ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

IFMR
INSTITUTE FOR FINANCIAL MANAGEMENT AND RESEARCH

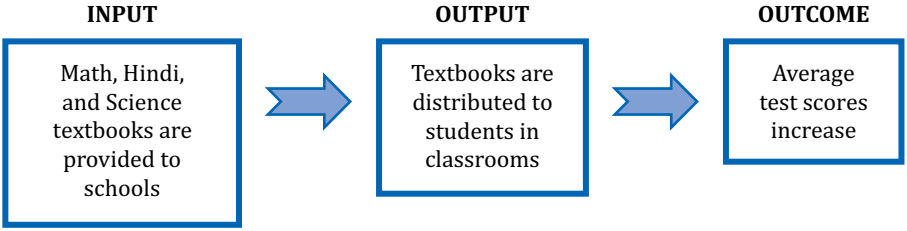
clear
Regional Centers for
Learning on Evaluation and Results

Impact Evaluation Methods

Why Randomise?

What is Impact Evaluation?

All development programmes have an objective or goal. For example, a programme that distributes free textbooks to primary school students might aim to increase students' test scores. A **Theory of Change** maps how a programme's inputs (textbooks) cause these larger outcomes (increased average test scores).



An **impact evaluation** determines whether or not a programme had an effect on a specific outcome and quantifies the magnitude of this impact. In our textbook example, an impact evaluation will ask:

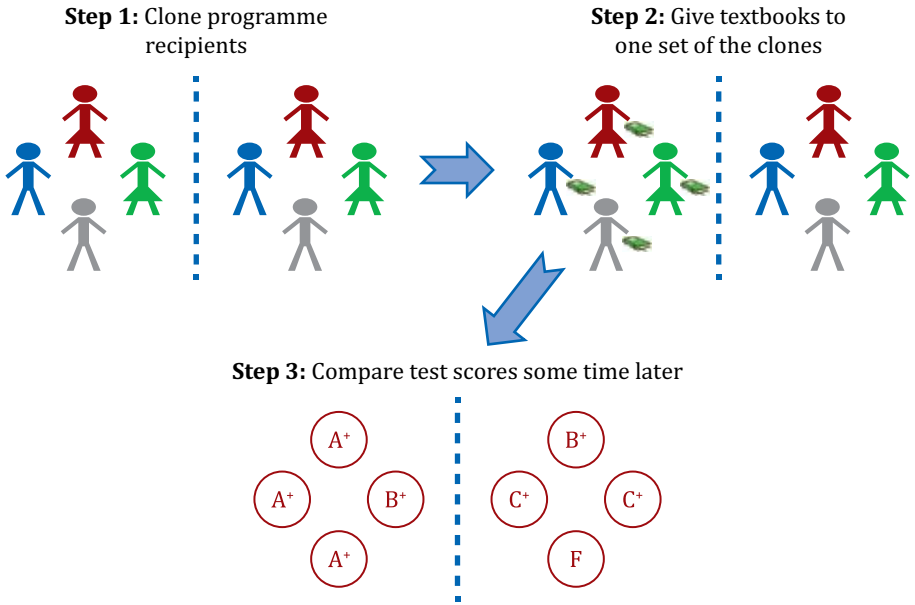
- Does providing textbooks to students increase average test scores of students?
- If yes, how large is the impact on average test scores?

How to Measure Impact?

The impact of a programme is measured by comparing the outcome some time after the programme has been introduced with what is known as the **counterfactual** – the outcome at the same point in time had the programme not been introduced. *The counterfactual represents the state of the world that programme participants would have experienced in the absence of the programme*. Mathematically, the impact of a programme can be expressed as:

$$Impact = Y_T - Y_C$$

Where Y_T is the outcome (e.g. average test scores) for the **treatment group**, or the group of participants receiving the programme, and Y_C is the outcome (e.g. average test scores) for the **counterfactual group**.



The perfect counterfactual would involve cloning programme participants. An impact evaluation using cloned participants would look as shown above.

The **problem** of course is that we cannot clone programme beneficiaries. In a real-world setting, we have no way of actually observing the counterfactual measure. To overcome this problem, impact evaluations rely on finding a **comparison group** of non-participants that provides an **estimate** for the counterfactual. Therefore, in practice, the impact of a programme is calculated as:

$$Impact = Y_T - \hat{Y}_C$$

Where $\hat{Y}_C$ is the estimate of the counterfactual outcome that is found using the **comparison group**.

