Guide on Measuring Decent Jobs for Youth
Monitoring, evaluation and learning in labour market programmes

Impact evaluation methods for youth employment interventions
Guide on Measuring Decent Jobs for Youth
Monitoring, evaluation and learning in labour market programmes

NOTE 5
Impact evaluation methods for youth employment interventions
Contents

The attribution challenge 2
Challenges specific to evaluating youth-focused active labour market programmes 7
Quantitative methods of impact evaluation 10
Randomization – lottery design 12
   How it works 12
   When can I use a lottery design? 15
   Advantages 15
   Disadvantages 15
Adapting random designs to different contexts 18
   Randomized phase-in design 19
   Randomized promotion design 20
Difference-in-differences (DID) 22
   How it works 22
   When can I use a DID design? 24
   Advantages 24
   Disadvantages 24
Propensity score matching (PSM) 26
   How it works 26
   When can I use PSM? 28
   Advantages 28
   Disadvantages 28
Regression discontinuity design (RDD) 31
   How it works 31
   When can I use RDD? 32
   Advantages 32
   Disadvantages 32
Simple comparisons: Before and after 34
Improving the relevance of quantitative impact evaluations 36
   Measuring a variety of impacts 36
   Combining quantitative and qualitative approaches 37
Key points 41
Key resources 41
References 42
Case study: Assessing rural micro-enterprise growth through different evaluation methods 43

Tables
   5.1 Sample size required to detect significant impacts 9
   5.2 Overview of different impact evaluation methods 11
   5.3 Categories of impact evaluation questions 38

Figures
   5.1 A visual illustration of intervention impact 3
   5.2 Consider all possible evaluation methods during a planning stage 10
   5.3 Steps in a lottery design 12
   5.4 Choosing samples for small and large programmes 14
   5.5 Evaluation design 16
   5.6 Take-up of the El Mashrou3 show 21
   5.7 Example of difference-in-differences analysis 23
   5.8 Comparing participants and non-participants 27
   5.9 Impact evaluation design (simplified) 29
   5.10 Impacts on labour market outcomes 30
   5.11 Sample discontinuity chart 31
   5.12 Discontinuity in the probability of districts participating in the programme 33
   5.13 Comparing before-and-after outcomes 35
NOTE 1.

DIAGNOSING, PLANNING AND DESIGNING YOUTH EMPLOYMENT INTERVENTIONS
Impact evaluation methods for youth employment interventions

Prerequisites:
A basic understanding of quantitative research methods would be helpful. This note describes a number of commonly used impact evaluation methods and explains the advantages and disadvantages of each, including theoretical and practical considerations.

Learning objectives:
At the end of this note, readers will be able to:

- appreciate the major considerations and challenges to be taken into account when seeking to establish impact by asking: “What would have happened to the same people/household/community if the intervention had not taken place?”
- construct a counterfactual to estimate the change in outcomes that can be attributed to an intervention, and identify the key characteristics that treatment and comparison groups must share to ensure internal validity
- weight the pros and cons of different evaluation techniques and how they aim to eliminate selection bias
- comprehensively understand different quantitative research methods, from fully randomized designs to quasi-experimental methods, such as difference-in-differences, propensity score matching and regression discontinuity design
- use qualitative methods, to not only find out “what” happened – determining the average treatment effect of the intervention – but “why”.

Keywords:
Attribution, before-and-after comparison, comparison group, counterfactual, difference-in-differences, external validity, internal validity, lottery design, process tracing, propensity score matching, randomized phased-in design, randomized promotion design, regression discontinuity design, treatment group.
This note provides practitioners with an overview of the different tools available for an impact evaluation and offers guidance on which ones to select under specific circumstances, and how to implement these tools to assess the effects of youth employment interventions. While impact evaluations can be based on both quantitative and qualitative methods, this note focuses primarily on quantitative methods and introduces qualitative methods as a valuable complement in the context of mixed-methods approaches.

The attribution challenge

Before moving on, we need to clarify what we mean by impact. In previous notes within this guide, we have used the term as synonymous with higher-level goals or outcomes relating to changes in a young person’s employment situation such as reducing unemployment or increasing the well-being of individuals and households. In the context of impact evaluations, however, we understand impact more narrowly as the change in outcomes (e.g. employment status, working time, earnings) that can be attributed to our intervention.

As discussed in Note 4, impact evaluations try to answer cause-and-effect questions; that is, whether an intervention (the cause) improves outcomes among beneficiaries (the effect). For example:

- Can observed changes in trainees’ likelihood of securing employment be attributed to our vocational training intervention?
- Does our job counselling intervention lead to a higher level of satisfaction among employers and a higher job retention rate?
- Does our start-up mentoring intervention foster business creation and sustainability?

The labour market outcomes that we are interested in are determined by many complex factors, such as the overall social and economic development context, changes in political and/or personal circumstances, etc. Hence, establishing the degree to which changes in such outcomes can be attributed to a particular intervention is challenging.

The purpose of impact evaluation is precisely to overcome this attribution challenge by measuring the extent to which a particular programme, and only that programme, contributed to the change in the outcomes of interest.

**DEFINITION**

**Attribution:** The ascription of a causal link between observed (or expected to be observed) changes and a specific intervention.

---

This note draws on materials originally developed by Duflo et al. (2006), Khandker et al. (2010), and Gertler et al. (2016), adapting some of the material and illustrations to the youth employment field and providing a more concise presentation of impact evaluation methods.
In other words, impact evaluations try to assess whether, to what extent and why observed changes in outcomes of interest can be attributed to an intervention or project.

The focus of this note is the so-called counterfactual framework on which quantitative impact evaluations are typically based. This approach defines the impact of an intervention as the difference between the observed outcomes under the intervention and the so-called counterfactual scenario: “What would have happened to the same people/household/community if the intervention had not taken place?”. Figure 5.1 visualizes the concepts of impact and counterfactual.

In practice, the real counterfactual is impossible to measure. Impact evaluation methods try to quantify causal effects through estimating or constructing the counterfactual typically – though not always – with comparison groups, sometimes known as control groups. The group of participants is known as the treatment group or participant group.

Depending on the intervention being evaluated and its context, an impact evaluation design that mixes qualitative and quantitative methods is usually more appropriate, as explained in greater detail below.

**DEFINITION**

**Counterfactual:** The counterfactual describes what a certain outcome would have been for a programme participant in the absence of the programme. By definition, the counterfactual cannot be observed directly. Therefore, it must be estimated, for example using comparison groups.

---

**FIGURE 5.1: A VISUAL ILLUSTRATION OF INTERVENTION IMPACT**

**EXAMPLE 1**

- **Outcome**
- **Time**
- **Without intervention (counterfactual)**
- **With intervention**
- **Impact**

**EXAMPLE 2**

- **Outcome**
- **Time**
- **Without intervention (counterfactual)**
- **With intervention**
- **Impact**
Treatment and comparison groups should share the same characteristics in at least three ways (Gertler et al., 2016):

1. They should be similar in terms of both observable and unobservable characteristics: Observable characteristics may include age, gender, level of education, socio-economic status, family characteristics, employment status, and the like. Unobservable characteristics could include motivation, interest, values and ideologies and the level of family support, among other factors. Not every person in the treatment group must be identical to every person in the comparison group, but both groups should share similar average characteristics.

2. Treatment and comparison groups should be expected to react to the intervention in a similar way: For example, outcomes, such as skills or income, should be as likely to increase for members of the treatment group as for those in the comparison group.

3. Treatment and comparison groups should have similar levels of exposure to other interventions: For example, both groups should have the same access to other support services provided by local government, NGOs, etc.

4. When the treatment and comparison groups share the similarities listed above, we can confidently infer that any differences we see in outcomes between the two groups can be attributed to the intervention. If, on the other hand, the comparison group differs from the treatment group in significant ways, comparisons of outcomes between the treated and comparison groups will reflect not only the impact of the intervention, but also the consequences of these differences. This is called selection bias.

Selection bias usually occurs when intervention participants and non-participants differ in characteristics that are not observed, which affect both the individuals’ probability of taking part in (and/or finishing) the intervention and the outcomes of interest.

In most youth employment programmes, it is likely that those who apply to participate differ in significant ways from those who do not apply, and that these differences cannot be easily observed. For example, participants of a job counselling project might be more motivated and have access to better information about how to find a job than non-participants, even before the intervention starts. In that event, participants might be more successful in terms of labour market outcomes but it would be unclear whether this was due to the intervention or because of their initial advantage in starting conditions.

One of the key objectives of the evaluation techniques presented here is to eliminate selection bias. In the absence of selection bias,

**DEFINITION**

A comparison group is a group used to estimate the counterfactual in an impact evaluation. In contrast to members of the treatment group, members of the comparison group have not been exposed to the intervention we want to evaluate. The terms “comparison group” and “control group” are often used interchangeably. For the purposes of this document, we will use the generic term comparison group throughout.

**DEFINITION**

Treatment group: The group of people that actively take part in an intervention is known as the treatment group or participant group.
observed differences in outcomes between treatment and comparison groups can be attributed to the intervention.

Just as counterfactuals are not observable, nor is selection bias. Unless care is taken in selecting treatment and comparison groups appropriately, a simple comparison of labour market outcomes of treated and comparison groups will include both the impact of the intervention and selection bias. It is, in general, impossible to know the exact extent to which this is due to one or the other.

The evaluation techniques presented below attempt to select treatment and comparison groups to eliminate selection bias, so that comparisons between treated and comparison groups reflect only the impact of the intervention.

