computer software, such as MS Excel. Randomization can occur at several levels (see box 6.1). By assigning our sample to treatment or comparison groups randomly, we select participants fairly, and we also develop a good counterfactual: if the sample size is big enough, youth in the treatment group have, on average, the same observable and unobservable characteristics as those in the comparison group.

BOX 6.1 Levels of randomization

Randomization can be conducted at the individual, group, or community level, according to program needs.

Individual level. Individual randomization is best for programs in which outcomes will be measured for each participant. There may be problems with this method, such as spillover, which occurs when individuals in the comparison group receive some of the treatment through informal means. For example, youth who received training or other information through our program may share their knowledge or resources with their friends in the comparison group.

Group level. Individual randomization is not always feasible or desirable. If there is not a list of people's names readily available, or if there is an expectation that people selected for the comparison group may receive the program anyway, then randomizing at a group level may be better. This works particularly well for programs that operate on a group level, targeting schools, vocational training centers, youth centers, and the like. In this case, groups of people are randomized into treatment or comparison cohorts. All individuals in the treatment group would receive the same intervention. Randomization at the group level can help reduce spillover effects and may be easier than randomizing on the individual level. Alternatively, it may also be possible to randomize at the subgroup level, such as classrooms in schools.

Village/community level. Programs may also choose to randomize at the level of villages, neighborhoods, communities, or even districts, when activities are implemented on that level, or when spillover effects are expected to occur beyond the group level. For example, if there are 100 villages in a district of interest and we don't have the resources to work with all of them, we may randomly choose to work with fifty of them, while keeping the other fifty villages as a comparison. All the youth within the respective treatment villages would then be eligible to participate in the program.

(continued)

BOX 6.1 (CONT'D) Levels of randomization

Implementing an intervention at a higher level, and, in turn, randomizing at that level, though it may reduce unwanted spillover effects, can also be problematic for the following reasons:

- The higher the level of randomization, the smaller the number of observations that can be compared with one other. Interviewing a number of people per area can mitigate this problem.
- The size of the evaluation sample increases with the scale of the intervention, which can have implications for the cost of the evaluation.
- Higher level units are more likely to experience different external influences over time, which has implications for the comparability between treatment and comparison group, and thus for the internal validity of the evaluation.

Program managers should therefore find the minimum scale of intervention at which the program can be implemented and randomized.

When Can I Use a Lottery Design?

A randomized lottery evaluation is used when the evaluation is planned in advance of implementation (prospective) and when the program can serve only a fraction of eligible youth. As long as resource constraints prevent the program from serving the entire eligible population, there are no ethical concerns in having a comparison group because a subset of the population will necessarily be left out of the program. In such a situation, comparison groups can be maintained to measure short-, medium-, and long-term impacts of the program (Gertler et al. 2011).

With any prospective evaluation, new data will need to be collected, suggesting cost implications. At a minimum, an endline survey (to be discussed in length in note 7) will be required for youth in both the treatment and comparison groups. In many cases, a baseline survey will be needed, as well. Despite the costs associated with collecting new data, a simple random lottery can be the cheapest option for an evaluation because it may require fewer surveys and lower numbers of respondents.

[Definition]

A **prospective evaluation** is one in which participants will be followed in the future, so these studies must be planned as the program is being designed.

Evaluations that look back on participants in programs that have already been implemented or even ended are called **retrospective evaluations**.

Advantages

- A lottery design is the most robust method for developing a counterfactual because it leads to a very well matched comparison group (relying on fewer assumptions than other methods). It is therefore considered the most credible design to measure impact.
- It is by far the analytically simplest of all evaluation methods. The impact of the program in a random trial is simply the mean difference in outcomes between treatment and comparison groups.
- It allows for communities to be directly involved in the selection process for a fair and transparent allocation of benefits.
- Since it is planned from the outset of the program, it can be designed to measure the average program impact and also to compare the effectiveness of different components, different lengths of programming, and so on.
- It is easy to implement and communicate to program staff.

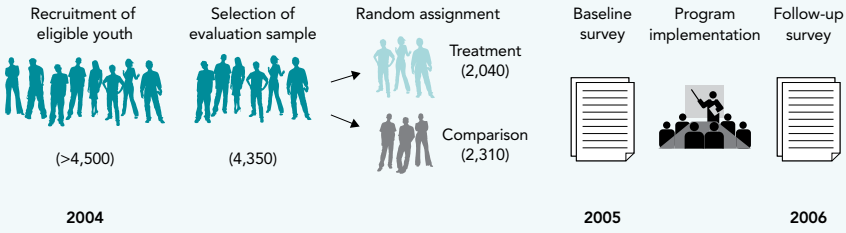
Disadvantages

- It requires a comparison group to be excluded from the program for the duration of the impact evaluation.
- Organizations must ensure that partners and local stakeholders consent to the method.
- The internal validity of a lottery design depends on the fact that the randomization works and is maintained throughout the study, which may not be easy to do. This condition may be threatened if randomization is done incorrectly, if treatment or comparison groups do not comply with their status (that is, if treatment individuals do not take up the program or comparison individuals receive the program), if participants drop out of the study prior to completion, or if there are spillover effects.

Box 6.2 provides an example of a lottery design.

BOX 6.2 Example of a lottery design

Attanasio, Kugler, and Meghir (2009) used a lottery design to study Jóvenes en Acción, a youth employment program in Colombia that provided three months of in-classroom training and three months of on-the-job training to young people aged 18–25 in the lowest socioeconomic strata of the population. The training providers were instructed to recruit more candidates than they had room for in their courses in case not everyone would eventually attend the training. Participants were then selected randomly from the pool of recruited candidates, and the remaining youth were waitlisted and used as the comparison group.



Attanasio and colleagues were concerned that despite randomization, the treatment and comparison groups might be different in ways that the researchers could not control. Using baseline data, they checked the comparability of the two groups and found that, on average, the treatment group had attended school three months longer than the comparison group and had about 5 percent more young women than the comparison group. Neither of these characteristics was thought to significantly influence the treatment outcomes.

The overall results were promising. On average, those who had gone through the program were more likely to be in paid formal employment, have higher incomes, and retain their jobs longer than those in the comparison group. The effects were generally stronger for women than for men.

