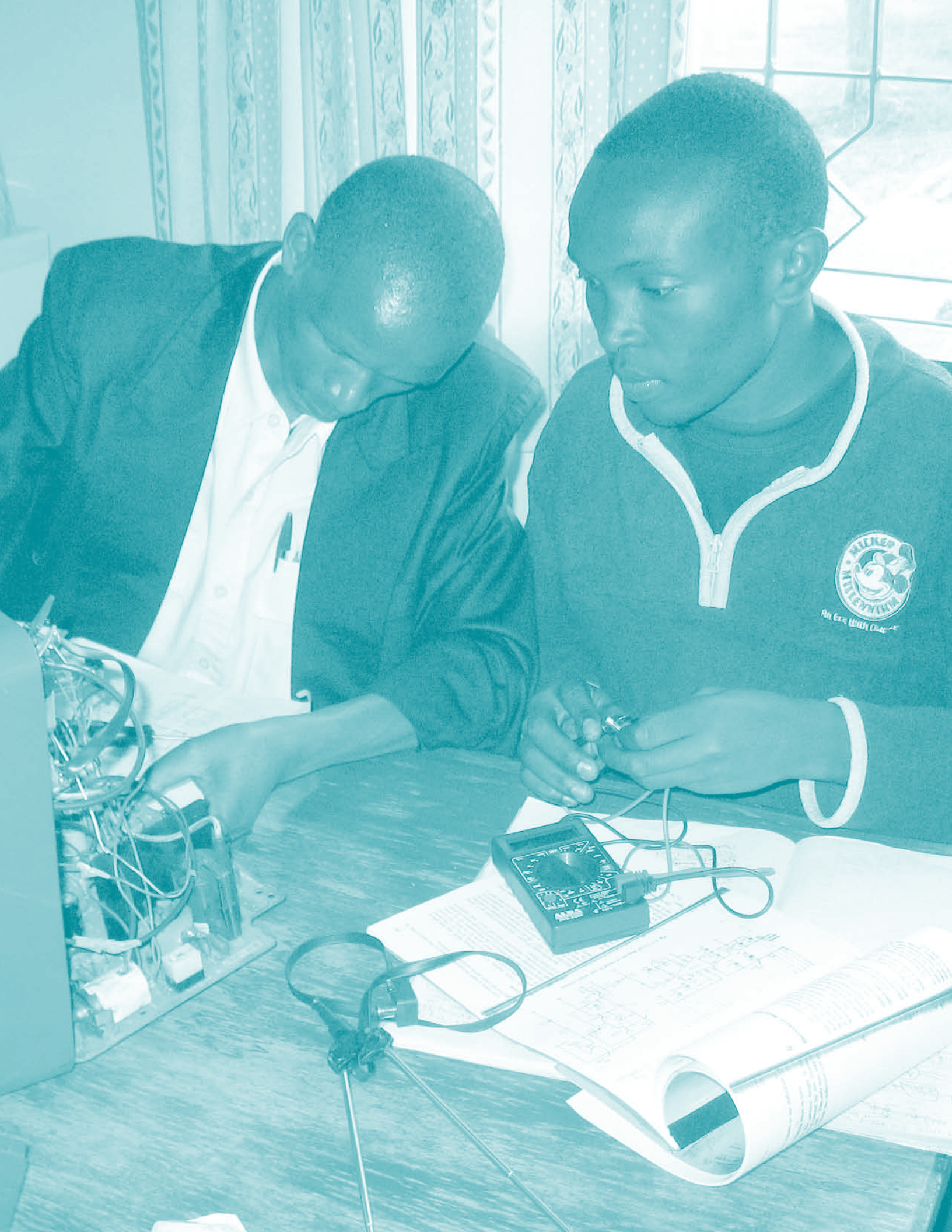


NOTE 5: Proving Program Impact

Rigorous skepticism is a creative force
because most damage is done by overconfident people
who thought they knew the answer when they didn't.

— William Easterly

Good intentions are not enough. Instead, we need to know that we are actually improving people's lives and not causing more harm than good without even being aware of it. Proof is provided by impact evaluations, which, unlike other evaluation types, provide scientific evidence of a program's effectiveness.