A good comparison group is essential for the internal validity of the evaluation which determines the reliability and credibility of the evaluation results. External validity comes into play when we start to think about the transferability of these results: Should the intervention be scaled-up to other communities, or implemented at a regional or nationwide level? Can we expect similar results if we design this programme in other contexts and/or for a different target population? These are usually questions of significant interest to policy-makers.

It is important to bear in mind that the conditions will never be exactly the same when replicating or upscaled an intervention. Hence, to achieve external validity, it is crucial to understand the complex aspects surrounding the programme in the specific time, place and context of its implementation, and their potential influence on the evaluation results. For example, employment services for young graduates might be effective in a region because there is corresponding demand from the

---

**DEFINITION**

**Observable and unobservable characteristics:** Observable characteristics can be measured through appropriate data collection methods (such as surveys). They often include age, gender, level of education, socio-economic status, family characteristics, employment status, etc. Unobservable characteristics are those factors that cannot be, or are not, measured in an (impact) evaluation and could include motivation, interest, values and ideologies and the level of family support. For many of those unobservable characteristics (imperfect) proxy measures have been developed.

**DEFINITION**

**Selection bias:** Selection bias occurs when the reasons for an individual’s participation in a programme are correlated with outcomes. This bias often occurs when the comparison group self-selects out of the programme (for example, drop-outs).

**DEFINITION**

**Internal validity:** To have internal validity, an impact evaluation must have a comparison group that provides a valid estimate of the counterfactual. An internally valid impact evaluation will be able to clearly attribute changes in outcomes to the intervention by controlling all possible differences between the treatment and comparison group. This can be achieved through appropriately applying experimental or quasi-experimental techniques.
In order to understand not only if something works, but why and in what context it can be expected to work, it is necessary to analyse the causal mechanisms underlying the observed results. Qualitative methods, e.g. those applied in the context of theory-based evaluation, are of fundamental importance in this work.

In the course of this note, different quantitative impact evaluation methods will be introduced, followed by an example of a qualitative method and remarks on how a mixed-methods approach can help to achieve both internal and external validity of evaluation results.

**External validity:** In impact evaluation, external validity means that the causal impact observed can be generalized to all eligible individuals. Therefore, for an evaluation to be externally valid, it is necessary that the evaluation sample is a representative sample of all eligible individuals.

---

### Box 5.1: ILO’s support for impact evaluation

ILO’s Evaluation Department (also known as “EVAL”) has developed a variety of resources to support impact evaluation (IE):

- **An Impact Evaluation Framework:** EVAL developed a position paper about how, when and why IEs should be considered and implemented, based on input from ILO staff. The position paper covers key issues, such as the specific use and purpose of IE; the match between evaluation research questions and appropriate methodology; use of a range of complementary and available methodologies; the feasibility and value of IEs; and the need to not only identify impact (what) but also the how and why.

- **An Impact Evaluation Review Facility (IERF):** EVAL established a review mechanism, which allows ILO staff to ask questions and request reviews of concept papers, full proposals, plans and reports to assist with planning, designing or implementing IEs (EVAL_impact@ilo.org). A Briefing Note on the operation of this facility is available.

- **An inventory of impact evaluations conducted at the ILO:** The inventory allows easier access to institutional knowledge in a variety of intervention areas.

- **A quality appraisal of ILO impact evaluations:** In order to monitor and report on the progress that the ILO is making in its use of IE and the quality of IE, EVAL will periodically commission a quality appraisal of IEs across the organization.

- **An Informal Impact Evaluation Network as a community of practice:** This informal group of colleagues who are involved with and interested in IEs meets on a regular basis to share experiences and provide peer review of IE, as required.

These resources are intended to support the ILO in further enhancing its capacity in terms of the use of IE, in documenting knowledge of what works and for whom, and in assessing impact.
Challenges specific to evaluating youth-focused active labour market programmes

The nature of active labour market programmes (ALMPs), specifically those which focus on targeting a youth population, affects many aspects of the design of a valid evaluation. As background for a more detailed discussion of design issues in Note 6, this section presents an overview of some of the most common features of youth focused ALMPs and describes certain evaluation design features which are particularly relevant for these types of programmes. Understanding which of these features are likely to be present in a given setting will help in formulating the appropriate evaluation design.

Mandatory or voluntary programmes

A fundamental characteristic of a youth employment intervention is whether the programme is mandatory or voluntary. Mandatory programmes are built into many public employment services, including those related to unemployment insurances and training programmes. In these settings, young people are required to participate in an ALMP which is linked to an unemployment benefit. That being said, and as will be explained in the next sections, mandatory participation in a youth employment intervention does create challenges for impact evaluation, where valid estimates of impact typically need a treatment and an equivalent comparison group. Some impact evaluation methods can only be applied to voluntary programmes that recruit participants from a wider pool of applicants who can decide whether or not to participate.

Non-compliance: No-shows and dropouts

In many voluntary youth employment interventions a substantial fraction of people who are assigned to the programme will either fail to register for the programme (so-called no-shows) or will drop out prior to completion of the programme (dropouts). This challenge is particularly relevant to young people who are highly mobile, tend to change address and place of work frequently and alternate between working and studying.

Indeed, Card et al. (2011) state that:

it is rare to achieve programme completion rates over 80 percent and rates as low as 50 percent are common and failure to anticipate the problems caused by no-shows and dropouts is one of the leading causes of a broken design in ALMP evaluations (2011, p. 13).

While non-compliance by members of either the programme group or the comparison group does not invalidate an evaluation design per se, it does complicate the interpretation of the results, and means that the evaluation has to collect data on the actual programme participation rates of the treatment group and the comparison group.

The validity of a randomized design relies critically on the equivalence between the observed outcomes of the comparison group and the counterfactual outcomes of the treatment

3 This section is based on Card et al. (2011).
Recruitment and screening

Because only some of all the young people recruited into an impact evaluation are assigned to actually receive the programme, intake for an evaluation may disrupt the normal flow of clients into an ongoing programme. This is not a particular concern in a setting where there are many more applicants than available slots: in these cases random selection serves as a convenient and objective rationing device. In settings where the regular flow of recruits is needed to fill the available programme slots, however, programme operators may object to having some of their potential clients allocated to the comparison group and may try to override the assignment process. It is extremely important to know in advance whether this is likely to occur. If so, planning for the evaluation may have to include a budget for extra recruitment efforts to increase the flow of new clients, and extra resources to closely monitor compliance with recruiting protocols. For example, ALMPs for young people may be limited to unemployed men and women between the ages of 16 and 30. Normally, the same eligibility screening procedures and rules should be used to select participants for the evaluation.

Sample sizes

Guidelines for the necessary sample sizes for an ALMP evaluation are based on a standard power calculation. The main ingredient for this calculation is an estimate of the plausible effect size of the programme (e.g., the effect of the programme on the outcome of interest, expressed as a fraction of the standard deviation of this outcome). Given this value, and standard choices for the statistical significance level (e.g., 5 per cent) and the adequacy of the power of the design (e.g., 0.80), it is straightforward to calculate the appropriate sample sizes for the treatment and comparison groups of a randomized design with equal-sized groups. Card et al. (2011) developed guidance (shown in table 5.1) showing detailing the sample size required to measure a range of impacts. Each row shows the employment rate of the comparison group, and each column represents the difference between treatment and control groups. For example, if the employment rate is 50% per cent in the comparison group, to detect a significant impact of 2.5 percentage points in employment, the required sample size is of 6,354 participants and the same number of non-participants.

---

For example, if 100 new clients present themselves at the programme sites each month, and there are 80 open programme slots each month, then, at most, 40 people per month can be recruited into the evaluation: 20 will be assigned to the programme (along with the other 60 new clients who are not part of the evaluation) and 20 to the comparison group.
In thinking about the effect size of interest for an ALMP, Card et al. (2011) recommend to place that these programmes be put in context. They state that

A very large body of research has shown that in most countries around the world each additional year of formal schooling is associated with a gain in earnings of about 10 percent. Arguably, a typical ALMP involves a smaller investment than a typical year of formal schooling, so an effect size of less than 10 per cent is reasonable, and for less intensive programs, effect sizes of no more than 5 per cent may be plausible (2011, p.19).

**Timing of follow-up surveys**

The timing for the follow-up survey (or surveys) is an important decision in terms of guaranteeing programme impacts. Many ALMP evaluations use a one-year follow-up survey, in part because the terms of the evaluation contract often require a final report within two or three years. On the other hand, the existing ALMP literature suggests that the impact of more intensive programmes, such as classroom training and on-the-job training programmes, only tends to manifest itself two or three years after entry into the ALMP, rather than after just one year (Card et al., 2011). Based on these studies, and consideration of the interruption effects of many ALMPs, a post-programme horizon of at least two years is desirable for ALMPs of longer duration.

There is, however, a trade-off between being able to observe long(er)-term impacts and ensuring a valid impact evaluation design: as young people are highly mobile and might move across the country (or even migrate) after finishing their education or training programmes, it becomes increasingly difficult to track down a sufficient number of programme beneficiaries as time progresses. When a large number of the young people who participated in the baseline survey of an impact evaluation cannot be contacted for follow-up survey(s) (high attrition rate), it becomes increasingly difficult, or impossible, to reliably detect and quantify impacts due to reduced statistical power and possible bias.
Recommended quantitative methods of impact evaluation achieve internal validity and avoid selection bias by comparing groups with and without treatment, which ideally differ only in this respect. This can best be achieved if we have control over who receives the intervention and who does not. In this case, experimental evaluation designs are possible, the most common of which is a randomized controlled trial. If the assignment to treatment and comparison group is totally random, the two groups will be, on average, very similar before the programme starts and we will have gone a long way towards assuring internal validity.