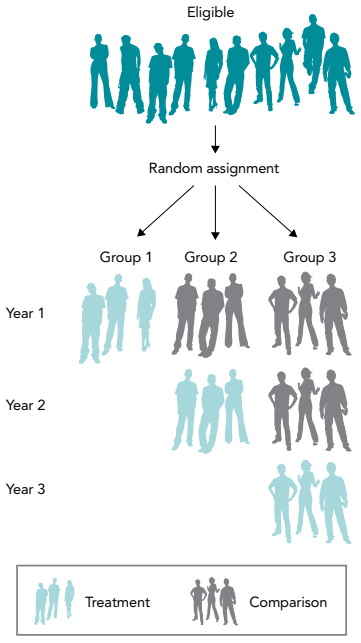
Method 2: Randomized Phase-In Design

Creating a pure comparison group in which youth are never given the program is sometimes impossible. Because many programs are in a community for years, never giving the program to a group of needy individuals can be both politically and programmatically difficult. A variation of the lottery design is the phase-in design. It applies to programs that are rolled out over time, and it uses the natural output flow to develop the treatment and comparison groups.

How It Works

The main difference between a phase-in design and a lottery design is the method of assigning people to treatment and comparison groups. When an intervention is delivered in several tranches over time, a phase-in design gives each eligible person or group the same chance of receiving the program under each of the tranches. One set of youth is then randomly selected to receive the treatment in the first period, while another group is selected to receive the program in the second period, a third group in the third period, and so on. For the time that certain groups are waitlisted, they can serve as the comparison group until they receive the program (see figure 6.4).

FIGURE 6.4 Treatment and comparison groups in phase-in design



Note: Treatment does not necessarily have to stop for the evaluation to work. Some interventions, once in place, will continue to be implemented. However, many programs, such as training, are offered over a limited period of time.

For example, an NGO may have the budget to train 1,500 youths, but it may not have the capacity to conduct all of the training at once. Instead, it chooses to train 500 people per year for three years. If it can identify all 1,500 participants in the beginning, a phased-in randomization may be the best evaluation method for them. The 1,500 youths are randomly split into three groups. In year one, while group 1 receives training, groups 2 and 3 are waitlisted and can serve as the comparison group. In year two, only group 3

remains for comparison. By year three, all three groups will have received training.

As individuals are selected at random for the different groups, it is possible to compare those offered treatment first with those offered treatment later. However, because everyone eventually gets the program, the phase-in design is usually not well suited to finding the long-term impact of a program because eventually there is no comparison group. Even large, longstanding programs will have difficulties asking participants to wait around for three or four years, so the time span of results is often limited to one or two years.

When Can I Use A Phase-in Design?

[Tip]

With a phase-in approach, it is critical to have enough time between each of the phases for the program to show effects. If a program officer believes it will take two years for the impact of the program to take effect, the time between the first and last phase must be at least two years. Small or short-run programs may not be suitable for this approach.

As with a lottery design, a phase-in evaluation is prospective and requires excess demand and the ability to assign participants randomly to treatment and comparison groups. The phase-in design is better suited than a lottery design to large programs that expect to rollout interventions over a number of years. Because the phase-in design requires a set plan for rollout, it also requires a dedicated program team that will be able to follow the rollout through the life of the program.

Phase-in designs do not differ significantly from the lottery design in data or cost requirements. An endline survey will need to be conducted, as well as a baseline survey, in many cases. One important difference is that the program implementation costs may increase because resources will be needed to ensure rollout is implemented in the manner required by the evaluation.

Advantages

- Phase-in designs produce a robust counterfactual, have a fair and transparent selection process, and allow for comparing the impacts of program alternatives.
- The method suits the natural rollout of many programs.
- Because everyone eventually receives the program with this method, phase-in studies can be politically expedient.

Disadvantages

- As with the lottery method, there are challenges to guaranteeing successful randomization and maintaining treatment and comparison groups over time.
- Participants may not wait to join in the program. If they do, there is a risk that they will change their behaviors in the meantime and therefore will not be a comparable comparison group. For example, they may stop looking for jobs in anticipation of joining the program.
- The phase-in method cannot estimate the long-term impact of the program.
- This method requires a clear rollout strategy, which may have operational implications.

See box 6.3 for an example of randomized phase-in design.

BOX 6.3 Example of randomized phase-in design

The World Bank's Economic Empowerment of Adolescent Girls program in Liberia provides six months of training and six months of follow-up activities with two different curricula: (1) skills training for wage employment, combined with job placement assistance; and (2) business development skills combined with links to microfinance. Mentorship is also provided to all beneficiaries starting from the third month of training.

To evaluate its impacts, the World Bank chose a phase-in evaluation design since this would allow for a quality randomized evaluation while also being able to eventually serve all girls who have been promised training. The evaluation took advantage of the natural rollout of the program and the operational constraints that did not allow for training everyone at the same time.

After the baseline survey, 1,273 participants were randomly assigned to the treatment group (receiving training during the Round I of the program in 2010) and 843 to the comparison group (receiving training during the Round II of the program in 2011). The follow-up survey was conducted at the end of each round and complemented with qualitative exit polls to collect information on the participants' views of their training, content, pedagogy, and trainers.

Because the program and evaluation targeted girls who specifically expressed interest in the training, results of the evaluation cannot be generalized to any young woman in the population. The evaluation helps us understand the impact of the training on those who chose to receive training and assistance for wage work or entrepreneurship.

Sources: World Bank (2008); Muzi (2011).

Method 3: Randomized Promotion Design

There may be cases where it is not possible or desirable to exclude any potential beneficiaries either because participation is voluntary and everyone can enroll if they desire or because the program has a sufficient budget to serve the entire eligible youth population immediately. In such cases, the randomized promotion method (also called encouragement design) may be suitable.

How It Works

Randomized promotion identifies the eligible population and chooses a sample just as in lottery or phase-in designs. But it differs in the randomization process. When it is not possible to randomly assign youth into a group that receives benefits and a group that does not, it may be possible to instead randomly promote the program. That is, rather than randomizing those who *receive* the benefits and services, we randomize who is *encouraged to receive* those benefits.

Random promotion is based on the premise that for many programs there will be three sets of potential beneficiaries:

- Youth who never enroll
- Youth who always enroll
- Youth who enroll only if they are encouraged to do so

No matter what the program offers, whether it is free savings accounts, vocational training, or media-based financial literacy programs, it is usually unlikely that every young person who is eligible will want to participate. Some may simply be distrustful of the intervention, others may face constraints such as time or transportation, and others

may just not know about the program.

Random encouragement may take many different forms. In the case of youth savings accounts, we may randomly advertise the initiative in selected schools. For a training program, we could hire a social worker to randomly visit homes of unemployed youth, describe the program, and offer to enroll youth on the spot. In the case of a financial literacy campaign, we may want to randomly send text messages to part of the target audience, but not to others. In all cases, there will still be people in the promoted group that will not take up our program, as there will be people in the non-promoted group who actually will. But the idea is that if the encouragement is effective, then the enrollment rate among the promoted group should be higher than the rate among those who did not receive the promotion. And if the promotion was done randomly, then the promoted and non-promoted groups share, on average, the same characteristics, allowing causal impact to be identified.

