

# BEET: Building Evidence Ecosystems about Employability and Skills Development through Training

## Training Session 2: Why Randomize

Nayera Adly Husseiny, Senior Policy Manager, J-PAL MENA

Institutional Host



Founding Partners



SAWIRIS FOUNDATION  
مؤسسة السويريس



Community  
Jameel

Additional Support



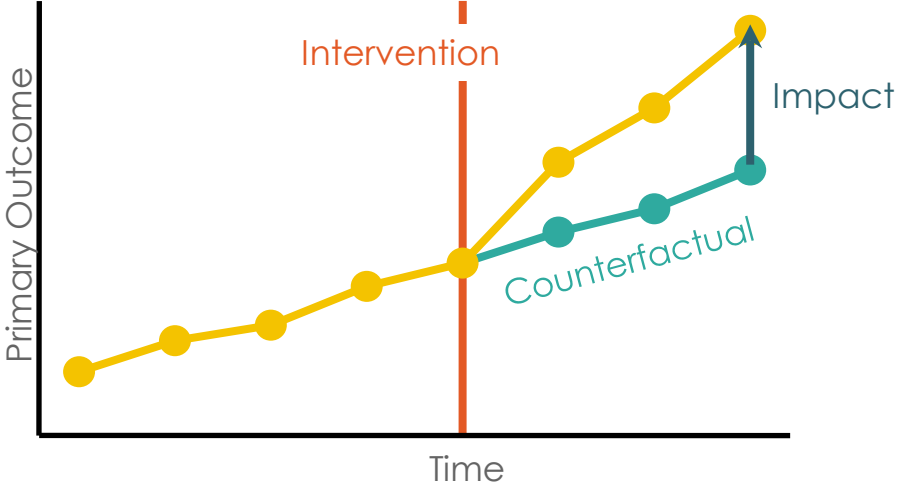
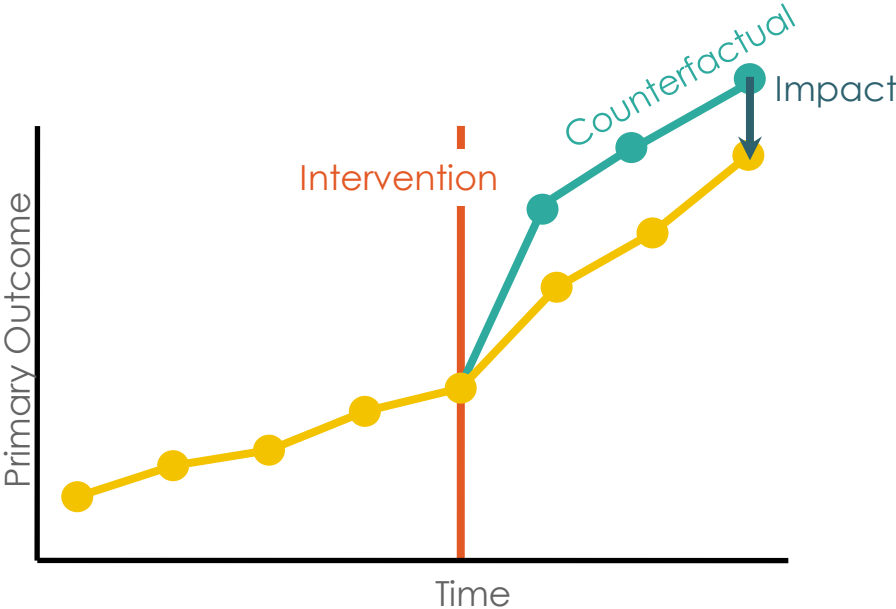
# Course Overview

1. Why Evaluate & Theory of Change?
- 2. Why Randomize?**
3. How to Randomize & Sample Size
4. Generalizability

# Lecture Overview

- I. **What is impact?**
- II. Why randomize: A case study
  - I. Non/Quasi-experimental methods
  - II. Experimental method
- III. When to randomize
- IV. Conclusion

# Quick Recap: What is Impact



# Impact: Definition

The impact of a program is defined as a comparison between:

**What actually happens** after the program has been introduced

**What would have happened** had the program not been introduced (i.e., the “**counterfactual**”)

# Impact: How Can we Measure it?

In order to assess the impact of a program, we need to understand the **counterfactual**, i.e., the state of the world that program participants would have experienced in the absence of the program

**Problem:** The counterfactual never happened so it cannot be observed

**Solution:** We need to “mimic” or construct the counterfactual

## Constructing the counterfactual:

## Determining a comparison group

Usually done by selecting a group of individuals that **did not** participate in the program

This group is usually referred to as the **control group** or **comparison group**

How this group is selected is a **key decision** in the design of any **impact evaluation**

# Constructing the counterfactual

## Comparing apples to apples

Determine a comparison group that **does not differ systematically** from the treatment group at the outset of the program/evaluation,

so that the difference that subsequently arises between them can be **attributed** to the program rather than to other factors.