For a variety of reasons, to be discussed below, randomization is not always possible or desirable. In that case, other methods can be used that seek to undertake internally valid comparisons by constructing a valid counterfactual. These are called quasi-experimental impact evaluation methods. The most commonly used ones are difference-in-differences (DID), propensity score matching (PSM) and regression discontinuity design (RDD), all of which will be briefly introduced in this section. It is generally an excellent idea to consider all possible impact evaluation efforts and to carefully weigh advantages and disadvantages before proceeding with the evaluation. Table 5.2 provides an overview over different quantitative impact evaluation methods.

**Quantitative methods of impact evaluation**

**DEFINITION**

**Experimental design:** Experimental designs rely on some element of randomization in the allocation of participants into treatment and comparison groups. They can produce highly credible impact estimates but are often costly and, for certain interventions, difficult to implement.

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.

**Quasi-experimental design:** Quasi-experimental design approaches are used to construct a valid comparison group by using statistical means to control for differences between the individuals treated with the programme being evaluated and those not treated.
### Table 5.2: Overview of different impact evaluation methods

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Description</th>
<th>Who is in the comparison group?</th>
<th>Required assumptions</th>
<th>Required data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before-and-after</td>
<td>Measures how programme participants improved (or changed) over time</td>
<td>Programme participants themselves before participating in the programme</td>
<td>The programme was the only factor influencing any changes in the measured outcome over time</td>
<td>Before-and-after data for programme participants</td>
</tr>
<tr>
<td>Comparing participants to non-participants</td>
<td>Measures difference between programme participants and non-participants after the programme is completed</td>
<td>Individuals who didn’t participate in the programme (for any reason), but for whom data were collected after the programme ended</td>
<td>Non-participants are identical to participants except for programme participation</td>
<td>After-programme data for participants and non-participants</td>
</tr>
<tr>
<td>Differences-in-differences</td>
<td>Measures improvement (change) over time of programme participants relative to the improvement (change) of non-participants</td>
<td>Individuals who didn’t participate in the programme (for any reason), but for whom data were collected both before and after the programme</td>
<td>If the programme didn’t exist, the two groups would have had identical trajectories over this period (would share the same “common” time trend)</td>
<td>Before-and-after data for both participants and non-participants</td>
</tr>
<tr>
<td>Propensity score matching</td>
<td>Individuals in the treatment group are matched with non-participants who have similar observable characteristics. The average difference in outcomes between matched individuals is the estimated impact</td>
<td>Non-participants who have a combination of characteristics which predict that they would be as likely to participate as participants</td>
<td>The factors that were excluded (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants</td>
<td>Outcomes as well as “variables for matching” for both participants and non-participants</td>
</tr>
<tr>
<td>Regression discontinuity design</td>
<td>Individuals are ranked based on specific, measurable criteria. There is a cut-off point to determine who is eligible to participate. Impact is measured by comparing outcomes of participants and non-participants close to the cut-off line</td>
<td>Individuals who are close to the cut-off line, but who fall on the side of that line where they (just) do not get the programme</td>
<td>After controlling for the criteria (and other measures of choice), the remaining differences between individuals directly below and directly above the cut-off score are not statistically significant and will not bias the results. A necessary requirement for this to hold is that the cut-off criteria are strictly adhered to</td>
<td>Outcomes as well as data of ranking criteria (e.g. age, index, etc.). Socio-economic background variables highly desirable.</td>
</tr>
<tr>
<td>Randomized evaluation</td>
<td>A sample of eligible individuals is randomly assigned into two groups; those who receive the intervention and those who do not. Impact is the difference in outcomes between the two groups. There are different ways of carrying out the randomization</td>
<td>Participants are randomly assigned to the treatment and comparison groups</td>
<td>Randomization is successful and complied with; that is, the two groups are statistically identical (in terms of both observed and unobserved factors)</td>
<td>Outcome data for comparison and treatment groups. Baseline data and background variables are desirable</td>
</tr>
</tbody>
</table>
A lottery is a simple and transparent way to assign youth to groups which will receive our services (the treatment group) and those which will not (the comparison group). This is the method used to design randomized controlled trials. If a large enough sample of people from the same population of interest is randomly assigned to one of two groups, then both groups will, on average, have similar observable characteristics (age, gender, height, level of education, etc.). Equally importantly, they will also, on average, share the same unobservable characteristics (such as motivation and state of mind).

Through randomization, the difference in outcomes that we observe between the two groups at the end of our intervention can be attributed to the intervention, because all other factors that could influence the outcomes are, in general, equal.

HOW IT WORKS

There are three steps to a lottery design (see figure 5.3).

**FIGURE 5.3: STEPS IN A LOTTERY DESIGN**

- **STEP 1**: Define eligible population
- **STEP 2**: Select sample
- **STEP 3**: Random assignment

Total population

Ineligible

Define eligible population

Select sample

Random assignment

Treatment

Comparison
Step 1: Define the eligible population

The first step in a randomized controlled trial is to find a group of eligible young people for an intervention. 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 enrol adults or elderly people in the intervention. Likewise, a job training programme 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 intervention and who will not.

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 purposes of our evaluation; this is called our sample. These will be the young people on whom we will collect data. While Note 6 provides more details about how to determine the sample and its size, the typical sample size for a youth employment intervention evaluated through a lottery design is somewhere between 500 and 2,000 study participants (usually with a roughly equal split between the treatment and comparison groups).

Choosing the sample for the evaluation can be done in two ways, depending on whether the intervention is large or small. A small intervention may find that there are 10,000 eligible beneficiaries, such as urban street youth aged 16–24 years old. The intervention may have sufficient budget to help 500 of them. Ideally, a comparison group will be similar in size to the treatment group, so 1,000 out of the 10,000 street youth will need to be selected for the intervention and evaluation (see figure 5.4, left-hand image).

Large programmes may be bigger than the sample size needed for an evaluation. If the job training is able to serve 4,000 young people, it is not necessary to find an additional 4,000 young people for comparison. Instead, only 1,000 may be needed. The intervention 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 intervention. The remaining 2,000 will then be randomly split between the intervention and the comparison group (figure 5.4, right-hand image).

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, participants will, on average, have similar characteristics to the total eligible population. Even though we include only a limited number of youth in the study, the potential impact of the intervention can be extrapolated to cover the entire eligible population, in this case, 10,000 young people.

**DEFINITION**

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 extrapolate results to the entire population.
Step 3: Randomize assignment

The next step is to assign the selected sample of youth to treatment and comparison groups which are roughly equal in size. In randomized controlled trials, every youth has the same chance of receiving the intervention. Randomization can be done 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 Microsoft Excel. Randomization

TIP

One way of obtaining 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 approach is not possible, randomly targeting areas where street youth interact, such as an urban centre, will produce a random sample. If young people are known to spend time at 50 different centres around a city or country, randomly selecting centres and then selecting a portion of youth at these centres to participate in the study is likely to result in a selection of youth with minimal bias. Note 6 will discuss sampling more in detail.
can occur at several levels. 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.

WHEN CAN I USE A LOTTERY DESIGN?

A randomized lottery evaluation can be used when the evaluation is planned in advance of implementation (prospective) and when the intervention can serve only a fraction of the eligible youth. As long as resource constraints prevent the intervention 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 intervention. In such a situation, comparison groups can be maintained to measure short-, medium- and long-term impacts of the intervention (Gertler et al., 2016). Importantly, the central advantage of randomizations – that treatment and comparison groups, on average, share the same characteristics – will only be maintained if we manage to follow up with (almost) all members of the treatment and comparison group. High attrition rates pose a severe threat to the internal validity of our results for every impact evaluation method, and methods that use randomization techniques are no exception.

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 simplest of all evaluation methods in analytical terms. The impact of the intervention 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.
- It is easy to implement and communicate to programme staff.

DISADVANTAGES

- Conducting a randomized experiment can be very cost- and time-intensive.
- No ex-post implementation of this method is possible. Planning the evaluation has to be part of planning the intervention (which is good practice in any case but does not always represent the reality in project work).
- It requires a comparison group to be excluded from the intervention for the duration of the impact evaluation. Political and/or ethical concerns might emerge in spite of the transparent allocation criterion of randomization (see more in the section “Adapting random designs to different contexts” below).
- 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 achieve. This condition may be threatened if randomization is implemented incorrectly, if treatment
or comparison groups do not comply with their status (that is, if treatment individuals do not take up the intervention or comparison individuals receive the programme), if participants drop out of the study prior to completion or if there are spillover effects: for example, young people who received the job training might transfer the acquired skills and knowledge to their peers, thereby blurring the clear separation between treatment and comparison groups. These cases are highly problematic, as they can substantially bias the results and thereby threaten the overall validity of the evaluation.

DEFINITION

**Spillover effects**: Spillovers are effects of an intervention on non-participants; for instance, if knowledge from a skills training spreads within a village, even to those who did not attend the course.

---

**Box 5.2: Evaluation of ILO’s intervention Start and Improve Your Business (SIYB)**

The evaluation was designed to test whether expanding access to capital via grants or loans would increase the profits of micro-enterprises owned by men or women, and whether the ILO’s SIYB entrepreneurship training could further increase impacts.

The research team surveyed 4,637 micro-enterprises from a census of businesses and selected 1,550 business owners to be included in the evaluation sample – based, among other criteria, on an expression of interest in receiving ILO training and participating in the loan intervention. The sample included small business owners interested in improving their businesses (for example, hair salons, retail shops and tailors). The sample was randomly split into five treatment arms, which received the following interventions: (1) a loan; (2) a cash grant; (3) business training and a loan; (4) business training and a cash grant; and (5) no intervention (the comparison group) (figure 5.5).