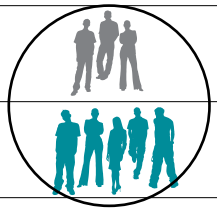
Random promotion evaluation may be suitable for

- *programs that distribute training vouchers.*
- *programs encouraging youth to open saving accounts.*
- *interventions leveraging mass-media based campaigns.*

Unfortunately, we cannot just compare the outcomes of those who participated in the program with the outcomes of those who did not. As discussed in note 5, people who choose to participate in a program are almost always different from those who do not, and many of these differences may not be observable or measurable. Even if promotion is random, participation in the program will not be random, so comparing participants to nonparticipants would be like comparing apples to oranges.

What we can do, though, is compare the outcomes of all those youth who received the promotion with the outcomes of those who did not receive the promotion (see figure 6.5). Let's consider an example of a job-training program in which 30 percent of eligible youth in the non-promoted group and 80 percent of eligible youth in the promoted group participated in the training (Gertler et al. 2011). One year after the program, we observe an average monthly income of $60 for the non-promoted group and $100 for the promoted group.

FIGURE 6.5 Estimating impact under randomized promotion

	Non-promoted Group	Promoted Group	Observed Change
Enrollment (% of eligible population)	30%	80%	50%
Type 1: Never enroll			
Type 2: Always enroll			
Type 3: Enroll only if promoted			
Average outcome (monthly income)	$60	$100	$40
Causal impact			**$80** (=$40/.5)

 *Those who actually enroll in each scenario*

Source: Adapted from Gertler et al. (2011, p. 75).

Given that the promotion is assigned randomly, the promoted and non-promoted groups have, on average, equal characteristics. Thus, the difference that we observe in average outcomes between the two groups ($40) can be attributed to the fact that in the group of people who enroll only if promoted take up the program. Though we cannot directly differentiate them from those who always enroll, we know that their share of the entire population is the difference in enrollment rates (50 percent, or 0.5). Thus, the average impact of the program on those who participated because of the encouragement is 40/0.5=$80.

When Can I Use Random Promotion?

Randomized promotion is well suited for prospective evaluations of programs that have universal eligibility or those in which we cannot control who participates and who does not. It works best when some sort of encouragement can significantly influence take-up. Random promotion is not a good option for services that are extremely popular, such as cash grants, which everyone will want to receive once they hear about it.

With this method, we calculate our average program impact based on people who joined the program as a result of promotional efforts. Because these participants are only a subset of the eligible population, we usually need very large samples for this type of evaluation in order to be sure our results are *statistically significant*. This increases the burden for data collection. If promotion is done on the community level, we may need to survey many more people in the community than we would have had to survey with a simple lottery design. As a result, our costs will likely be higher than costs associated with other types of evaluations. Other conditions are shown in box 6.4.

[Definition]

In statistics, a result is called **statistically significant** if it is unlikely to have occurred by chance. Statistical significance does not tell us anything about the magnitude of the effect size (economic significance); that is, the impact of a program could be statistically significant, yet very small.

BOX 6.4 Necessary conditions for promotion design to produce valid impact estimates

The promoted and non-promoted groups must have comparable characteristics. This can be achieved by randomly assigning outreach or promotion activities to individuals, groups, or communities in the evaluation sample.

The promotion campaign must increase enrollment by those in the promoted group substantially above the rate of the non-promoted group. "Substantially" is a relative concept based on statistical power needs. In general, a program should increase participation by 40 percent or more to be cost effective. 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.

It is important that the promotion itself does not directly affect the outcomes of interest. If the promotion itself changes behavior, it is not possible to determine whether the changes observed in people are due to the program or the promotion. This is most likely to happen if the promotion is done in conjunction with training programs. In most cases, it is most important to know the effect of the program, not of the promotion.

Source: Adapted from Gertler et al. (2011, p. 73).

Advantages

- Randomized promotion campaigns never deny anyone the program, but instead allow people to make their own decisions about whether or not to take up the program.

- This type of evaluation produces a high-quality comparison group with, on average, the same characteristics as the treatment group, just like any of the other randomization methods described above.

Disadvantages

- This method can be used only for specific programs.
- It often needs larger sample sizes than other methods, which increases costs.
- Advanced statistical techniques are required to calculate the program impact.
- Researchers must be careful when interpreting results because the impact estimate is valid only for those who participated in the program because they were encouraged; results cannot be generalized to other groups of potential beneficiaries.

An example of a randomized promotion design is in box 6.5.

BOX 6.5 Example of a randomized promotion design

In South Africa, a randomized promotion design was used to evaluate the impact of entertainment education that aims to enhance the knowledge, attitudes, and behavior regarding sound financial decision making, with a particular focus on managing debt. The program consists of including financial capability storylines in the South African soap opera *Scandal!*, which has been running for several years.

Evaluating the impact of a soap opera on behavior and attitudes is quite challenging. First, it is difficult to separate the effect of the soap opera's message from other messages on similar issues that individuals and families may receive from other sources. Second, certain types of individuals may self-select into watching a particular soap opera, and hence any subsequent behavior change is confounded by these selection attributes. Third, since access to TV is basically universal, it is difficult to establish a good comparison group of individuals who do not receive the financial capability messages.

To overcome these issues, the following randomized promotion methodology was designed: After the study population was identified (approximately 1,000 people), about half the population was provided a financial incentive (about $10) to watch *Scandal!* This was the randomly selected treatment group. Encouragement to watch the program took place through calls before a total of three shows over a period of three months alerting individuals of their financial incentive to watch that particular show. During those calls, treatment group members learned the conditions under which they could receive their incentive and they were asked a number of questions to establish prior knowledge about financial issues. After the show aired, individuals were called and awarded the incentive if they answered several questions about the nonfinancial content of the show correctly. During the same call, they were asked a number of questions on financial knowledge and attitudes.

The same financial incentive was provided for the other half of the population—the randomly selected comparison group—to watch a similar soap opera, one that was aired about the same time and, importantly, did not have a financial literacy component. The mechanism for awarding the incentive was identical to the treatment group. The comparison group was asked the same questions on financial literacy as the treatment group.

The theory was that, if the financial education component of *Scandal!* was successful, those who were encouraged to watch the show would score better on financial questions than the group who was encouraged to watch the other soap opera. Immediate effects on knowledge and attitudes were captured through the short survey after the end of the show; long-term effects were captured through multiple follow-up surveys.

Source: World Bank (2011).