In this note, we explore the fundamental impact evaluation question: "How can we be sure that the changes in outcomes we see result from our intervention?" We show that measuring impact requires estimating what would have happened in the absence of the program. These estimates can be made by identifying a comparison group through experimental or quasi-experimental evaluation techniques. We also show why the two most common techniques—comparing participants before and after the intervention and comparing participants with subjectively selected nonparticipants—cannot provide reliable estimates of program success.

The Attribution Challenge

Impact evaluations help us answer very specific questions about our program. As discussed in note 4, they try to answer whether an intervention (the cause) improves outcomes among beneficiaries (the effect). For example:

- Does our vocational training program increase trainees' incomes?
- Does our school-based entrepreneurship curriculum increase secondary school completion rates and students' interest in higher education?
- Does our start-up mentoring program foster business creation and sustainability?

Establishing causality between intervention activities and the outcomes we observe can be complicated because other factors may also influence the outcomes we are interested in. For instance, simply observing that business creation increased after our entrepreneurship program was implemented is not proof of our program's success because other factors such as local economic conditions or regulations about starting a business may have improved during the life of our program and contributed to business creation. Similarly, an observed decrease in business creation after our intervention does not necessarily mean that our intervention *caused* a decline in business start-ups; instead it may reflect a worsening of other external conditions.

The purpose of impact evaluations is precisely to overcome this attribution challenge by measuring to what extent a particular program, *and only that program,* contributed to the change in the outcomes of interest.

What Exactly Is "Impact"?

[Definition]

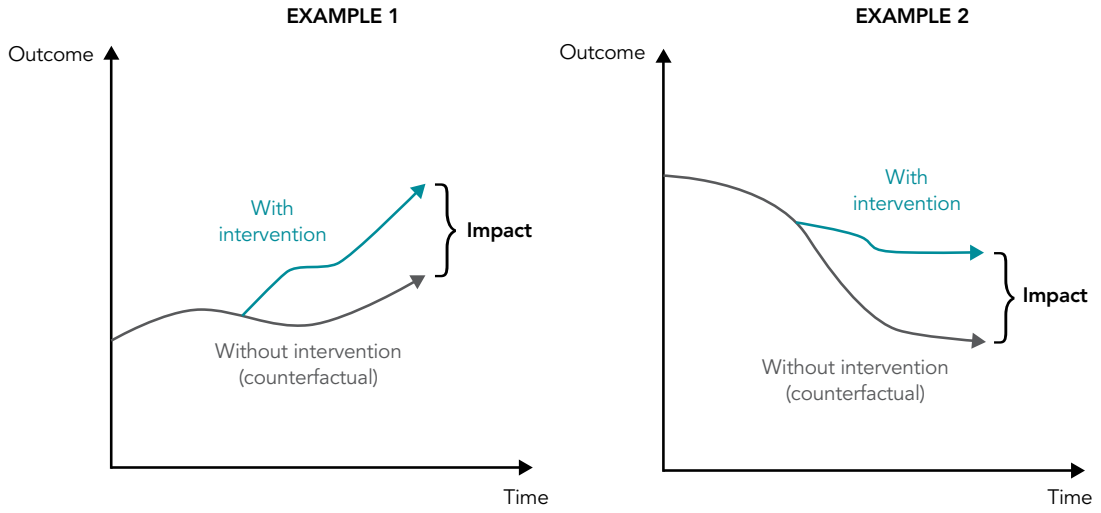
Outcome with the program
– Outcome in the absence of the program
= **Impact**

First, we need to clarify what we mean by *impact*. Often the term refers to higher-level program goals or outcomes relating to changes in overall living standards, such as reducing poverty or increasing the wellbeing of individuals and households. In the context of impact evaluations, however, *impact* is understood more narrowly as the change in outcomes that can be directly attributed to our program. The focus here is on "directly attributed," meaning that we want to know that the changes in outcomes we observe are truly due to our intervention and nothing else.

Simply speaking, as illustrated in figure 5.1, the impact of an intervention is the difference between

- the observed outcomes with the intervention, and
- the observed outcomes for the same individual, household, community, or other unit of observation without the intervention. The outcomes in the absence of the intervention is what we call *counterfactual,* referring to *what would have happened to the beneficiary if the program had not taken place.*

FIGURE 5.1 A visual illustration of program impact



For obvious reasons, it is impossible to observe the same person (household, school, etc.) with and without the intervention. Although we can observe outcomes for those youth that participate in our program, it is impossible to know what their situation would have been in the absence of the program. That is, we cannot know with certainty what would have happened to them if they had not participated in our program. As a result, we will never be able to get the real counterfactual, so an estimate must suffice.