---

**FIGURE 5.5: EVALUATION DESIGN**

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Number of Businesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cash grant</td>
<td>167</td>
</tr>
<tr>
<td>Loans</td>
<td>406</td>
</tr>
<tr>
<td>Business training and loans</td>
<td>401</td>
</tr>
<tr>
<td>Business training and cash grants</td>
<td>219</td>
</tr>
<tr>
<td>Comparison group</td>
<td>357</td>
</tr>
</tbody>
</table>

The full sample of businesses included 1,550 businesses.
The main interventions which were delivered to the business owners were:

- **Business training:** The Start Your Business (SYB) training programme targets starting (or nascent) entrepreneurs and consists of a five-day training course, followed by fieldwork and group-based and individual counselling sessions. The trainees prepare their detailed bankable business plan and action plan (see [www.ilo.org/siyb](http://www.ilo.org/siyb)).

- **Unconditional cash grants,** valued at US$200, were delivered via free bank accounts at a local microfinance institution (MFI). The business owners were given free choice in the use of the loan.

- **Semi-conditional loans,** valued at US$180 to US$220, were offered at a discounted annual interest rate of 20 per cent by the MFI. Loans had to be paid back to the MFI, but there were no consequences in the event of misuse of the money.

The size of the grants and loans is equal to approximately 1.5 times the monthly profits of the average businesses.

To check whether the randomization “worked”, the evaluation compared business owners in the treatment group with those in the comparison group with respect to 26 different variables and found that, for virtually every characteristic, treated and non-treated enterprises, on average, looked alike before the intervention.

The business owners were surveyed before the intervention (baseline survey) and six months, nine months and two years after the intervention (three follow-up surveys).

The main outcome variable of interest was business profit and the evaluation found a significant increase in earnings after the intervention, but only for male business owners. None of the interventions led to sustained increases in profits for female entrepreneurs. Women with high initial profits also saw negative effects through all interventions. While the initial response to the grants was positive, this increase disappeared entirely and even became negative over time. Women who received the grant made 35 per cent less profit than their peers who received no intervention. After nine months, women were either not better off or were even worse off than their counterparts in the comparison group.

The evaluation also found that the proximity of family members represents a positive force on business for men and a negative one for women. Married women with family living in the same district experienced large and significant decreases in their profits.

Source: Fiala, 2015.
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) from needy individuals, which would be unethical. Still, for many interventions, demand considerably exceeds what can be supplied and, as further elaborated in box 5.3, randomization may in fact be more ethical than other selection methods.

Adapting random designs to different contexts

Box 5.3: Is randomization ethical?

Sometimes randomly assigning potential beneficiaries into treatment and comparison groups is considered unethical. These concerns might be valid in certain cases, for example when a policy or intervention that is likely or proven to work can be extended at little cost to a large population. However, more often than not one of the following situations arise:

- **Uncertainty of project impact.** For most programmes, it is not clear if the intervention has a positive and sizable impact on the individual and the community that justifies the resources being spent. For instance, programmes geared toward girls at the exclusion of boys may increase gender violence. A microfinance intervention for youth may leave participants worse off if they are not able to repay their loans. A poorly designed training programme may actually decrease job prospects. An increase in incomes (e.g., a US$100 per participant) may come at a very high cost (e.g., US$1,000 per person). 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 intervention based on randomly assigned treatment and comparison groups.

- **Budget constraints.** In reality, because of limited resources, it is rarely possible to serve everyone in need. That is, most programmes 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 intervention has a limited number of available spots, then some young people 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 programme managers to allocate scarce places in their interventions 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.
Nevertheless, creating a pure comparison group by random lottery assignment in which young people are never given the intervention is sometimes impossible. Both random phase-in and random promotion designs avoid the strict separation into treatment and comparison groups and might be a viable alternative for an experimental impact evaluation design when lottery designs are not feasible or desirable.

RANDOMIZED PHASE-IN DESIGN

Because many programmes are active in a community for years, never giving the intervention to a group of needy individuals can be both politically and programmatically difficult. A variation of the lottery design is the phase-in design. The main difference between a phase-in design and a lottery design is the method of assigning people to treatment and comparison groups. In practice, potential beneficiaries are randomly divided into two or more groups. The intervention is then rolled-out over time, so that individuals of group one participate in the intervention first, followed by group two, group three, and so on. During the time when groups are on the waiting list, they can serve as the comparison group until they receive the intervention.

For example, a non-governmental organization (NGO) may have sufficient budget to train 1,500 youths, but it may not have the capacity to conduct all of the training simultaneously. Instead, it chooses to train 500 people per year over three years. If it can identify all 1,500 participants in the beginning, a phased-in randomization may be the best evaluation method to adopt. The 1,500 youths are randomly split into three groups. In year one, while group 1 receives training, groups 2 and 3 remain on the waiting list and can serve as the comparison group. In year two, only group 3 remains for comparison purposes. 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. This method often suits the natural roll-out of many programmes.

However, because everyone eventually benefits from the programme, the phase-in design approach is usually not ideal for finding the long-term impact of an intervention because eventually there is no comparison group. Even large, longstanding programmes will have difficulty in asking participants to wait for three or four years before their turn comes, so the time span of results is often limited to one or two years. Moreover, there is a risk that participants may change their behaviour while waiting to join the intervention. This could invalidate their ability to serve as a good comparison group. For example, they may stop looking for jobs in anticipation of joining a skills training intervention.

TIP

With a phase-in approach, it is critical to have enough time between each of the phases for the intervention to show effects. If, for example, an intervention officer believes that it will take two years for the impact of the intervention to take effect, the time between the first and last phase must be at least two years. Small or short-term programmes may not be suitable for this approach.
RANDOMIZED PROMOTION DESIGN

There may be cases where it is not possible or desirable to exclude any potential beneficiaries and where the intervention is not rolled out over time. In such cases, the randomized promotion method (also called encouragement design) may be suitable. When it is not possible to randomly assign young people into a group that receives benefits and a group that does not, it may be possible instead to randomly promote the intervention. That is, rather than randomizing those who receive the benefits and services, we randomize who is encouraged to receive those benefits.

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 programme, we could hire a social worker to randomly visit homes of unemployed young people, describe the programme and offer to enrol young people on the spot. In the case of a financial literacy campaign, we may want to randomly send text messages to one part of the target audience, but not to another. In all cases, there will still be people in the promoted group that will not take up our intervention, 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 enrolment rate among the promoted group should be higher than the rate among those who did not receive the promotion.

To assess the impact of the intervention, we cannot, unfortunately, simply compare the outcomes of those who participated in the intervention with the outcomes of those who did not. People who choose to participate in an intervention 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 intervention will not be random, so comparing participants to non-participants would be like comparing apples to oranges.

We can, however, compare outcomes between everyone who received the encouragement and all young people in the comparison group. 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 can be attributed to the fact that those people only enrolled in the intervention because they received the promotion.

A key advantage of this design is that randomized promotion campaigns never deny anyone the programme, but instead allow people to make their own decisions about whether or not to take up the intervention. However, these studies often need larger sample sizes to provide reliable impact estimates, which increases costs.
El Mashroua is a reality TV show designed to promote entrepreneurship among young adult viewers and broadcasted on one of the most popular Egyptian television channels. To evaluate the impact of the show, a research team, supported by the ILO, chose a randomized promotion design. From the study sample of 9,277 individuals, a randomly selected treatment group received SMS reminders about the show that were designed to encourage recipients to tune in.

The follow-up survey that was conducted approximately 1.5 years after the broadcast clearly showed that young people from the treatment group (those who received messages) were more likely to have heard of the show and to have watched at least one episode compared to youth from the comparison group (those who did not receive reminders), see also figure 5.6.

These statistically significant differences can be exploited to estimate the actual impacts of the show on the viewers’ attitudes and labour market outcomes. This is possible because two assumptions can be made:

1. Because of the randomization, treatment and comparison group do not differ systematically in any observable or unobservable characteristics that could be correlated with the outcome variables.

2. Because having received the messages alone does neither affect attitudes nor labour market outcomes, any difference between treatment and comparison groups can be attributed to the difference in likelihood of having watched the show.

The study finds that having watched the show did not have impacts on young people’s propensity to start a business, but that it significantly reduced gender-discriminatory attitudes held by men against women.

Source: Barsoum et al., 2017.
For the reasons already explained, it is sometimes not possible or desirable to employ experimental evaluation methods. In that case, there is a range of quasi-experimental impact evaluation methods, which can also deliver robust, internally valid results. One of the most commonly used techniques is the difference-in-differences (DID) approach, which compares the change in outcomes experienced by the treatment group with the change in outcomes experienced by the comparison group.

HOW IT WORKS

**Identifying the comparison group:** DID designs rely on having a comparison group whose development in our key outcomes of interest we can reasonably assume would be the same as the development of the treatment group over the time period of the intervention. To this end, it is desirable to choose groups with similar characteristics.

Let us imagine a six-month job training intervention for young people, for which we want to evaluate impacts on labour market outcomes. Randomly distributing training places is not possible. Instead, we take a sample of young people of similar age, education level, socio-economic background and labour market situation from another community as the comparison group.