Method 4: Discontinuity Design

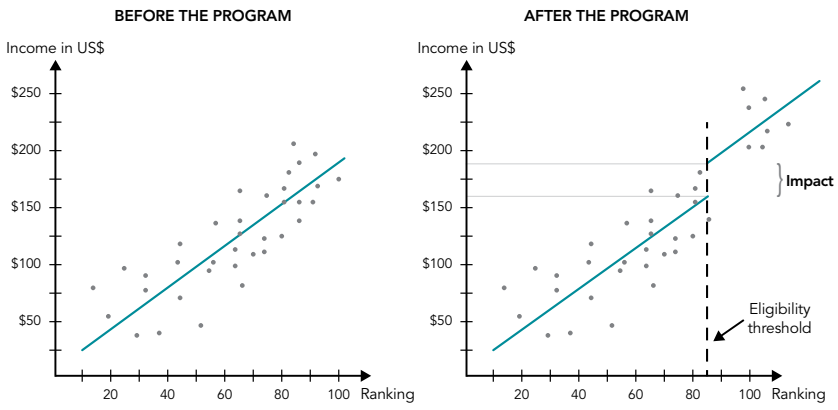
The reality is that in many cases we are not able to plan the evaluation during the program design, and even when we are, it may be impossible to use any form of randomization to obtain a valid counterfactual. In these cases, we may be able to use other targeting rules of the program to obtain a good comparison group. In fact, many programs use a continuous ranking of potential beneficiaries, such as test scores, credit scores, poverty index, or age, and have a cutoff point for acceptance into the program. For example, applicants to a business plan competition or a microfinance bank may be given a score based on a set of criteria and assigned a grade 1–100. If youth score at or above the minimum threshold, say 85 and above, they receive start-up financing. If they score below, they are not accepted into the program. Eligibility rankings like these can be used for an impact evaluation.

How It Works

The premise of discontinuity (or eligibility-index) evaluation designs is that the people who score just above and just below a defined threshold are not very different from one another, or at least the difference may be continuous across the scores. For instance, are applicants who receive a score of 86 much different from those who receive an 84? Probably not. Or are 18-year-olds, who may be eligible for cash-for-work programs, very different from their 17-year-old peers, who may not be eligible? If we have a situation in which some of those youth who receive the program (those just above the threshold) and some of those who don't (those just below the threshold) are not fundamentally different from one another, then comparing the outcomes of these two groups, in turn, would allow us to analyze program impact.

Figure 6.6 illustrates what we may find when analyzing the impact of a youth microcredit initiative. The left graph indicates that, at the time of applying to the program, those who achieved better scores already tended to have higher incomes. There may be many reasons for this, such as that those with somewhat better education are already earning more and that their education also helped them secure better scores. Or those who are more motivated in starting a business were already more entrepreneurial, reflected in higher incomes, and that motivation also helped them convince the jury to support them. Many other explanations are possible, which we do not necessarily need to understand to apply this method.

FIGURE 6.6 Sample discontinuity chart



When starting the program, the local microfinance bank decided that the threshold to receive a loan was 85, and all applicants were accepted or denied support accordingly. Now we'd like to know whether the microcredit program had any impact on incomes. As illustrated in figure 6.6 (right graph), we assume that those who received a score below 85 have the same outcomes as previously, while the income of those with a score of 85 and above increased across the board. From this information, it is possible to identify the impact of the program, which will be represented by the difference in outcomes (that is, the discontinuity of the linear relationship) near the cutoff.

When Can I Use a Discontinuity Design?

The discontinuity design can be used for both prospective and retrospective evaluations. That is, unlike the randomized techniques discussed above, it can also be used when the program is already underway or completed. The main requirement for this method is that program participation is determined by an explicitly specified targeting rule; in other words, by a continuous scale or score. For this method to work, however, we need many observations in the region immediately above and below the cutoff point in order to have sufficient numbers of youth that we can compare with one another. Unless the evaluation is done without baseline data or can take advantage of existing program records, a discontinuity design requires similar data collection as a lottery design, and thus has a similar cost.

Advantages

- The discontinuity method takes advantage of existing targeting rules and does not require any change in program design.
- It provides unbiased estimates for participants near the cutoff.
- It does not require randomization of any kind, so it may be more politically acceptable than other methods.
- It identifies potential effects of marginal scaling. For example, if a program is considering lowering the eligibility threshold from a score of, say, 85 to 75, a discontinuity evaluation can indicate what impact this will have on participants, providing information for a cost-benefit analysis of the proposal.

Disadvantages

- The method requires a very specific threshold for determining groups.
- Impact estimates are valid only for the margin near the cutoff and cannot be generalized to people whose scores are further away from the threshold. The technique does not provide an average impact for program participants.
- It requires large evaluation samples since only the observations around the cutoff can be used.
- As discussed in Duflo, Glennerster, and Kremer (2006), in developing countries, eligibility rules are rarely enforced strictly in the first place, and so there is a high chance that groups may not be distinct, which makes it difficult to obtain valid data using this method.

All in all, the discontinuity method is a good solution when the evaluation starts late or when randomization is not possible. However, it can be applied only in specific circumstances. Box 6.6 presents an example of a discontinuity design.

BOX 6.6 Example of a discontinuity design

Klinger and Schuendeln (2007) use a discontinuity design to study the role of entrepreneurial training on enterprise formation and enterprise outcomes in the context of the business plan competitions run by the NGO TechnoServe in Central America. The program provides training and business development services to help participants prepare a business plan, and it funds a selected number of the best plans.

The evaluators take advantage of the fact that to enter the program there is first a preliminary screening process that assigns applicants a score characterizing their potential entrepreneurial ability. The number of applicants that are admitted into the program is fixed before the competition begins. Applicants are accepted to the workshop if their score falls above the cutoff; if not, they are rejected. This allows for comparing beneficiaries who just received a passing score with those who failed to enter the program by a small margin. Since both groups have similar scores just above and just below the cutoff, it is fair to assume that they also share similar unobservable characteristics, which in turn allows for a high-quality counterfactual.

Statistical analysis confirmed that the eligibility rules were respected—that is, people were selected properly based on their score—and that outcome characteristics of applicants were continuous along their scores prior to the program. After the program, evaluators found a more pronounced change in outcomes around the cutoff. Based on the discontinuity design, in turn, they were able to show that the training increased the probability of opening a business by approximately 10 percent and the probability of expanding a business by more than 20 percent.