How Can We Estimate the Counterfactual?

To estimate counterfactuals, we identify *comparison groups*, sometimes known as *control groups*. The group of program participants is known as the *treatment group*. A good comparison group has the same characteristics as the treatment group, except for the fact that comparison group members do not benefit from the program.

According to Gertler and colleagues 2011, treatment and comparison groups should share the same characteristics in at least three ways:

1. **They should be identical in terms of *observable* and *unobservable* characteristics.** Observable characteristics refer to age, gender, level of education, socioeconomic status, family characteristics, employment status, and the like. Unobservable characteristics include motivation, interest, preferences, the level of family support, and other factors. Although not every person in the treatment group must be identical to every person in the comparison group, both groups should be the same on average.
2. **Treatment and comparison groups should be expected to react to the program in the same way.** For example, outcomes, such as skills or income, should be as likely to increase for members of the treatment as for those in the comparison group.
3. **Treatment and comparison groups should be equally exposed to other interventions.** For example, both groups should have the same access to other support services provided by local government, NGOs, and so on.

When the above conditions are equal between the groups, then only the existence

[Definition]

A **comparison group** is a group that shares the same characteristics as the group of participants, except for the fact that the people in the comparison group do not benefit from the program. The terms comparison group and control group are often used interchangeably, though strictly speaking the latter is applicable only in the context of experimental evaluations (see below). For the purpose of this document, we will use the generic term *comparison group* throughout.

of the intervention will explain any differences in outcomes. In this case, the causal impact of the program can be demonstrated. If, on the other hand, the comparison group differs from the treatment group in significant ways, we are facing *selection bias*, which will make our impact measures invalid. Selection bias refers to the fact that underlying differences between the treatment and comparison groups by itself explains why we see different outcomes. Selection bias often occurs when the comparison group is made up of individuals who are either ineligible for the program (based on observable characteristics) or who chose not to participate (for unobservable reasons).

[Definition]

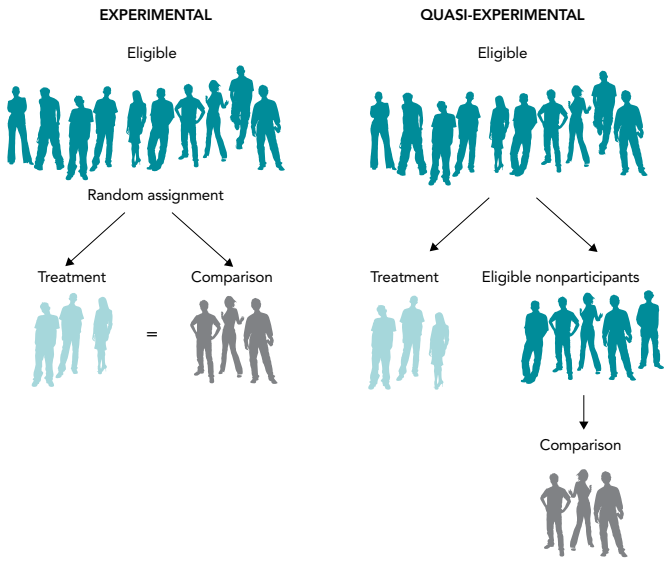
Selection bias usually occurs when program participants and nonparticipants differ in characteristics that cannot be observed, which affect both the individual's decision to participate in the program as well as the outcomes of interest.

In skills training and livelihood programs, it is likely that those who apply to participate are different from those who do not apply, and that these differences cannot be easily seen by the researcher. For example, applicants may be more motivated or have better information than non-applicants. These differences may also mean that applicants, on average, are more successful in the labor market than non-applicants *regardless of the training*. In that case, the better outcomes among training recipients may be due to these underlying differences and not to the training they received in the program.

Techniques to Find Good Comparison Groups

In general, there are two ways to make sure that the treatment and the comparison groups are as similar as possible: (1) with experimental techniques, and (2) with quasi-experimental techniques (see figure 5.2).