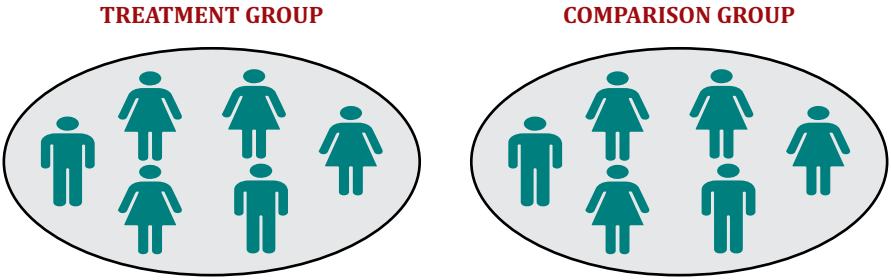
Causality

In an accurate impact evaluation we can be sure that the programme itself and not any other factor caused the changes in the outcome we observed. If an impact evaluation is able to identify the changes in outcomes that are directly caused by

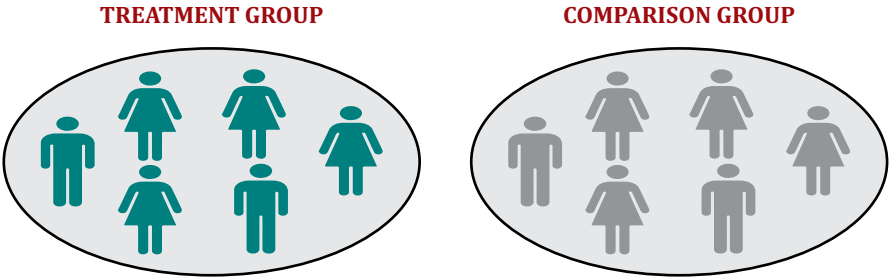
the programme itself, we say that the impact evaluation provides a **causal** estimate of the impact of the programme.

So which impact evaluations are accurate and causal and which are not? Whether or not an impact evaluation is accurate and provides a causal estimate depends on how closely the comparison group resembles how the programme participants would have been without the programme, the counterfactual.

If the comparison group, on average, has similar characteristics to the counterfactual, then the Impact Estimate is accurate (CAN claim causality).



If the comparison group, on average, does NOT have similar characteristics to the counterfactual, then the Impact Estimate is NOT accurate (CANNOT claim causality).



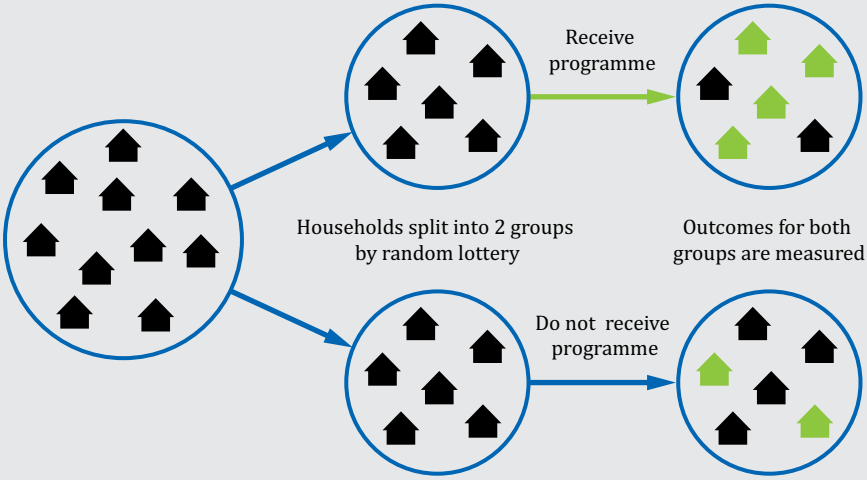
Impact Evaluation Methods

The question then is **how to construct a comparison group that accurately mimics the counterfactual?** There are in fact many types of methods for constructing a comparison group. Some methods create a more accurate comparison group than other methods. These "other methods" tend to produce misleading results because they rely on assumptions that are often unrealistic.