Method 5: Difference-in-Difference

In many programs, the selection of target areas and beneficiaries does not follow clear criteria. This can lead to highly selective targeting. For example, we may have prior knowledge about a specific community, better access to some places than to others, or existing partners that already have basic infrastructure in place that we would like to build on. Although there is nothing wrong with this in principle, such subjective targeting rules make it harder to develop a good counterfactual. Nevertheless, we may be able to get a rough estimate of a program's impact by using a difference-in-difference evaluation design.

How It Works

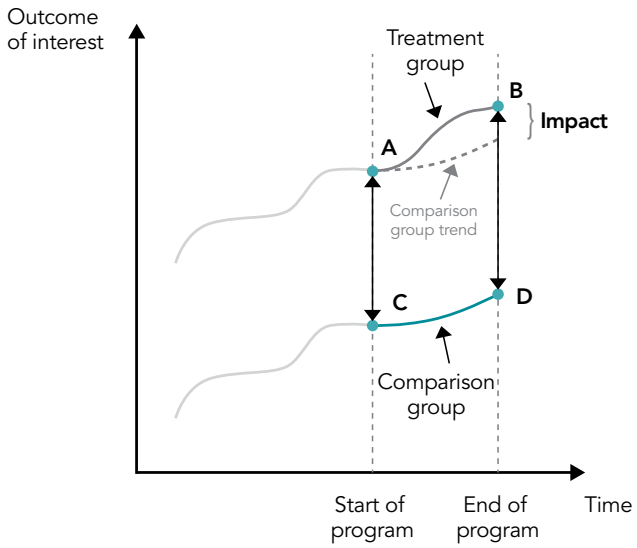
Identifying the comparison group. The difference-in-difference design is basically structured like "a pre-test/post-test randomized experiment, but it lacks its key feature, the random assignment" (Trochim 2006). In the difference-in-difference design, we try to identify a comparison group that we *believe* is similar to our pre-defined treatment group. For example, in center-based youth livelihood interventions, we may pick two comparable training centers or classrooms. In community-based programs, we may use two similar neighborhoods or districts. Either way, we always try to select groups that we think are as similar as possible so we can adequately compare the treated group with the comparison group. However, since the selection is not done at random, we can never be sure the groups are truly comparable—remember that there are unobservable characteristics that we cannot control for—thus, this methodology is also known as the *non-equivalent groups design* (Trochim 2006).

Estimating the impact. As we saw in note 5, simply comparing the outcomes of participants and subjectively selected nonparticipants does not give us the program's

impact, since both groups are most likely different from each other. Similarly, comparing program participants before and after an intervention is problematic as well because many other factors are also likely to influence the participant outcomes over time. But what if we combined both techniques and compared before-and-after changes in outcomes of both a group that enrolled in our program and of a group that did not participate?

Let's imagine a job-training program for youth. To apply the difference-in-difference evaluation technique, we need to measure outcomes (monthly income, for example) for both the treatment and comparison groups before the program begins (see figure 6.7, points A and C) and measure the outcomes of both groups after the program (points B and D). Since both groups are likely to be different from the outset, their incomes at baseline may also be different, but this does not immediately disqualify the method. The difference-in-difference technique compares the difference in outcomes between both groups at the end of the intervention (B minus D) with the difference in outcomes between both groups at the beginning (A minus C). Alternatively, we could compare the difference in outcomes for participants (B minus A) with the difference in outcomes for nonparticipants (D minus C). Subtracting these differences from each other yields a rough idea of the program's impact; it shows whether and how much the training program increased income for participants relative to those who did not participate.

FIGURE 6.7 Example of difference-in-difference analysis



Source: Adapted from Gertler et al. (2011).

[Tip]

A good test for whether it is realistic to assume equal trends between participants and nonparticipants is to compare their changes in outcomes before the program is implemented. If the outcomes moved in tandem before the program started, we can be more confident that their outcomes would continue this trend during the program. If, however, pre-program trends are different, the equal trend assumption may not be correct. Yet, knowing the difference in trends would at least allow us to control for that difference when computing the analysis.

Source: Adapted from Gertler et al. (2011).

The "equal trends" assumption. The underlying assumption of this method is that although the observed and unobserved characteristics of the treatment and comparison groups may be somewhat different (reflected in different levels of income at the beginning), their *differences are constant over time*, or time-invariant. This allows us to use the trend of the comparison group as an estimate for what would have happened to our treatment group in the absence of the intervention.

Is such an assumption realistic? Many observable characteristics, such as year of birth, gender, parent's education, and the like will probably not change over the

course of the evaluation. However, the same cannot be said about several unobservable characteristics, such as personality traits, an individual's intrinsic motivation, risk preferences and so on, which have been shown by numerous studies to change over time, especially in connection with development programs (see, for example, Robins et al. 2001, and Roberts, Caspi, and Moffitt 2003). Therefore, we can never be certain that the differences between the groups do not change over time, which, in turn, could bias our impact estimates. Even if the differences in participant characteristics remained constant, these differences could lead to interaction effects over time. If participating youth are, on average, more motivated than nonparticipants, then they could take better advantage of the program and, in turn, secure higher returns from their participation than nonparticipants would have. Moreover, external factors may influence both groups to a different extent during the implementation period. This would be the case if the municipality starts a new program in our treatment community but not in our comparison community, for example.

When Can I Use a Difference-in-Difference Design?

This design is best used in the absence of a clear targeting mechanism (such as random assignment or eligibility rankings). Since it assumes that the differences of participants and nonparticipants are constant over time, this method is most reasonably used when there are good data at multiple periods before the program begins. There should be at least three data collections: two prior to treatment, and at least one endline. This means that unless the data on participants and nonparticipants are available through other channels, such as an existing household survey, the costs of such an evaluation can be much higher than with other impact evaluation techniques.

Advantages

- The difference-in-difference design provides a way to account for differences between participants and nonparticipants.
- It controls for many individual effects.
- It does not require a prospective evaluation if the necessary data have already been collected.
- It is useful when combined with other methods to increase statistical power.

Disadvantages

- It produces less reliable results than randomized selection methods.
- It cannot be used alone without assuming the treatment and comparison groups change over time in the same way.
- It requires at least three data collections, whereas other methods need only two, so it can be more expensive.

See box 6.7 for an example of this design.

BOX 6.7 Example of a difference-in-difference method

Almeida and Galasso (2008) studied the short-run effects of a program to promote self-employment among workfare beneficiaries in Argentina. Following the severe economic crisis in 2001, the Argentinean government introduced a large workfare program, *Jefes*, including a program initiative to promote self-employment called *Microemprendimientos Productivos* (Productive Microenterprises). The microenterprise program provided in-kind grants to finance inputs and equipment as well as technical assistance through periodic visits of tutors.