FIGURE 5.2 Experimental versus quasi-experimental techniques



Experimental Techniques

Experimental evaluation designs *randomize* who will be in each group. That is, if we have a group of potential beneficiaries (let's say 500 youth, 500 schools, etc.), we randomly select some of them (for example 250) to receive the program. This is the treatment group. The others will not receive the program; this is the comparison group. If randomization is carried out correctly, it is likely that both groups are very similar (1) in observable and unobservable characteristics, (2) in the way they would respond to the program, and (3) in their exposure to other interventions. Evaluations using this

technique, or variations of it, are commonly referred to as randomized controlled trials. See box 5.1 for ethical considerations of randomization.

BOX 5.1 Is randomization ethical?

Some programmers are reluctant to randomly assign potential beneficiaries into treatment and comparison groups. The general concern is that the evaluation leads to withholding seemingly obvious benefits (such as training opportunities) to needy individuals, which would be unethical. In reality, however, it is wrong to assume that one would be denying a benefit if a program has not yet been properly evaluated. In programs that have not been evaluated, random assignment may in fact be more ethical than other selection methods for the following reasons:

- **Uncertainty of program impact.** For most programs, it is not clear if the program has a positive impact on the individual and the community, or if that impact is of a size that justifies the resources being spent. An intervention may in fact have zero impact or even unintended negative side effects. For instance, programs geared toward girls at the exclusion of boys may increase gender violence. A microfinance program for youth may leave participants worse off if they are not able to repay their loans. Even a training program, if designed poorly, may actually decrease job prospects. Where a positive impact is achieved (e.g., a $100 increase in income per participant), it may come at a very high cost (e.g., $1,000 per person), suggesting that the money would be much better spend elsewhere. Thus, in the case of interventions whose impact and cost-benefit structure has not yet been sufficiently proven, it is well justified to evaluate the program based on treatment and comparison groups.
- **Budget constraints.** In reality, because of limited resources, it is rarely possible to serve everyone in need. That is, most programs provide benefits and services only to a limited number of beneficiaries, thereby excluding others, whether this is made explicit or not. For example, if a youth training program has a limited number of available spots, then some youth will receive the training while others will not. Similarly, if an intervention is carried out in one particular district, eligible youth in other districts are excluded. Randomization allows program officers to choose from the universe of potential participants in a way that is fair and that gives the same chance for participation to everyone. If the randomization is done in an open manner (for example as a lottery during a public event), it also enhances transparency in the selection process and may reduce fears in the population that selection was based on personal or political preferences.

It is also important to note that randomized evaluations do not necessarily require denying services to anybody. Note 6 will provide details on different evaluation techniques.

Quasi-Experimental Techniques

Randomization is not always feasible or desirable (see box 5.2). In such cases, quasi-experimental techniques may be used to isolate the effect of our intervention. Although they are usually less reliable than the experimental methods, quasi-experimental designs try to simulate the counterfactual by identifying nonparticipants that are as similar as possible to the treatment group. To do this, quasi-experimental methods usually rely on statistical tools and analysis. Some of the common methods are called discontinuity design, difference-in-difference, and matching (see note 6 for a detailed discussion).

BOX 5.2 Selected examples of when randomization is not possible

- The program has already started; beneficiaries have already been selected.
- Available resources are sufficient to serve all eligible members of the population. It may then be unethical to deny benefits or services only for the purpose of the study.
- We cannot select a comparison group or exclude anyone from the program. For example, a media campaign for financial literacy via TV or radio potentially reaches every household and it is impossible to monitor who listens and who does not.
- The intervention targets a limited number of groups or communities with unique characteristics.
- There is political opposition to providing an intervention to one group and not another.

When the conditions for a good comparison group are met, we say that the impact evaluation has internal validity (see box 5.3).

BOX 5.3 Internal and external validity

Ideally, impact evaluations will satisfy two requirements:

1. They will be *internally valid,* which means we will be able to show causality. To do so, we control for all possible differences between the treatment and comparison group, and are able to clearly attribute changes in outcomes to the intervention. To guarantee this, we use experimental or quasi-experimental techniques (discussed in detail in note 6).