Randomisation Provides the Most Credible Impact Estimate

Randomised evaluations give the most credible impact evaluation estimates because randomisation ensures there are no systematic differences between participants and the comparison group. In a randomised experiment, one essentially flips a coin to determine who receives the programme and who does not. When a large enough number of individuals are randomised, the resulting treatment and comparison groups will have similar characteristics on average, meaning that any difference in average outcomes must be due to the programme itself. This is the power of randomisation and is the reason why these evaluations provide the most accurate impact estimates.

To further understand why a randomised evaluation is the most credible, consider two other commonly used impact evaluation methods that suffer from poor counterfactual measures: a **before/after** and a **simple difference**.

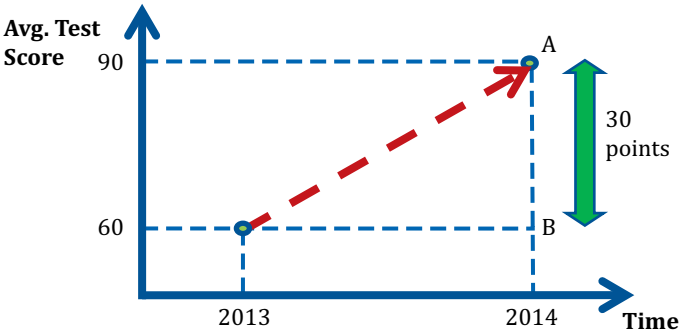


Features of Randomisation:

- Groups are statistically similar before the programme
- Two groups continue to be similar, except for the programme
- Any difference in outcomes between the groups can be attributed to the programme

i. Before/After

- In a before/after estimation, impact is measured by comparing the outcomes (e.g. test scores) from the **same** group of programme recipients both before and after the programme.
- For a group of students who received the textbook programme in 2014, a before/after study would take the difference between test scores for these students at the end of 2014 with test scores for the same group of students in 2013, i.e. before they received textbooks. The impact of the programme would thus be the difference in test scores between 2014 and 2013 (90 - 60 = **30 points**)



- **For this before/after impact estimate to be causal, we need to assume that between 2013 and 2014 the programme was the only factor that could have led to the increase in test scores**. However, many other factors besides the textbooks could have caused an increase in test scores over that time period. One of these factors could be the increase in students' general knowledge associated with being 1 year older. There is no way to know whether the 30 point increase in test scores is due to textbooks, or the gains in maturity and knowledge accumulated through a year of growing older.

ii. Simple Difference

- A simple difference study measures impact by comparing ex-post outcomes between a group receiving the programme and a group not receiving it. In the textbook example, this would mean taking the difference between test scores for students who received textbooks and test scores for a group of students who did not receive textbooks. Suppose the test score results after the programme were the following:

Table of Test Score Results

Average test score for students receiving textbook programme	83
Average test score for students without textbook programme	68
Difference	**15 points**

- Before we can accept that the impact of the textbook programme was 15 points, we must **assume that children who did not receive textbooks were identical, on average, to the children who did receive the textbooks except for the textbook programme itself.**
- Differences between treatment and comparison groups can invalidate this assumption. If say the students receiving textbooks are more likely to be from private schools than comparison group students, then this is likely to invalidate our assumption. Students from private schools may come from families with higher average incomes, which may allow them to invest in resources that would have made them more likely to score higher test scores even in the absence of the programme. In this case we can't say that the programme caused the 15 point impact because this gain is due to both the textbook programme **and** income differences between students.

A **randomised evaluation** overcomes the problems in both a before/after and a simple difference study. A randomised evaluation of the textbook distribution programme would involve identifying a sample of schools that are eligible to receive the programme, and then randomly assigning eligible schools into a treatment group that receives the textbook programme and a comparison group that does not.

- Given a large enough sample of schools, the treatment and comparison schools would be similar on average along all possible characteristics (*e.g. percent of children in private schools, education level of teachers, teaching effort and motivation, etc.*). Since groups are similar on average, any difference in the outcome between the treatment and comparison groups **must** be due to the programme itself and not any other factor. In other words, **since the comparison group is exactly similar to the treatment group, besides the programme itself, the comparison group is a very near approximation of the counterfactual**. In this way a randomised evaluation provides an accurate and causal impact of the textbook programme.

CLEAR South Asia
J-Pal South Asia at IFMR
AADI, 2 Balbir Saxena Marg, Hauz Khas,
New Delhi – 110 016