**Estimating the impact:** To apply the DID evaluation technique, we need to (a) measure our outcomes of interest (for example, labour market status, see Note 2) for both the treatment and the comparison groups before the job-training intervention begins and (b) measure the outcomes of both groups at a given time after the intervention took place. Even though we tried to identify a comparison group of young people that appear similar to the treated youth, it is likely that there are differences between the two groups prior to the job training, and that these differences remain afterwards. Figure 5.7 shows a situation where the comparison group has a considerably lower outcome indicator (say, employment status) at baseline. However, this does not affect the method. The DID technique compares the difference in outcomes between both groups at the end of the intervention (point B minus D) and adjusts it for the difference in outcomes between both groups at the beginning (A minus C). Subtracting these differences from each other (that is, taking a difference from two differences, which gives the method its name) yields an idea of the programme’s impact; it shows whether and to what extent the training intervention increased employment status for participants relative to those who did not participate. The scenario in figure 5.7 indicates a moderate positive impact of the job-training intervention.
NOTE 5. IMPACT EVALUATION METHODS FOR YOUTH EMPLOYMENT INTERVENTIONS

DIFFERENCE-IN-DIFFERENCES (DID)

The “common trend” assumption: The assumption underlying this method is that, although the observable and unobservable characteristics of the treatment and comparison groups may be somewhat different (reflected in different levels of income at the beginning of the intervention), 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. We therefore do not have to assume that without the intervention outcomes would have remained constant but rather that the treatment and the comparison groups share the same trend over time. This is what we refer to as the “common trend” assumption.

Coming back to the job training example above, in order to be able to apply DID, we have to be sure that over the next six months there will be no factors that systematically influence the outcomes of youths from the “treated community” differently to the outcomes of those from the “comparison community”, apart from the training assignment. For example, faster economic growth, a new local policy providing incentives to companies for the employment of young people or a major employer closing down in only one of the two communities would violate that assumption and consequently bias our evaluation results.

A good test to establish whether it is realistic to assume equal trends between participants.

Source: Adapted from Gertler et al., 2016.

FIGURE 5.7: EXAMPLE OF DIFFERENCE-IN-DIFFERENCES ANALYSIS

Impact = (2. Difference) - (1. Difference)
and non-participants is to compare their changes in outcomes before the intervention is implemented. This approach requires multiple data points prior to the intervention. As several baseline surveys can quickly become very costly, this test can more easily be carried out if administrative data on our key outcome indicators are available at little cost (for example, employment status from public employment agencies or test scores from previous school years). If the outcomes of the two communities moved in tandem before the intervention started, we can be more confident that their outcomes would continue this trend during the intervention. If, however, pre-intervention trends are different, the equal trend assumption may not be correct.

WHEN CAN I USE A DID DESIGN?

Since it assumes that the differences between participants and non-participants are constant over time, this method is most usefully applied when there are good data available at multiple points before the intervention begins. To improve the credibility of impact estimates it is preferable to have at least three rounds of data collection: two prior to treatment, and at least one at the end of the intervention (see above). This means that, unless data on participants and non-participants are available through other channels, such as an existing household survey, the costs of such an evaluation can be much higher than those of other impact evaluation techniques.

ADVANTAGES

- This method provides a way to account for both observable and unobservable differences between participants and non-participants. More precisely, it controls for all individual effects that remain constant over time, or that share the same course of change over time (i.e. treated and comparison groups show similar trends in the outcomes of interest).
- Even if the method is not experimental, it allows for a (partial) check of the assumption that renders it internally valid. This implies that we can have a sense of whether our estimated impacts are valid or not. If good administrative data is available, the method can be applied fairly easily and even ex-post, based on before and after data from the programme.

DISADVANTAGES

- This method produces less reliable results than randomized selection methods.
- In order to test the key assumption of “common trends”, at least three data collections are required, so the implementation can be expensive if data are not available initially.
The Plan Jefes programme is a conditional cash transfer programme introduced during the Argentinian economic crisis of 2001–2002. Reforms of this programme following the recovery after the crisis included the implementation of the Training and Employment Insurance (Seguro de Capacitacion y Empleo, SCE) in 2006, in order to provide support in skills upgrading, vocational training, jobseeking and job placement to the eligible participants of Plan Jefes.

Participants in the SCE receive a monthly stipend and are provided with the following activation measures:
- assistance for the completion of primary and secondary education
- vocational training and apprenticeships
- labour intermediation services
- indirect job creation measures (e.g. employment subsidies)
- promotion of self-employment and micro-enterprise creation.

The ILO studied the effect of implementing these active labour market tools for beneficiaries of the Plan Jefes programme on their labour market status and job quality with DID estimators. In order to isolate the effect of these tools, a comparison group with similar features to those of the SCE participants had to be identified.

As the transfer from Plan Jefes to the new programme was gradual, the researchers could select participants in Plan Jefes who met the requirements to be beneficiaries of the SCE but had not yet been transferred to the new programme. An important key assumption of this identification strategy was that the transition between the programmes was not influenced by any factors which might be driving differences in the outcomes of interest. A total of 1,149 non-participants were selected, based on data from Argentina's Permanent Household Survey – a survey conducted quarterly by the Argentinian National Institute of Statistics (INDEC) which contains questions about individuals' personal characteristics, education and labour market performance. The selected participants and non-participants were similar in gender, age and level of educational attainment.

The evaluators compared a range of outcomes between the two groups at two different moments in time (baseline and follow-up). This approach allowed causal effects of the SCE programme to be identified while controlling for selection bias due to observable and unobservable characteristics of the participants.

The panel structure of the survey allowed the researchers to gather data on both participants and non-participants at several points in time, both before and after programme participation. This allowed the assumption that, in the absence of the programme, the outcomes of participants and non-participants would have changed in the same way (common-trend assumption) to be tested and confirmed.

The study results showed that the programme had a positive effect on the participants’ job quality, i.e. the probability of having a formal job and higher hourly wages, and a lower probability of having a low-paid job and working an excessive number of hours. It did not affect their employment status (i.e. the probability of being employed). The evaluation also showed heterogeneous effects, revealing that the programme had a higher impact for young beneficiaries, but no effect for women.

---

5 In order to correct for observed differences, the researchers also applied the propensity score matching (PSM) method. See the following section for a more detailed description of this methodology.
Propensity score matching (PSM) is a very commonly used approach among the quasi-experimental evaluation methods. Its basic principle is to construct a comparison group by matching participants with similar non-participants, based on their predicted probability of participating in the intervention. This is called the propensity score, which is calculated based on a range of observed characteristics.

HOW IT WORKS

A range of potentially relevant covariates have to be selected in order to calculate the propensity score for non-participants, based on their probability of being treated. The aim is to include in the propensity score calculation all covariates that affect both programme participation and outcomes. Non-participants are then matched with participants based by their respective scores. There are different ways of matching procedures, the most common approach being nearest neighbour matching, where each participant is matched to the non-participant with the closest propensity score. The closer the score, the better the matching quality. Balancing tests can be conducted to assess how well the matching worked. Consequently, the average difference of the two groups in the relevant outcomes of interest is equivalent to the impact of the intervention.

As an example, consider a skills training programme targeting rural youth which has 1,000 participants. Pre-programme data on key characteristics of the participants are available, e.g. sex, age, education and key aspects of their labour market history. Existing secondary survey data can be used to construct a comparison group based on their propensity score, estimating the probability of treatment for a large number of individuals based on the abovementioned characteristics. A total of 1,000 people with the best matching propensity scores will be selected as a comparison group for the intervention. Post-intervention data from the comparison could be gathered through the same secondary data source (if it is a panel data set, regular waves might be collected).
Box 5.6: Comparing participants and non-participants

Sometimes, although we might be able to identify a comparison group, we might only have available data on a few key outcome variables and no information on covariates, such as socio-economic background, knowledge, skills, etc. In these cases, we can use a simple impact evaluation methodology and compare outcomes of participants and non-participants. Thus, the counterfactual is estimated by the outcome of people who did not participate in the programme. However, this method is unlikely to yield either credible results or useful information about the true effect of our programme.

In particular, if non-participants (comparison group) differ from participants (treatment group) in ways that are relevant to the outcomes, this type of comparison will not be valid and will feature selection bias. More precisely, this method relies on two strong assumptions. First, we need to assume that programme participants and non-participants had, on average, similar outcomes at the beginning of the programme.

The right-hand side of figure 5.8 depicts a situation where participants already had a higher income at the beginning of the intervention than non-participants. This case leads to an overestimation of the true impact of our intervention.

Second, we must assume that, in the absence of the intervention, both groups would have developed similarly over time. This requires the assumption that, on average, participants would have reacted in the same way as non-participants to all external factors. Note that in the situation described on the right-hand side of figure 5.8 this assumption holds true. The black dotted line, which describes how non-participants developed over time, and the red dotted line, which describes how participants would have developed in the absence of the intervention, move parallel over time. In order to obtain accurate impact estimates through this method, both assumptions must hold.
WHEN CAN I USE PSM?

PSM is a particularly useful method when large and rich amounts of secondary data are available, as these are necessary to define a good propensity score and to match sufficient numbers of participants and non-participants with similar scores, i.e. to find a large enough region of common support. Furthermore, PSM relies on the assumption that only observed factors influence both participation and outcomes (conditional independence assumption). Thus, PSM should only be applied if there is a good understanding of the drivers of programme participation and the outcomes of interest, and should be avoided if unobservable characteristics can be expected to affect those variables. In any case, careful consideration is needed before the decision can be made on how many, and which specific variables to select for the estimation of the propensity score.