2. They will be *externally valid,* which mean we will be able to generalize findings. That is, we can expect the same results if we provided the program to different or larger groups. To guarantee this, we need an appropriate strategy for choosing the sample of people we work with (this will be discussed in note 7).

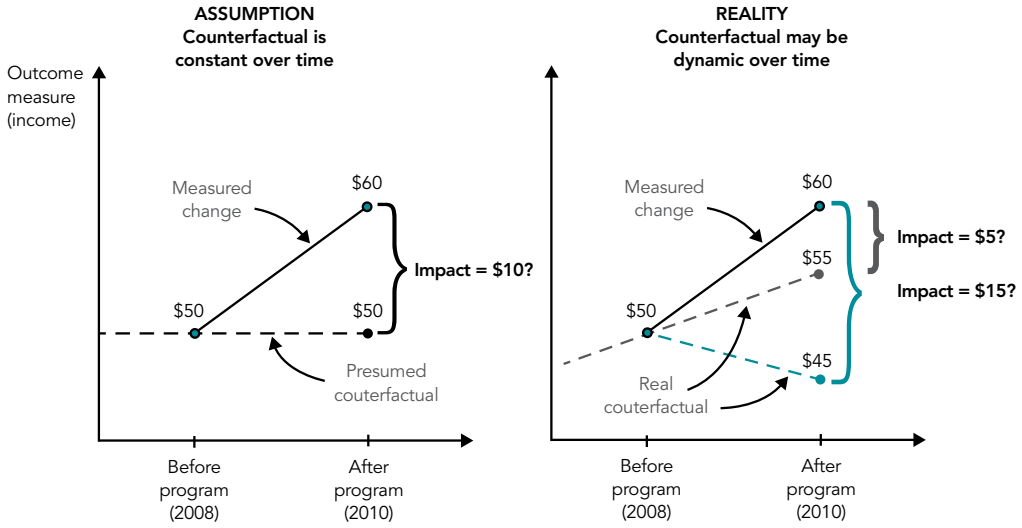
Counterfeit Counterfactuals

The two most common techniques for measuring success in our programs are comparing participants before and after the intervention, and comparing participants with subjectively selected nonparticipants. These techniques fail to identify a quality comparison group. As a result, they cannot be considered proper impact evaluation methods and their impact estimates are usually not credible. Here is why.

Counterfeit Counterfactual 1: Comparing Participants Before and After

In this technique, we use the pre-intervention outcome to estimate the counterfactual. Thus, we assume that if the program had never existed, the outcome for participants after the program would have been exactly the same as before the program. In the example of a training program, we may observe that the monthly income of participants increased from $50 before the program to $60 after the program. We may thus conclude that the impact of the program was $10 per month per person (see figure 5.3, left graph).

FIGURE 5.3 Risks in comparing before-and-after outcomes



The Problem

The assumption that in the absence of the program nothing would have changed is simply unwarranted in most cases. Many things can happen during the implementation period, particularly when programs last several years. For example, local economic conditions may improve, raising the number of available jobs and average incomes; positive weather conditions could raise yields and incomes in agriculture; or the local government could implement its own cash-for-work program, increasing incomes for many youth. If, indeed, the external environment improved independently of the program, then youth would have an increase in income anyway (say, $55 per month), and the real impact of our intervention would likely to be much smaller than estimated by a simple before-and-after comparison. In our example, the gain would be $5 instead of $10 (see figure 5.3, right graph). Conversely, if conditions actually worsened (say youth would earn only $45 in the absence of the program), then we would underestimate the true program impact using a before-and-after comparison.

Conclusion

Many factors can affect the outcomes of youth livelihood interventions over time. As a result, a pre-program outcome measure is almost never a good estimate of the counterfactual. For this reason, a before-and-after comparison is not considered a quality technique to demonstrate impact.

[Tip]

A specific case in which before-and-after comparisons can provide a fairly solid counterfactual is for targeted short-term interventions aimed at improving specific attitudes, knowledge, and skills (see, for instance, the before-and-after evaluation of the ILO *Know About Business* program in box 5.4 below). In that case, the outcomes can sometimes be realistically attributed to a selected intervention. However, other potential program impacts, such as behavior change, employment, and income are influenced by many factors and can thus not be accurately estimated by a simple before-and-after comparison.