Treatment



Source: freepik

Comparison



# Impact evaluation methods

- Different **impact evaluation methods** estimate the counterfactual in different ways
- As we will see, they rely on different assumptions to be able to construct a counterfactual
- Impact evaluation methods answer **cause-and-effect** questions in the form of: What is the effect of [program, policy, or intervention] on [desired outcomes]?

# Overview of impact evaluation methods

## Non/Quasi-experimental methods

- Pre-post comparison
- Simple difference
- Statistical matching
- Regression discontinuity design
- Difference-in-difference
- Instrumental variables

## Experimental method: Randomized evaluations

Also known as:

- Randomized controlled trials (RCTs)
- Random assignment studies
- Randomized field trials
- Social experiments
- Randomized (controlled) experiments

What is an impact evaluation question that is relevant for the programs and policies you work on?

How have you tried to estimate the impact of your program or policy previously?

# Lecture Overview

- I. What is impact?
- II. Why randomize: A case study**
  - I. Non/Quasi-experimental methods
  - II. Experimental method
- III. When to randomize
- IV. Conclusion

# Case study: Principal training to improve student achievement in Puerto Rico

- Problem: Low student achievement and high drop-out rates
- Two-week intensive management training for school principals, followed by biweekly workshops
  - Focused on personnel management, instruction planning, goal setting and monitoring, school culture, and personal leadership

Study: [« Principal Training to Improve Student Achievement in Puerto Rico »](#) (Bobonis et al., ongoing)

# Thought experiment: Designing an impact evaluation of the principal training program

- Imagine you want to design an impact evaluation to answer the following question:

*What is the effect of the principal training on student learning?*

- How would you construct a counterfactual? Where can we find a good **comparison group**?

# Lecture Overview

- I. What is impact?
- II. Why randomize: A case study**
  - I. Non/Quasi-experimental methods**
  - II. Experimental method
- III. When to randomize
- IV. Conclusion

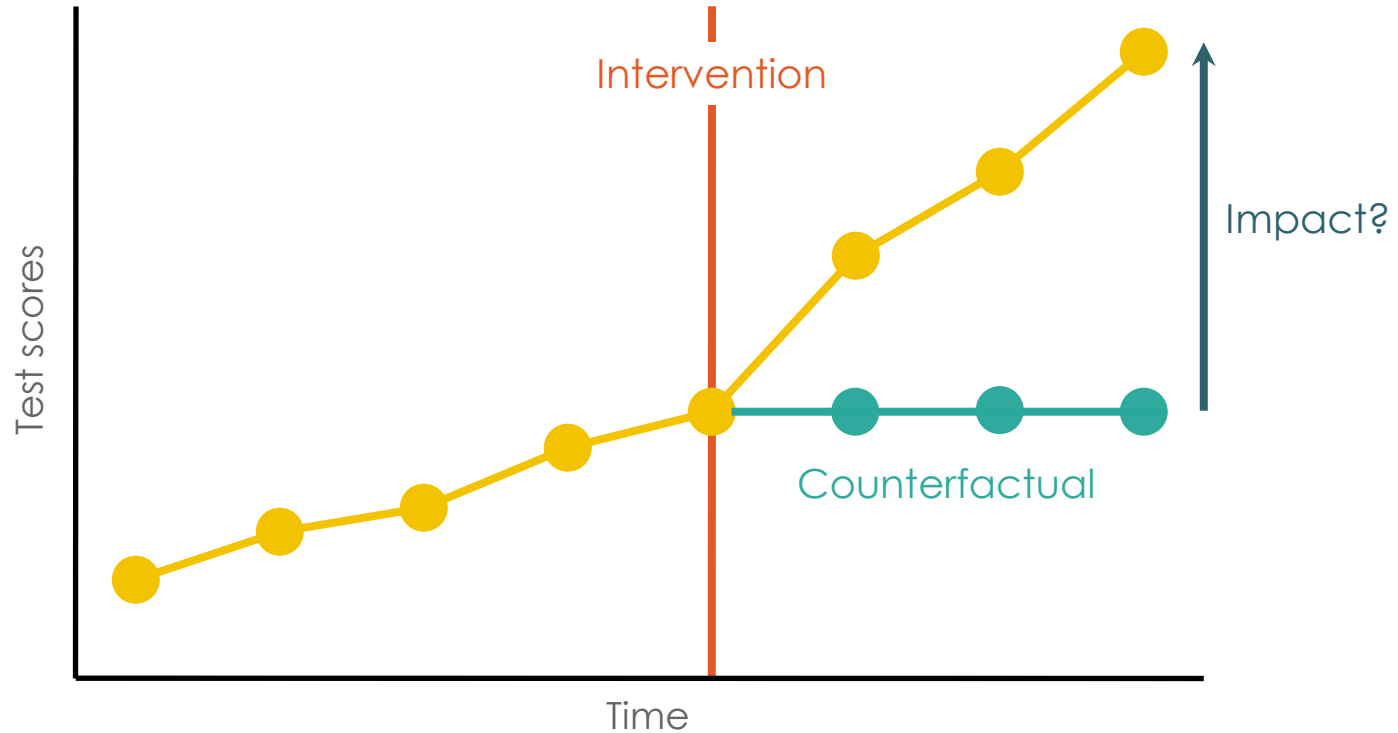
# Non/Quasi-experimental impact evaluation methods