ADVANTAGES

PSM is a robust impact evaluation methodology which, if its assumptions are met, can help to remove selection bias and provide internally valid results. As in the case of the other quasi-experimental methods, it can be applied based on existing data sources and no random assignment of the intervention is necessary. By matching on the propensity to receive treatment, PSM reduces the number of dimensions on which to match participants and comparison units to one, and thereby makes matching relatively straightforward.

DISADVANTAGES

- The application of PSM usually requires large data sets.
- Matching can only be conducted on observable characteristics. Hence, the risk remains that selection bias due to unobservable characteristics driving programme participation can affect the evaluation results.
- The application of PSM is statistically complex and requires a corresponding level of expertise.
Note 5: Impact Evaluation Methods for Youth Employment Interventions

From 2013 to 2014 the Population Council implemented the Neqdar Nesharek (meaning “We can participate”) programme in rural Upper Egypt. The programme targeted 4,500 young women aged 16–29 years old, adopting the “safe spaces” livelihood approach by addressing community-specific needs of vulnerable women. The intervention aimed to empower young women by providing them with business and vocational skills training and supporting them in starting a business or seeking employment. The training programme consisted of three main components: (1) business skills training, (2) vocational training and (3) life skills, legal rights and civic education.

The intervention was accompanied by an impact evaluation to assess the effect of Neqdar Nesharek on young women’s labour market outcomes and social empowerment measures. The evaluation used a PSM design. Impacts were calculated by matching women who participated in the programme with women with similar socio-economic characteristics from villages in the comparison group and comparing key programme outcomes between them (see figure 5.9).
The evaluation found a significant impact of the programme on the economic empowerment of programme participants, as measured by their engagement in income-generating activities. Programme participants were 4.5 percentage points more likely to be engaged in an income-generating activity than women in the comparison group. As shown in figure 5.10, most of the positive impact was driven by an increase in participants’ engagement in self-employment activities. In contrast, the level of participation in wage work did not significantly change for women in the treated group.

Source: ILO, 2017
Regression discontinuity design (RDD)

Regression discontinuity designs (RDDs) are often used when eligibility for a labour market intervention is based on some form of continuous ranking of potential beneficiaries or applicants, for example a cut-off age.

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, 25-year-olds, who may be eligible for a youth skills training programme, are not likely to be very different from their 26-year-old peers, who may no longer be eligible. If we have a situation in which some of those youth who receive the programme (those just above the threshold) and some of those who do not (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 analyse programme impact.

Figure 5.11 illustrates what we might find when analysing the impact of a youth microcredit initiative. The left-hand graph indicates that, at the time of applying to the programme, those who achieved better scores already tended to have higher incomes. There may be many reasons for

**FIGURE 5.11: SAMPLE DISCONTINUITY CHART**
this, for example, that those with a slightly higher level of education are already earning more and that their education also helped them to secure better scores. Or those who are more motivated to start a business are already more entrepreneurial, reflected in their higher incomes, and that motivation also helped them to convince the jury to support them. Many other explanations are possible, which we do not necessarily need to understand to apply this method.

When starting the programme, the local microfinance bank decided that the threshold for receiving a loan was 85, and all applicants were accepted or denied support according to their ranking relative to that threshold. Now we would like to establish whether the microcredit programme had any impact on incomes. As illustrated in figure 5.11 (right-hand 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 programme, which will be represented by the difference in outcomes (that is, the discontinuity of the linear relationship) near the cut-off point.

WHEN CAN I USE RDD?

In many cases we are not able to plan the evaluation during the programme design. Sometimes, however, we may be able to use the targeting rules of the programme to obtain a good comparison group ex-post. Some programmes use a continuous ranking of potential beneficiaries, such as test scores, credit scores or poverty index, and have a cut-off point for acceptance into the programme. In the case of youth labour market interventions, there is often a binding age cut-off. Only youth under a specified age are eligible for the programme. This eligibility rule can be used for conducting an impact evaluation based on RDD.

ADVANTAGES

- The RDD can be applied ex post, if sufficient administrative data are available.
- It can take advantage of an existing rule for assignment to construct a valid comparison group and thereby does not require the exclusion of an eligible group from the intervention.

DISADVANTAGES

- The main requirement for using discontinuity designs is that programme participation is determined by an explicitly specified targeting rule; in other words, by a continuous scale or score. For this method to work, we need many observations in the region immediately above and below the cut-off point in order to have sufficient numbers of youth to compare with one another. Unless the evaluation is done without baseline data or can take advantage of existing programme records, a discontinuity design requires similar data collection to a lottery design, and therefore bears a similar cost.
- The informative value of the results is limited to the sample around the cut-off point. This might be relevant, for example, in discussions regarding whether a programme should be scaled-up to include other age groups or regions.
Public works programmes are an increasingly popular policy tool in developing countries. From 2007 to 2011 the Government of Peru implemented the programme Construyendo Perú with the primary objective of supporting unemployed individuals in situations of poverty. The programme provided them with access to temporary employment and skills development training through the financing of public investment projects with intensive use of unskilled labour.

The ILO evaluated the medium- to long-term effects of the programme using a regression discontinuity approach. The evaluation exploits an interesting assignment rule of the programme at the district level that consisted in selecting beneficiary districts by ranking them according to the FAD (Factor de Asignación Distrital) index. The FAD is a composite index that combines demographic information with an index of human development shortcomings and a poverty severity index. As such, districts with an FAD index above a certain threshold (i.e. whose with higher poverty and development shortcomings) were allowed to participate in the programme and districts below that threshold did not participate in the programme. This is an example of a fuzzy regression discontinuity design. As shown in figure 5.12, districts just above the cut-off point were considerably more likely to participate in the programme than those just below the threshold.

The reasoning behind the evaluation is therefore to estimate the causal impact of the programme by comparing outcomes of individuals around the cut-off point of the FAD index. The evaluation found that over the medium-term (three to five years) the intervention helped to increase employment and reduce inactivity for women and less-well educated programme participants. However, the programme was not able to improve the prospects of lower-educated participants in terms of job quality and, in fact, had a detrimental impact on job quality perspectives of women and more highly educated individuals (for example by increasing the probability of informal employment).

Source: Escudero, 2016
Simple comparisons: Before and after

Sometimes randomization is not possible and, moreover, the conditions for a valid quasi-experimental evaluation do not hold; for example, if we cannot find a suitable comparison group with baseline information available and/or if the common trend assumption cannot be confirmed. In these cases, it is advisable to consider whether it is worth conducting a quantitative impact evaluation at all.

If it is not possible to include a comparison group in an impact evaluation, the most basic approach relies on simply comparing the outcomes of programme participants before and after the intervention. This simple approach can give an idea of the change that occurs over the course of an intervention but should be regarded as part of a monitoring system rather than as providing evidence of the causal impact of an intervention since there is no way of knowing if an observed change should be attributed to the intervention in question or to other circumstances.

Taking the example of a training programme, we may observe that the monthly income of participants increased from $50 before the intervention to $60 after the intervention and therefore conclude that the impact of the intervention was $10 per month per person (see figure 5.13, left-hand graph). However, if in the absence of the intervention the income level could have increased anyway due to a change in circumstances (i.e. the situation we are facing corresponds to one of the scenarios shown in the right-hand graph of figure 5.13), we will not be able to obtain an accurate estimate of the intervention.

Since the real counterfactual scenarios (the dotted lines in the figure) cannot be observed, there is no way of knowing if the case that applies in a particular evaluation is the one shown on the left-hand side of figure 5.13 or the one on the right. It is therefore impossible to have a sense of whether the impact we estimate with this method is the true impact of our intervention or a “contaminated” one.
Performing before-and-after comparisons could make sense if there are reasons to believe that, in the absence of the treatment, outcomes would, on average, remain unchanged. This could apply to interventions that (a) are delivered over a short period of time (for example, short skills training interventions, job-counselling services or events that aim to change the attitudes of participants) and (b) are expected to have effects of interest in the short term. However, the above-mentioned limitations persist and, unlike well-implemented experimental and quasi-experimental methods, simple before-and-after comparisons cannot be considered robust impact evaluations. Their level of robustness can be improved, first, by controlling for potential confounding factors in a regression model (instead of simply comparing the outcomes) and/or, second, by the complementary application of qualitative methods in order to work out the causal mechanisms underlying the observed change in outcomes.

**FIGURE 5.13: COMPARING BEFORE-AND-AFTER OUTCOMES**

**ASSUMPTION**

Counterfactual is constant over time

Outcome measure (income)

Before | After

Measured change

Presumed counterfactual

$50 | $50

Impact = $10?

**REALITY**

Counterfactual may be dynamic over time

Outcome measure (income)

Before | After

Measured change

Real counterfactual

$50 | $55

Impact = $5?

$50 | $45

Impact = $15?
Improving the relevance of quantitative impact evaluations

As shown in the preceding section, there is a range of good quantitative methods to provide an internally valid answer to the basic evaluation question, “Did the project work?”; that is, “Did it affect the outcomes of interest as defined in our intervention and learning objectives?”. The question of whether the intervention as a whole had an impact is an important one, but it is by no means the only question we may want to ask.

In order to gain a detailed and holistic understanding of how and why the effects of a youth employment programme unfold, we need to “dig deeper”. Having a clear understanding of the heterogeneity of impacts and the causal mechanisms leading to observed effects also helps us to derive valuable lessons from our evaluation and gain a better insight into whether a programme is likely to work in other settings.

This brings us back to the important issue of internal and external validity (see the beginning of this note). In order to achieve internal validity, our methods need to be robust and properly implemented. In order to achieve external validity, we need to understand the relevant contextual factors of our programme and their potential effect on the evaluation results. It is difficult to reach achieve objectives using a single method.