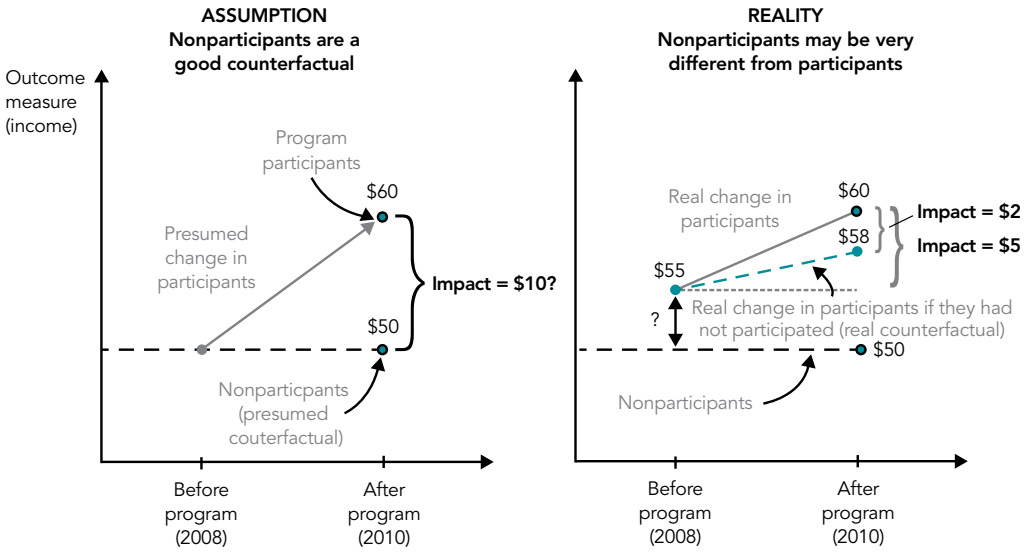
Counterfeit Counterfactual 2: Comparing Participants and Nonparticipants

In this technique, we observe the outcomes of subjectively selected nonparticipants at the end of the intervention to estimate the counterfactual. When comparing participants with nonparticipants, we assume that these groups are very similar in nature. For example, we trust that both groups share the same observable and unobservable characteristics, would react to the program in the same way, and are equally exposed to other interventions.

Using our example of a training program, we would measure the level of income of both participants and nonparticipants at the end of training. Assume we find that

participants earn $60 per month, while nonparticipants earn $50 per month. We then may conclude that our program impact was $10 per month per person (see figure 5.4, left graph).

FIGURE 5.4 Risks in comparing participants with nonparticipants



The Problem

There are two major problems with this approach. First, the assumption that both groups have equal levels of outcome at the beginning of the program may not be true. Participants may have been better or worse off before the program than the subjectively selected nonparticipants. If we measure outcomes only at the end on the program, we may not be able to learn baseline conditions. Participants may already have had a higher income at the beginning of the program than nonparticipants (e.g., $55) and thus the real change compared with our observation at the end of training ($60) would be $5 instead of $10 (see figure 5.4, right graph).

Second, an assumption that participants and nonparticipants are very similar is usually not true. Let's just think about our criteria for selecting young people in the program. Maybe it is on a first come, first served basis. In this case, those with better access to information about the existence of the program, those who live nearby, those who get encouraged by their parents, or simply those who are more motivated to participate would likely end up being part of the program. Alternatively, clear selection criteria such as test scores, interviews, or the quality of a business plan indicate that we explicitly want participants to be different from nonparticipants. In either case, and whether desired or not, participants and nonparticipants are likely to be different from one another on average; therefore, it is misleading to compare the two groups. In reality, given their potentially higher motivation, better access to information, proximity to services, and the like—characteristics that may not always be obvious to us—young people who participated in our program may very well have improved their situation even without the intervention. Going back to our example, if participants would have earned $58 after a certain period even without participating in our program, then their total earnings following the training ($60) would reflect a program impact of only $2, not $10 (see figure 5.4, right graph).

Conclusion

There are usually underlying reasons why some people participate in a program and some don't. These reasons make both participants and nonparticipants fundamentally different from one another, whether we can observe it (test scores) or not (family support, motivation). As a result, subjectively selected nonparticipants almost never represent a good counterfactual to understand how participants would have done in the absence of the program. Therefore, a simple comparison of participants and nonparticipants without using experimental or quasi-experimental techniques is not considered a quality technique to demonstrate impact.