To evaluate the impacts of the program in the absence of experimental data, Almeida and Galasso used a difference-in-difference framework. This approach compared the labor market outcomes for program participants before and after the intervention with those of nonparticipants. In order to identify a valid comparison group, they took advantage of the program's promotion campaign, during which *Jefes* beneficiaries could sign up to declare interest in the program. By restricting the comparison group to those who had shown interest in the microenterprise initiative (but eventually did not participate), the authors aimed to minimize the problems of comparing individuals interested in self-employment (for example, due to their entrepreneurial ability or motivation) with those who were not.

A baseline household survey was administered to 309 participants and 244 nonparticipants in November 2004. SIEMPRO, the Argentinean public monitoring and evaluation agency for poverty programs, administered the survey. The same households were re-interviewed one year later, at the end of 2005. With only two data collections available, the evaluators had to assume that in the absence of the program, participants and nonparticipants would have had comparable trends in labor market outcomes (the "equal trends" assumptions).

The findings indicated that, given the relatively low participation rate, jumpstarting self-employment through start-up capital and business training is not necessarily an attractive option for all workfare beneficiaries. Moreover, although the program increased the number of working hours of participants, it failed to significantly increase their average income. Finally, not everyone benefited from the program to the same extent, with positive effects measured only for the more educated participants.

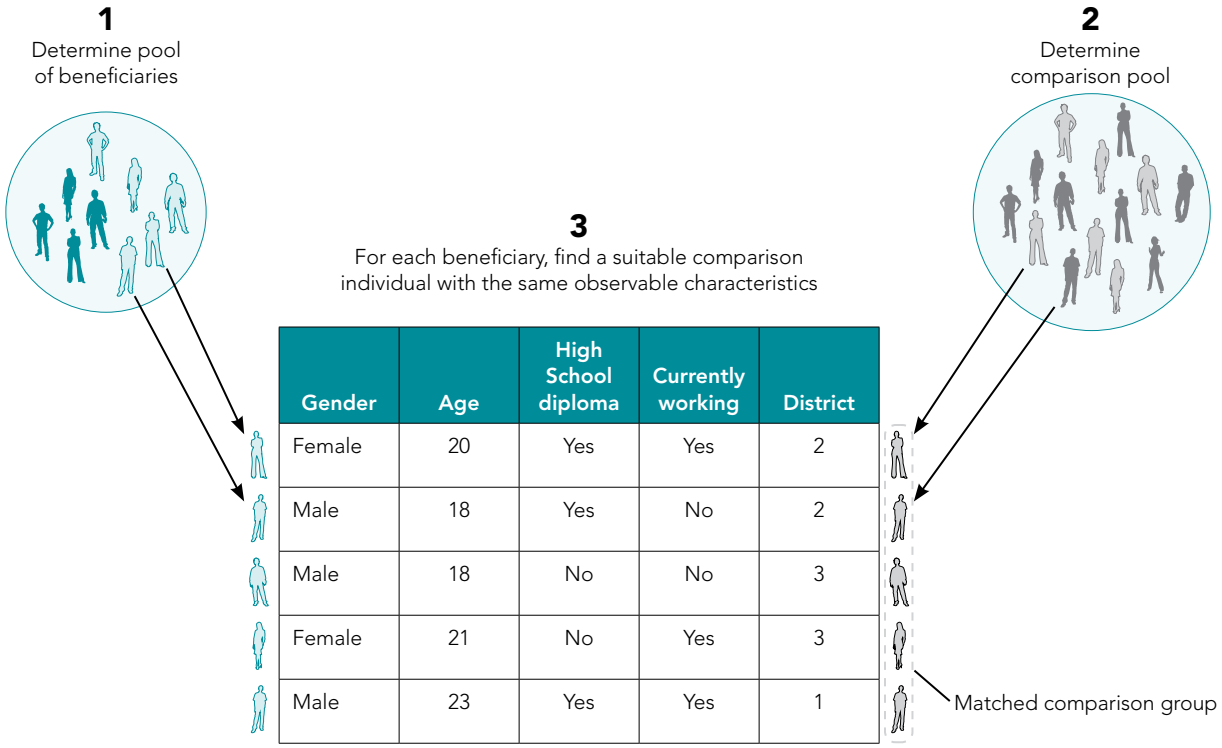
Method 6: Matching

As with the difference-in-difference design, matching is used in the absence of other strict program assignment rules. In the past, matching was popular with program evaluation specialists, but it has become eclipsed by more robust methods, such as those described above.

How It Works

The matching method pairs youth participating in a program with nonparticipants based on observable characteristics (age, gender, level of education, employment status, residency, and other factors). That is, for every individual youth (or group of youths) in the treatment cohort, matching constructs an artificial comparison unit that has as many similar characteristics as possible (see figure 6.8). This statistical technique tries to simulate a comparison group that otherwise does not exist. Ultimately, the average outcomes of those receiving treatment can be compared with the outcomes of the comparison group, and their difference yields the impact of the intervention.

FIGURE 6.8 Exact matching on five characteristics



Identifying a good match for each program participant requires finding those characteristics that explain an individual's decision to enroll in the program. Unfortunately, this is not as easy as it may sound. As Gertler and colleagues (2011) point out, if the list of relevant characteristics is small (as in figure 6.8, above), we will probably find a match for each youth of the treatment group, but each match may not be particularly precise and we run the risk of leaving out other potentially important criteria. If, on the other hand, we want to match based on a large number of characteristics (adding, for example, parents' level of education, test scores, and income level), it may be hard to identify a match for each of the units in the treatment group unless the number of observations in our database of comparison youths is very large.

When Can I Use Matching?

Matching techniques can be used in a variety of settings, regardless of a program's coverage or targeting criteria. In practice, it is often used when none of the other evaluation designs is feasible, especially when the evaluation starts after implementation. Given its inability to control for unobserved characteristics, however, matching is preferably used with one of the other evaluation techniques. Also, in order to match properly, we usually need a large sample size to ensure a matchable comparison group can be found (see box 6.8). If data required have not been collected through other channels, the evaluation may be significantly more costly than other methods described in this note.

[Tip]

The challenge of finding pairs in treatment and comparison groups with many comparable characteristics can be overcome by using a technique called propensity-score matching. Instead of matching treatment and comparison units based on the same characteristics for all selected criteria, propensity-score matching computes the likelihood (the propensity score) of each youth enrolling in the program based on several observed characteristics. Once the propensity score (a number between 0 and 1) has been computed for all participants and nonparticipants for whom data are available, participants are matched with those nonparticipants that have the closest score. These matched nonparticipants then form the comparison group.