For example, experimental evaluations, if properly implemented, can give us credible information about the impacts that can be uniquely attributed to our project, but can tell us very little about their replicability in other settings. Critically, quantitative impact evaluations tell us “what” happened – the average treatment effect – but they do not tell us “why”. For this purpose, the complementary application of qualitative methods is required.

MEASURING A VARIETY OF IMPACTS

First, it may be useful to have a more nuanced picture of the programme’s actual impact. This can be partly achieved within the quantitative designs described above. Relevant questions to ask could be:

- Do outcomes vary across different groups of beneficiaries (e.g. young men benefit, but young women do not)?
- What is the short-term versus the long-term impact of the intervention?
- Does the intervention have positive or negative spillover effects? Are there any intended or unintended outcomes beyond the actual target group?

Second, we may also be interested in testing cross-cutting designs, testing how the effectiveness of our intervention changes as we modify the design. These designs allow us to investigate the following questions:

- Is one intervention design more effective than another? We may want to compare alternative interventions (providing start-up grants versus start-up loans for young entrepreneurs, for example), or test the
most effective combination of programme components (training alone, training plus internship, or training plus internship and mentoring).

- What is the most effective dosage of the intervention? For example, should we provide 20, 50 or 100 hours of training (see table 5.3 for further impact evaluation questions)?

Cross-cutting designs help to identify more than just the overall impact of a project; they also evaluate specific intervention features and why these do or do not work. For example, a programme may provide vocational and entrepreneurial skills training, such as carpentry or tailoring, along with a small amount of start-up capital for businesses. The provision of cash grants could be expensive or politically difficult, and so the programme director may wonder whether the start-up capital is necessary, or if participants are able to implement their training without the capital. A cross-cutting design can help to determine the best project design in this case. In practice, this requires us to compare the outcomes of different treatment groups to a comparison group and to each other.

COMBINING QUANTITATIVE AND QUALITATIVE APPROACHES

Furthermore, we may be interested in shedding light on the channels through which the impact operates – that is, understanding why and how an impact unfolds. For example, we might want to answer the following questions:

- How and why did things happen as observed?
- Why did a project (or part of it) not work as we expected?
- What can we learn from failure?

If such potential avenues of investigation are envisaged at the design stage, theories can be tested partially within the quantitative methods above. To achieve this end, the surveys must include questions designed to capture the different factors (intermediate outcomes) through which the impact is hypothesized to operate in order to verify if the intervention affects these intermediate outcomes.

However, many outcomes of youth employment interventions (such as mental health, empowerment or household relations) are complex and multidimensional and may not be captured with quantitative methods. Mixed methods allow for tracking qualitative indicators and provide selected case study analysis to help develop a better understanding of the dynamics and results of the intervention. For example, structured and semi-structured qualitative interviews, in which participants are free to express real-life stories that fall outside categories of quantifiable information, can help to round out an understanding of a programme’s impact (Bamberger et al., pp. 6–7; Leeuw and Vaessen, 2009).

Qualitative data collection methods might be particularly useful for collecting information about how well the intervention was implemented (see Note 4 on performance evaluations). Understanding the implementation process is crucial to discovering how the intervention implementation affected results and correctly interpreting findings to determine whether disappointing results are due to weaknesses in intervention design or in implementation. Furthermore, qualitative techniques might shed light on why specific findings transcended and, in particular, why effects differed across the target population (for example, between rural and urban young people or between young women and young men).
### Table 5.3: Categories of impact evaluation questions

<table>
<thead>
<tr>
<th>Question</th>
<th>Description</th>
<th>Additional data requirements</th>
<th>Sample evaluation result and interpretation</th>
</tr>
</thead>
</table>
| What is the overall intervention impact on outcomes A, B, and C in group X? … in context Y? | Interventions often affect groups differently (heterogeneity of impacts). Measuring only average impact may hide these differences, so we need to break down impacts by population group | • Socio-demographic information of participants and comparison group (age, gender, income level, etc.)  
• To be able to disaggregate the results, the number of people covered by the evaluation (the sample size) needs to increase with each category of information that is to be analysed | The average increase in income is $40 for boys and $0 for girls. Older youth benefit more than younger youth ($30 versus $10, on average). Therefore, the intervention is not equally effective for all participants. We need to understand why groups benefit to a different extent and possibly adapt the programme's targeting and design to accommodate particular groups. |
| Do the outcomes vary across population groups?                          | This is the standard impact evaluation question.                                                                                               | n/a (standard data collection based on the method chosen)          | The average impact of the training intervention on the income of youth is +$20 per month. The intervention has a positive impact on participants' income |
| What is the short-term versus the long-term impact of the programme?    | The change in outcomes may not be constant over time. Short-term effects may vanish, while long-term effects may not manifest themselves until years after the intervention has ended | Data over an extended period of time (in practice, this often means following treatment and comparison groups for several years) | At the end of the programme, we observe an average monthly income for participants of –$5 (a loss) compared with the controls. Two years after the programme, the average increase in monthly income for the treatment group is $20. Those who participated in the training were not able to work as much as their peers during the course of the training, so they lost income. Over time, however, the training paid off and participants were able to secure higher incomes than their counterparts who did not participate. Looking only at short-term outcomes may provide misleading results |
| Does the intervention have spillover effects?                          | The intervention may have indirect effects on non-participants (positive and negative)                                                       | • Data beyond the treatment and comparison group, to include family or community members  
• Several treatment groups (one receives design A, one receives design B, etc.)  
• The number of people covered by the evaluation needs to be large enough to be able to create more than one treatment group as well as a comparison group | Not only do participants have a $20 higher average income, their neighbours also experienced a $5 increase. Participants apparently passed on the knowledge to others |
| Is intervention design A or intervention design B more effective?       | There is often ambiguity about the best possible intervention design. Questions can relate to comparing alternative interventions or combinations of programme components | • Several treatment groups (one receives design A, one receives design B, etc.)  
• The number of people covered by the evaluation needs to be large enough to be able to create more than one treatment group as well as a comparison group | The average increase in income is $5 for those who received training and $20 for those who received training and an internship. Thus, providing practical work experience in addition to training appears to significantly improve impact |
| What is the most effective dosage of the intervention?                 | More is not always better; finding the right balance of how much service to provide is important to maximize impact on the one hand and minimize costs on the other | • Several treatment groups (one receives design A, one receives design B, etc.)  
• The number of people covered by the evaluation needs to be large enough to be able to create more than one treatment group as well as a comparison group | The average increase in income is $0 for those who received 1 month of training, $20 for those who received 3 months, and $20 for those who received 6 months. Although 1 month of training was insufficient, 6 months of training had no additional benefit compared with 3 months of training. The optimal length of the training seems to be about 3 months |
| Why did the intervention (not) work? Why did it only work for part of the target population when the intervention had a certain duration? | Along with assessing the impact itself, it is crucial to understand how and why it took place as observed | Both quantitative and qualitative data, ideally triangulated to establish reasonable causal connections. For example, in-depth interviews with training participants, trainers and employers | Employers valued specific skills which could realistically be acquired by participants in 3 months' time. Remaining longer in training did not add additional value to the employees' skillsets. Employers did not want to incentivize longer training duration as they would lose employee working time during the training. Training only led to an increase in income for boys as employers tend to assign different tasks to girls, in which the skills transferred in the training are less in demand |
Process tracing involves the in-depth analysis of the different events linking an intervention to one or more intermediate or final outcomes, and their causal relations. Often (but not always) process tracing methods aim to develop and test theoretical mechanisms, which can be generalized to cover other interventions and contexts. In summary, process tracing is applied as follows:

1. **Developing a hypothesized causal mechanism for how change happens**

As a first step, we have to build the narrative of the process we are going to assess. This can be a project's theory of change, including the people and activities involved in it. It is important that the process is set out in its smallest individual elements, which should all be both essential for it to work and measurable.

For example: “Teachers conduct skills training for unemployed youth”; “Students attend skills training”; “Students acquire new knowledge and skills about how and where to look for jobs”; “Students search more and more efficiently for jobs”; “Students have a higher probability of finding a job”.

In order to make plausible claims regarding the causal linkages between the different parts of this mechanism, it is necessary to identify possible alternative explanations for the occurrence of each individual element and to look for evidence to confirm or rule out those explanations. For example: “Students acquired the knowledge on how and where to look for jobs independently”; “Students find jobs because of an improvement in the local labour market situation”.

2. **Defining and collecting the required evidence**

After defining the mechanism, or our theory, we need to define the empirical evidence required to analyse each link in the causal chain. This applies both for “our” mechanism and for the competing alternative hypotheses. Consequently, the previously identified evidence will be gathered through primary and/or secondary data collection. Sources for this evidence can be stakeholder interviews, programme documents, survey data, meeting minutes, and statistics, among others. Evidence should be collected in such a way that it can either confirm or refute the different competing hypotheses. It is good practice to triangulate methods, i.e. to use different methods to assess the same element from different angles.

3. **Assessing the evidence and drawing a conclusion**

The collected evidence is then examined in a procedure similar to that used in a criminal trial. In process tracing, we aim to establish a case that offers sufficient proof to reasonably assume that each element of the mechanism took place due to another element and that together they caused certain outcomes.

There are different tests for assessing the strength of the evidence for each hypothesis. For example, the “smoking gun” test refers to convincing evidence directly referring to the mechanism in question. So, for example, a statement from a skills training participant such as: “Thanks to the things I learned in the training, I have much more confidence to apply for jobs and I send out more applications than before” can make us fairly confident that this participant did not increase his or her job-search behaviour – which could be an intermediate outcome variable measured quantitatively – due to other reasons.