Although the above counterfeit counterfactuals may not be useful to estimate impact—that is, to answer cause-and-effect questions—they may still be of value to our programs. In fact, collecting descriptive information about participants and even nonparticipants over time can be important for program management, since it may help us better understand the dynamics of our program. It is absolutely legitimate to use these types of comparisons as part of our monitoring or performance evaluation, as long as we are aware of what their results can and cannot tell us (see box 5.4 for examples.)

BOX 5.4 Selected examples of non-experimental evaluations

Technique: Before-and-after comparison

ILO Know About Business, Syria

Assessing the Effect of Know About Business (KAB) on the Knowledge and Attitudes of Secondary School Students (2007)
http://www.syriatrust.org/site/images/files/KAB_Schools_Report_0708.pdf

Technique: Comparing participants and nonparticipants

Junior Achievement, USA

The impact on students of participation in JA Worldwide: Selected cumulative and longitudinal findings (2004)
http://www.ja.org/files/long_summary.pdf

Key Points

1. The impact of a program is the change in outcomes that can be directly attributed to the intervention. Understanding impact requires that we isolate the effects of the program from other factors influencing beneficiary outcomes.
2. Measuring program impact requires a counterfactual, knowing what would have happened to our program participants in the absence of the intervention.
3. In order to estimate what would have happened to beneficiaries in the absence of the program, we construct comparison groups that share as many characteristics with the beneficiaries as possible. If a good comparison group can be identified, comparing outcomes between the comparison group and the beneficiaries (treatment group) yields the impact of the program.
4. Impact evaluation techniques to find valid comparison groups can be classified as one of two types. Experimental techniques randomly separate the eligible population into those who receive the program and those who don't. Quasi-experimental techniques try to find a valid comparison group among nonparticipants, mirroring the treatment group as closely as possible.

5. Simple before-and-after comparisons as well as comparing participants with subjectively selected nonparticipants do *not* provide credible impact estimates. The first fails to control for changes in external factors over time, the second fails to control for (often unobservable) characteristics that influence program placement. However, both can be useful for providing descriptive information as part of our monitoring system.

NUSAF Case Study: Identifying a Counterfactual

Identifying the counterfactual was an especially important concern for the Youth Opportunities Program evaluation. Are NUSAF participants different from the general population? If so, how could a counterfactual be drawn from them?

The government and research team expected that there would be important differences between the NUSAF participants and the general population. One clue was that individuals were supposed to form groups and submit proposals. This meant the applicants would need to be at least somewhat educated, implying they are better off. In addition, those who submitted proposals to the program had to want to be engaged in business, so they were probably very motivated. This is not a characteristic that is easily measured.

To verify potential differences, NUSAF looked at the characteristics of program participants collected at baseline and compared them to other youth surveyed around the same time. It was found that Youth Opportunities Program members owned significantly more assets and were much more educated than the general population. Additionally, women were highly underrepresented in the program (33 percent) compared with the general population (51 percent). Households in the study were five times more likely to own a radio or bicycle and three times more likely to own a mobile phone or cattle than the general population. There were also disparities in education among program participants and the general population.

In addition, by comparing rates of poverty in the general population with those of program participants, striking differences were found: at least 50 percent of the participants were above the defined levels of poverty. Thus, whether considering relative or absolute difference between the general population and Youth Opportunities Program participants, it was evident that applicants to the program, on average, were part of higher socioeconomic strata than a representative sample of youth in the region.

The differences across groups underlined why a careful impact evaluation was necessary. Using the general population as a counterfactual would greatly overestimate the effect of the program, as there were already major differences in the sample populations even without the program.

To identify a valid counterfactual, the evaluation team could take advantage of the fact that there was a very high demand for the program, but few remaining funds. The problems with identifying an appropriate comparison group outlined above, along with the lack of funds to ensure everyone who was eligible could participate, led to the decision to use randomized methods in order to select the comparison group.

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

Key Reading

Gertler, P., Martinez, S., Premand, P., Rawlings, L., and Vermeersch, C. 2011. *Impact Evaluation in Practice*. Washington, DC: The World Bank. See Chapter 3. http://www.worldbank.org/ieinpractice

Notes