BOX 6.8 Steps for applying a matching technique

1. Identify youth that enrolled in the program and that did not.
2. Collect in-depth information on observable characteristics (such as age, gender, level of education, employment status) of enrolled and non-enrolled youth through a baseline survey or by consulting existing data.
3. Using a statistical matching technique such as propensity-score matching, match each participant with a similar nonparticipant.
4. Compare the outcomes of the enrolled youth and their matched comparisons. The difference in outcomes is the impact of the program on that particular individual.
5. Calculate the estimated average impact of the program by taking the mean of the individual impacts.

Advantages

- Matching allows for comparison of outcomes between similar people.
- It can be used with other techniques to validate the quality of the comparison group.

Disadvantages

- Because matching requires direct comparisons of people, a large sample survey may be needed in order to draw an appropriate comparison group.
- Matching can be performed on observable characteristics only. Unobservables, or traits that are very hard to observe, such as personality, motivation, family support, and so on, cannot be incorporated in this technique. It therefore requires an assumption that there are no systemic differences in unobserved characteristics between treatment and comparison groups, which is often implausible. If this assumption does not hold, matching may lead to bias in estimating the impact of the program.
- It may not be possible to find an appropriate match for everyone in the treatment group, impairing the external validity of the impact estimate.

For an example of matching, see box 6.9.

BOX 6.9 Example of matching

Jaramillo and Parodi (2003) used propensity-score matching to evaluate the youth entrepreneurship program implemented by the Peruvian NGO Colectivo Integral de Desarrollo. To estimate the impact of the business plan competition and the subsequent support services consisting of training, follow-up support, and internships on participants, the evaluators constructed a comparison group consisting of those youth who had participated in preparatory activities of the program (pre-training) but either did not join the business plan competition or did not present winning proposals.

The evaluators calculated the probability of an individual's participation in the program based on observable characteristics such as age, gender, level of education, and marital status. Each beneficiary was then matched with someone from the comparison group that had a similar propensity score. The comparison of outcomes (in terms of business sustainability, number of jobs created, income) between the beneficiaries and their matched peers was then used to estimate the impact of the intervention.

However, since the matching could be based only on observable characteristics, there was a realistic chance that the positive effects identified in the evaluation were an overestimate of the actual impact of the intervention. In fact, youth who successfully participated in the business plan competition were likely to be different from youth in the comparison group, for example, in terms of their motivation or skills level, and may have been more successful entrepreneurs than their peers even without participating in the entrepreneurship program.

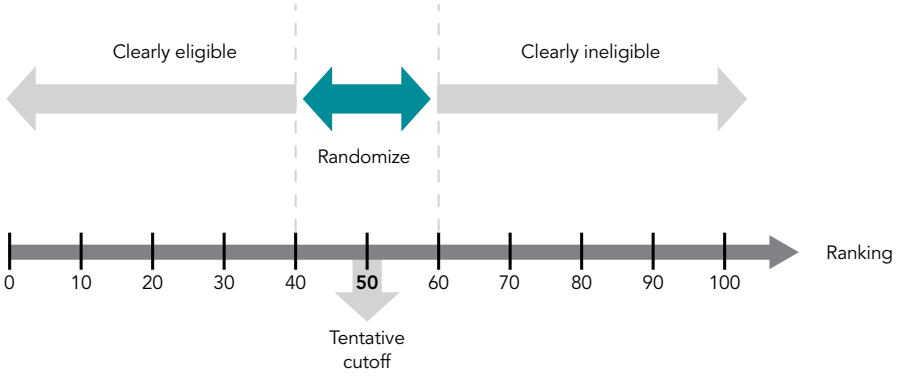
Combining Methods

As we have seen, some methods are stronger in constructing a counterfactual than others. In particular, it may be hard to find good comparison groups when the evaluation is not planned from the beginning of the program. Combining methods may offset some of the weaknesses of a single technique and increase the validity of the estimated counterfactual.

Randomized Discontinuity Design

This technique combines a discontinuity design with randomized assignment. If a cutoff is not clearly designated, or if it is not sufficiently justifiable, it is possible to randomize around the cutoff. In this case, those youths who are clearly eligible are still given the program, while those clearly not eligible are not given the program (see figure 6.9). Only a group near the threshold is selected for randomization. With this method, some of those who otherwise may not have received the program may now receive the program, and vice versa. As in a normal discontinuity design, the results are valid only for those participants at the margin of acceptability. However, given the partial randomization, we can be more confident that the treatment and comparison groups share the relevant characteristics, and we need a smaller sample size to find statistically significant results. The analysis is then done in the same way as any randomized design. The average outcome of those in the treatment group is compared with the average outcome of those in the comparison group, and the difference is the causal impact of the program on those selected.

FIGURE 6.9 Spectrum of eligibility (example of a poverty score ranking)



Note: A lower score represents a higher level of poverty.

Difference-in-Difference or Matching Combined with Randomization

The difference-in-difference technique assumes that those in the treatment and comparison groups are very similar, or at least that their differences are constant over time. Likewise, matching assumes that having similar observable characteristics justifies a comparison between two individuals. Randomization does not require either of these assumptions in order to estimate the impact of a program. However, randomization can be improved when used in conjunction with either or both of these methods. By minimizing differences between those compared, both difference-in-difference and matching methods increase statistical power without the need to increase the number of participants. By combining nonrandom methods with random methods, survey costs can be reduced.

Difference-in-Difference Combined with Matching

If no type of randomization or discontinuity design is feasible, another possibility is to combine the difference-in-difference with the matching technique, thereby mitigating some of the weaknesses both methods have when used on their own. Since the difference-in-difference technique cannot guarantee that treatment and comparison groups are equivalent, combining it with simple matching or propensity-score matching can at least ensure that both groups are very similar in terms of observable characteristics.

For an overview of the standard evaluation methods, see table 6.1.

[Tip]

In practice, the lead evaluator must assess whether it would be useful to combine methods. Practitioners therefore do not need to worry about the details of combined approaches but should be aware that this may be a way to get more reliable impact estimates.