Let's look at different non-experimental methods of estimating the impacts using the data from the schools where the program was implemented

1. Pre – Post (Before vs. After)
2. Simple difference
3. Matching
4. Regression Discontinuity Design
5. Difference-in-difference

# Method 1: Pre – Post

Compare test scores before and after the program



Is this a good comparison group? Are we confident that any difference between the groups resulted from the program?

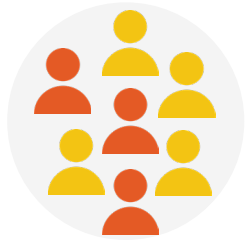
---

### **Not Likely!**

- Relies on **very strong assumption** that test scores of students who followed the training would not have changed over time in the absence of the program.
- Other things likely influence these outcomes over time.

# Method 2: Simple difference

Compare schools of participants with those of non-participants



Motivated principals  
apply to the program

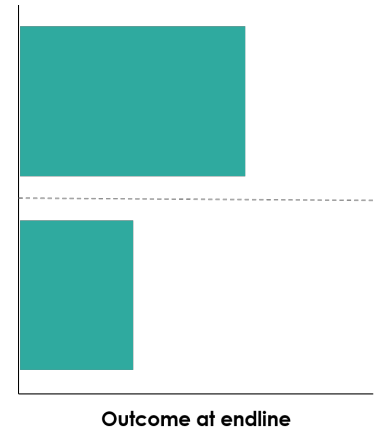


Other principals in the  
program area that did  
not apply

Training  
program

continue under business as usual

Compare outcomes  
at the end of the  
program



Is this a good comparison group? Are we confident that any difference between the groups resulted from the program?

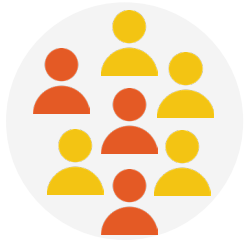
---

### Not Likely!

- Principals who apply might be more informed, motivated, or invested in student outcomes.
- This creates “**selection bias**”: participants may differ from non-participants in important ways.
- Hard to disentangle whether the changes in the outcomes of students are due to the program or due to some other aspects of the principals / their schools.
  - E.g. potentially hard-to-capture underlying personality traits or other so-called “unobservable” factors

# Method 3: Matching

Try to identify similar principals among non-participants



Participants in the program

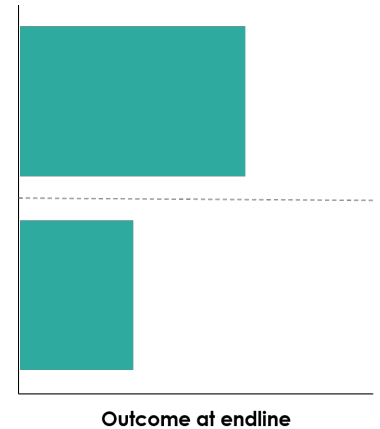


continue under business as usual



Non-participant pool

Compare outcomes at the end of the program



# Is this a good comparison group? Are we confident that any difference between the groups resulted from the program?

---

Yes, **if** we find individuals in the group of non-participants

- that are very similar to our participants across observable characteristics (something we can verify)
- that are also similar across so-called non observable characteristics (something we cannot test).