When assessing the evidence for different competing hypotheses, it is important to bear in mind that the strength of the overall evidence for a certain mechanism is always only as robust as the weakest evidence for one individual link. Finally, based on the conclusions from this exercise, the hypothesized mechanism, as well as the alternative hypotheses will be confirmed or ruled out.

---

For more tests and details on their application, see Bennett, 2010.
Rather than being substitutes for a quantitative impact evaluation, several of the abovementioned evaluation strategies can contribute to assessing a specific intervention. Using a mixed-methods approach therefore allows us to combine the strengths and offset the weaknesses of both qualitative and quantitative evaluation tools, allowing for a stronger evaluation design overall.

Employing a mixed-methods design for an impact evaluation in practice implies collecting both qualitative data, for example through field visits, key informant interviews or focus group discussions, and quantitative data, relying for instance on administrative data, surveys or secondary data sources, such as household surveys (see also table 3.5 in Note 3).

Mixed-method evaluation designs are also closely related to and inform theory-based impact evaluations. As White and Phillips (2012) observe, theory-based impact evaluations aim to establish causal links “by collecting evidence to validate, invalidate, or revise the hypothesised explanations, with the goal of documenting the links in the actual causal chain”. They often seek to combine all available quantitative and qualitative evidence to establish beyond reasonable doubt that an intervention impacted its participants. Box 5.9 introduces the methodology of process tracing as a theory-based impact evaluation methodology.
KEY POINTS

1. Impact evaluations answer cause-and-effect questions to determine whether and to what extent an intervention caused observable change. Understanding impact requires us to isolate the effects of the intervention from other factors influencing beneficiary outcomes.

2. Quantifying impacts of interventions requires estimating the counterfactual: that is, what would have happened to beneficiaries in the absence of the intervention. To this end, most quantitative impact evaluation designs rely on having a comparison group that shares as many characteristics with the beneficiaries as possible.

3. Observational impact evaluation designs include difference-in-differences and matching methods. They can be applied in a broad range of contexts and based on secondary data sources, but for some interventions these methods might not be able to estimate impacts credibly. Experimental designs that rely on some degree of randomization can produce highly credible impact estimates but can be costly and difficult to implement for certain interventions.

4. To maximize learning about “why” interventions worked, or did not work, used mixed-method approaches which build on qualitative and quantitative data and make use of several methodologies for analysis.

KEY RESOURCES


REFERENCES


Case study

ASSESSING RURAL MICRO-ENTERPRISE GROWTH THROUGH DIFFERENT EVALUATION METHODS

Disclaimer: This is a fictional case study. All information contained within has been invented for learning purposes.
Learning objectives

By the end of this case study, readers will be able to demonstrate the following learning outcomes:

- identify impact evaluation methods without being told the specific employed method
- explore the problem of producing estimates of the causal impacts of a development programme, and the various ways of estimating the impacts using comparison group designs
- develop an intuitive understanding of when and how impact evaluation methods will produce biased results by learning about the concept of selection bias and how comparison group designs are only as good as their ability to eliminate selection bias.

Introduction and case study context

Micro-enterprises are vital in rural areas with limited formal employment options, both for providing informal employment and for ensuring household economic security for business owners. However, once a business has been started, there are a number of challenges to growth.

What can be done to help develop rural businesses? The Training for Rural Economic Empowerment (TREE) programme tests some of these constraints to understand what kind of financial and training services have impacts on enterprise growth, for whom and why.

The International Labour Organization (ILO) conducted the training tested here using their TREE methodology, a development approach which ensures that women and men living in poverty gain the skills and knowledge they need to improve their incomes and take a more active role in shaping their communities. Moreover, a local microfinance organization delivered loans to individuals of US$200 at a discounted annual interest rate of 20 per cent (reduced from the standard 25 per cent).

This case study focuses on 400 rural micro-enterprise owners who were offered the chance to participate in a skills training programme and to receive loans. In total, 144 out of the 400 business owners took part in the training and received the loan.
Comparing different impact evaluation methods:
Did TREE work?

Did the TREE programme work? Did the programme improve business profits? What is required in order for us to measure whether a programme worked, or whether it had an impact?

In general, to ask if a programme works is to ask if the programme achieves its goal of changing certain outcomes for its participants, and ensure that those changes are not caused by some other factors. We need to simultaneously show that, if the programme had not been implemented, the observed changes would not have occurred (or would be different). But how do we know what would have happened? Measuring what would have happened in the absence of the programme requires entering an imaginary world in which the programme was never offered to these participants. The outcomes of the same participants in this imaginary world are referred to as the counterfactual. Since we cannot observe the true counterfactual, the best we can do is to estimate it by mimicking it.

The key challenge of programme impact evaluation is constructing or mimicking the counterfactual. We typically do this by selecting a group of people that resemble the programme participants as much as possible but who did not take part in the programme. This group is called the comparison group, which ideally differs from the group of beneficiaries only insofar as they did not participate in the programme.

We then estimate “impact” as the difference observed at the end of the programme between the outcomes of the comparison group and the outcomes of the programme participants.

Importantly, the impact estimate is only as accurate as the comparison group is successful at mimicking the counterfactual. Therefore, the method used to select the comparison group is a key decision in the design of any impact evaluation.

That brings us back to our questions: Did the project work? What was its impact on the outcome being evaluated, namely business profits?

In our case, the intention of the programme is primarily to “improve enterprise growth”, and profits (measured in US$) are the key outcome indicator. So, when we ask if this project worked, we are asking if it improved business profits. The impact is the difference between profits after the businesses have been exposed to the intervention and what their profits would have been if the intervention had never existed.

What comparison groups and impact evaluation methods can we use? The following (fictional) experts illustrate different methods of evaluating impact. Table 5.2, at the beginning of this note, presents an overview of different evaluation methods for your reference.
Part I: News release: “Training for Rural Economic Empowerment” programme helps businesses grow

TREE celebrates the success of its programme. It has made significant progress in its goal of helping businesses grow through provision of loans and skills training. The achievement of the TREE programme demonstrates that providing skills training to business owners, combined with loans to ease capital constraints, can produce significant gains.

Just before the programme started, businesses were making profits of $286, on average. But after spending just a few months in the programme, profits for these businesses doubled!

### Discussion topics

1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the challenges with this type of evaluation?

Part II: Opinion: The “Training for Rural Economic Empowerment” project not up to the mark

With an estimated outreach of 6 million trainees, a continuously growing network of more than 17,000 trainers and 200 master trainers in 2,500 partner institutions, TREE is one of the biggest training systems used for the support of micro- and small enterprises (MSEs) currently on the market. But do the profits of its businesses actually double in size, as suggested in the first example? Recent evidence suggests otherwise.

An independent team of evaluators was hired to verify these findings. The team compared profits of TREE businesses to profits of other businesses in nearby villages. They found that TREE businesses grow their profits by only a meagre $64, and not $286 as originally estimated. That's only a 12 per cent increase in profits after 6 months of the TREE programme paired with loans. It seems that income estimates were severely overestimated and that ILO's assurances about the successes of the programme were false.
CASE STUDY: ASSESSING RURAL MICRO-ENTERPRISE GROWTH THROUGH DIFFERENT EVALUATION METHODS

NOTE 5. IMPACT EVALUATION METHODS FOR YOUTH EMPLOYMENT INTERVENTIONS

Discussion topics

1. What type of evaluation does this opinion piece imply?
2. What represents the counterfactual?
3. What are the challenges with this type of evaluation?

Part III: Letter to the Editor: Independent evaluators should consider evaluating fairly and accurately

There have been several unfair reports in the press concerning programmes implemented by the ILO. A recent article by an independent evaluator claims that TREE is, in reality, not helping businesses grow. However, their analysis uses the wrong metric to measure impact. It compares the profits of TREE business with other businesses in the village – not taking into account the fact that TREE targets those whose profits are particularly low initially. If TREE simply recruited the biggest businesses into their programmes, and compared them to their smaller counterparts, they could claim success without conducting a single training session or providing a single loan. But TREE does not do this. And, realistically, TREE does not expect its smaller businesses to overtake the bigger businesses in the village. It simply tries to initiate an improvement over the current state.

Therefore the indicator should be improvement in profits – not the final profit level. When we repeated the analysis using the more appropriate outcome measure, the TREE businesses improved at twice the rate of the non-TREE businesses (US$283 profit increase compared to US$162). Had the independent evaluators thought to look at the more appropriate outcome, they would recognize the incredible success of TREE. Perhaps they should enrol in some TREE training themselves.

Discussion topics

1. What type of evaluation does this letter imply?
2. What represents the counterfactual?
3. What are the challenges with this type of evaluation?
Part IV: Designing your own evaluation to assess the impact of TREE

As discussed earlier in this case study, there are challenges and reservations with respect to all three of the evaluation methods detailed above. It is now your turn to design an impact evaluation for the TREE programme, assuming that the programme is yet to be implemented.

To begin, assume that your research team has surveyed several thousand micro-enterprises from a census of businesses and selected 1,600 business owners to be included in the evaluation. All of these business owners have expressed an interest in receiving the ILO training and in participating in the loan programme. However, due to resource constraints, your project manager tells you that the training and loan programme can only be offered to a maximum of 800 businesses.

We also assume that the key outcome of interest remains “business profits”.

1. How would you design the evaluation? In particular, how would you select a comparison group?

2. When would your research team collect data and from which businesses (all 1,600 or only a subset)?

3. Why do you think this is a reliable impact evaluation method that overcomes some or all of the shortcomings of the three methods discussed above?