TABLE 6.1 Overview of impact evaluation techniques

Methodology	Description	Comparison Group	Required Assumptions	Data Needed	When to Use?
Lottery	A sample of eligible individuals is randomly assigned into those who receive the intervention and those who do not. Impact is the difference in outcomes between the two groups.	Those selected by lottery	• Randomization is successful and complied with. • The two groups are statistically identical on observed and unobserved factors.	• Post-intervention data for treatment and comparison groups • Baseline data are desirable	• If study can be designed before the program begins • If resources are scarce and it is important to ensure that fair methods are used to enroll needy people into the program • If the comparison group will never get the program for the length of the evaluation
Random Phase-In	Eligible individuals are assigned to treatment tranches and receive the program sequentially.	Those waitlisted	• Randomization is successful and complied with. • Those in the later program phases will not significantly change their behavior while waiting to participate in the program.	• Post-intervention data for treatment and comparison groups • Baseline data are desirable	• If study can be designed before the program begins • If resources are scarce and it is important to ensure that fair methods are used to enroll needy people into the program • If program is rolled out over time
Random Promotion	A random set of individuals or groups is encouraged to enroll in the program. Impact is measured by comparing the average outcomes of those who were encouraged to participate with the outcomes of those who were not.	Those who did not receive the promotion	• Randomization is successful and is complied with. • There will be differential take-up between those who receive promotion and those who do not.	• Post-intervention data for treatment and comparison groups • Baseline data are desirable	• If study can be designed before the program begins • If nobody can be excluded from the program • If participation is voluntary and more people will participate in the program if the program is promoted to them

TABLE 6.1 (CONT'D) Overview of impact evaluation techniques

Methodology	Description	Comparison Group	Required Assumptions	Data Needed	When to Use?
Discontinuity	Individuals are ranked based on specific, measurable criteria. There is a cutoff that determines who is eligible to participate. Outcomes of participants and nonparticipants close to the cutoff line are then compared, and the eligibility criterion is controlled for.	Those close to the cutoff who were not eligible	• Those near either side of the cutoff are very similar in observed and unobserved characteristics.	• Post-intervention data for those nearest the cutoff • Baseline data are desirable	• If randomization is not possible or the evaluation starts after the program begins • If the selection is based on a continuous ranking with cutoff
Difference-in-Difference	Outcomes of program participants and nonparticipants are compared before and after the intervention. The relative change in outcomes is the impact of the program.	Non-equivalent group of individuals who did not participate in the program, but for whom data were collected	Over time, those in the treatment group do not change in a fundamentally different way than those in the comparison group.	• Baseline and follow-up data • Data from at least three time-periods desirable (at least two must be before the program begins)	• If the study starts after the program begins • If nonparticipants who are similar to participants can be identified • If fairness of selection into the program is not considered an issue • Best used in combination with other methods
Matching	Individuals in the treatment group are matched with nonparticipants who have similar observable characteristics.	Exact matching: For each participant, at least one nonparticipant who is identical on selected characteristics Propensity-score matching: nonparticipants who have a mix of characteristics that predicts that they would be as likely to participate as participants	Researchers can identify all of the relevant characteristics of people through a survey.	• Large survey (census, DHS, LFS, etc.), ideally in combination with program-based household survey (ideally two observations of both)	• If the study starts after the program begins • If nonparticipants who are similar to participants can be identified • If fairness of selection into the program is not considered an issue • Best used in combination with other methods

Key Points

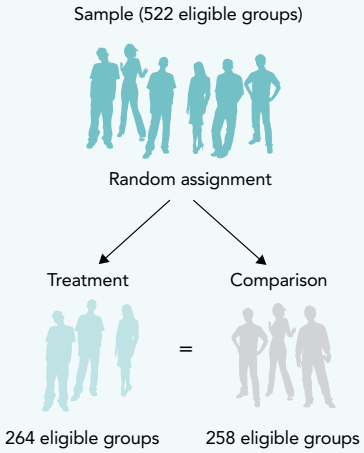
1. Only a selected range of impact evaluation methods allow for obtaining a reliable counterfactual and trustworthy results.
2. Lottery designs, randomized phase-in, randomized promotion, and discontinuity-designs all produce estimates of the counterfactual through explicit program assignment rules. Difference-in-difference and matching methods offer the evaluator additional—though less accurate—tools for impact evaluation when the evaluation starts after implementation and when eligibility criteria are less clearly defined.
3. No single method is best for every program. The best method depends on the operational context (i.e., timing, coverage, and targeting) of the program. Therefore, program managers need to discuss the programmatic constraints with the evaluation specialist because these constraints will affect the feasibility of different evaluation designs.
4. Whenever possible, it is highly desirable to plan the impact evaluation before the program is implemented. Retrospective evaluations tend to be less robust and may not be possible at all if the necessary data was not collected through other channels.
5. In some cases, the methods described here may not be feasible because of budget requirements, timing constraints, or political issues.

NUSAF Case Study: Selecting a Lottery Design

The NUSAF Youth Opportunities Program impact evaluation was developed during program preparation. Because the number of eligible applicants to the program far exceeded the program's funding capacity, the impact evaluation design hinged on the availability of a large pool of eligible but unfunded applications that had been submitted for Youth Opportunities Program funding. Given this large oversubscription to the program, NUSAF management and the program coordinators determined that selection of beneficiaries through a lottery system was not only feasible but also provided a fair and transparent mechanism to allocate funding among equally qualified youth group applicants.

NUSAF District Technical Officers were instructed to verify applications for the minimum set of technical criteria required for eligibility and to conduct field appraisals on programs that would be selected for funding. A list of eligible and verified programs was sent to the Project Management Unit for onward submission to the impact evaluation team, which conducted the lottery for selection of funded proposals. In each district, 30–60 percent of the eligible groups were selected for funding, dependent on budget limitations for that particular district.

Once the complete list of applicants was received from the District Technical Officers, the random assignment of applicants to treatment and comparison groups was completed all at once for each district. Each applicant group was assigned a random number using a random number generator. Groups were then sorted from first to last based on the random number. The sum of the program costs was calculated. Starting from the first randomly selected project, projects were awarded funding until the pools of available resources for that district were exhausted. All other projects remained unfunded and were assigned to the comparison group.



Through this process, a total of 264 projects were selected for funding, comprising the treatment group. The remaining pool of 258 eligible projects not selected for funding made up the comparison group. For the purposes of the impact evaluation, the generation of an equivalent comparison group allowed for the estimation of the counterfactual, the condition that the treatment group would have experienced in the absence of treatment.

Source: Blattman, Fiala, and Martinez (2011).

Key Reading

Duflo, E., Glennerster, R. and Kremer, M. 2006. "Using Randomization in Development Economics Research: A Toolkit." BREAD Working Paper No. 136. (For advanced readers.) http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf.

Gertler, P., Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. (Chapters 4–8 are relevant to this note.) http://www.worldbank.org/ieinpractice

Khandker, S., Koolwal, G., and Samad, H. 2010. *Handbook on Impact Evaluation: Quantitative Methods and Practices*. Washington, DC: The World Bank. (For advanced readers. Chapters 3–7 are relevant to this note.) http://www-wds.worldbank.org/external/default/WDSContentServer/IW3P/IB/2009/12/10/000333037_20091210014322/Rendered/PDF/520990PUB0EPI1101Official0Use0Only1.pdf

Notes