---

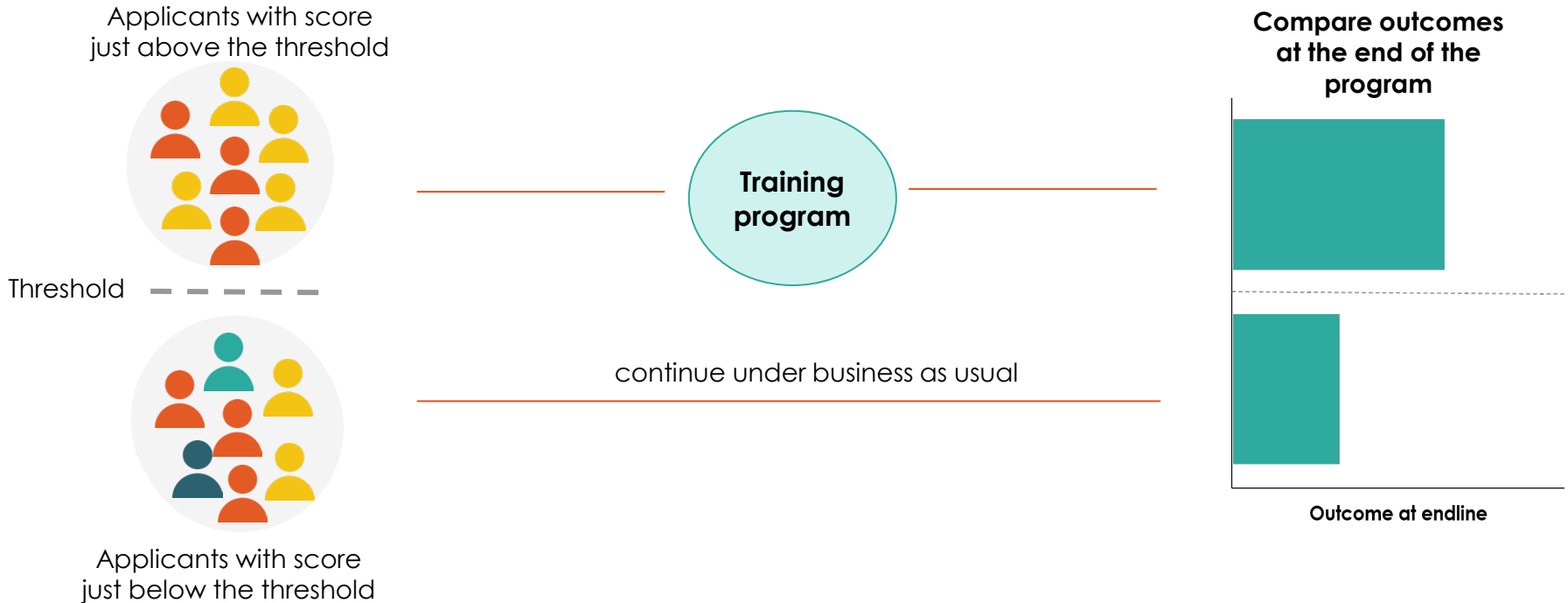
No, if we do not find individuals in the group of non-participants

- that are very similar to our participants across observable characteristics, or
- if they are not similar across characteristics that we cannot observe (something we cannot test).

# Method 4: Regression discontinuity design

## Compare accepted applicants just above and below a threshold

Imagine the program is oversubscribed and you admit principals based on a performance threshold.



Is this a good comparison group? Are we confident that any difference between the groups resulted from the program?

---

Yes, **if**

- individuals just below and just above the eligibility threshold are indeed similar
- we have enough individuals close to this score (sample size)
- **Note:** Impact estimate only valid for those around the eligibility cutoff.

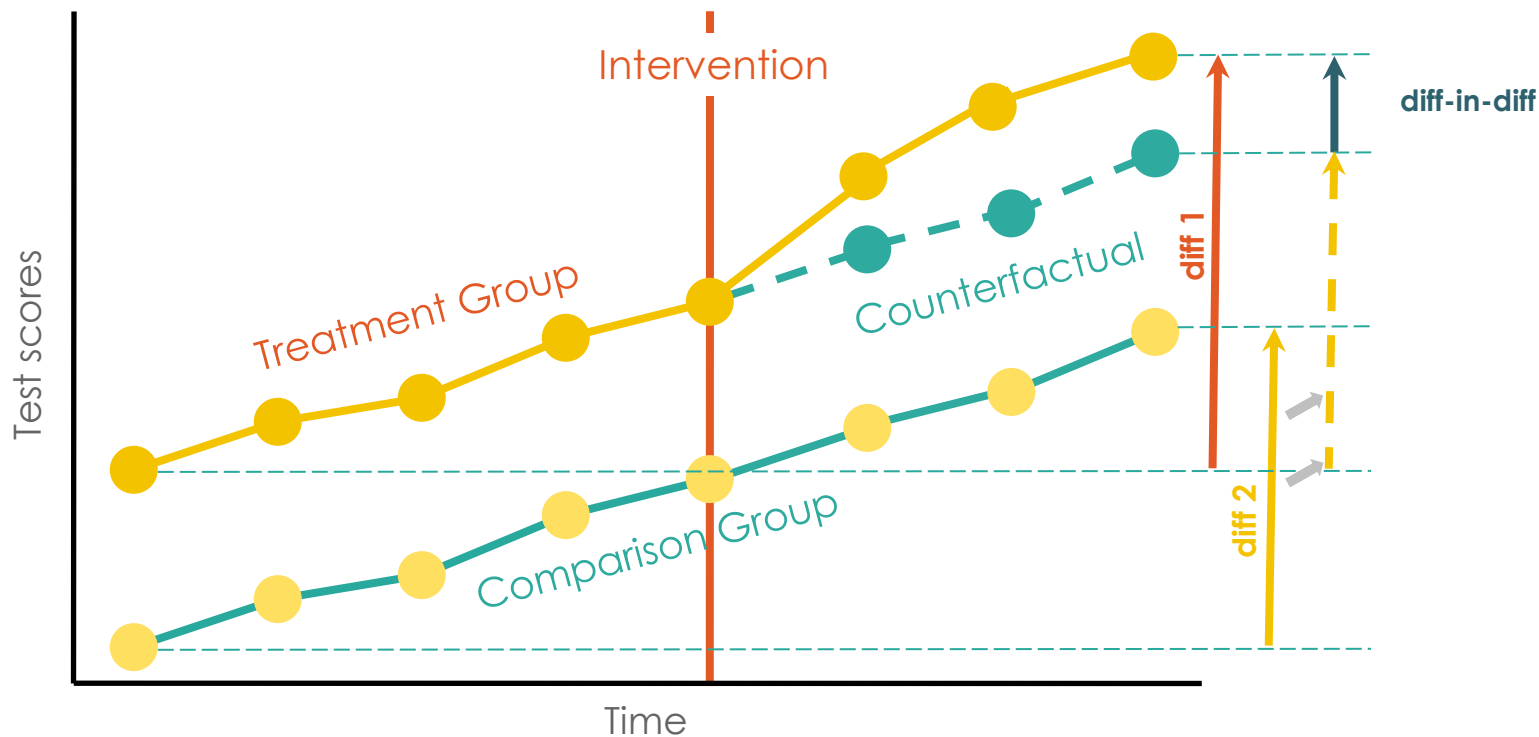
---

No,

- if there are not enough individuals around the eligibility threshold,
- if some individuals can manipulate their score in order to qualify for the program, or
- if other things happen at the cut-off.

# Method 5: Difference-in-differences

Find a group with a similar trend and compare *changes* over time



# Is this a good comparison group? Are we confident that any difference between the groups resulted from the program?

---

Yes, **if**

- the outcomes of the two groups would indeed have developed in parallel in the absence of the program, i.e.
- other factors that may have affected the outcome over time are the same for both groups.

---

No,

- If the trend in the treatment group would have been different from that in the comparison group in the absence of the program (something we cannot test).

# Non- and quasi-experimental methods rely on being able to “mimic” the counterfactual under certain assumptions

**The methods we just discussed are often referred to as non-experimental or quasi-experimental methods.**

**They rely on assumptions that need to hold to create a credible counterfactual.**

**Challenge: Many of these assumptions are not testable. The credibility of the evaluation will depend on the credibility of the assumptions.**

# Lecture Overview

I. What is impact?

## II. Why randomize: A case study

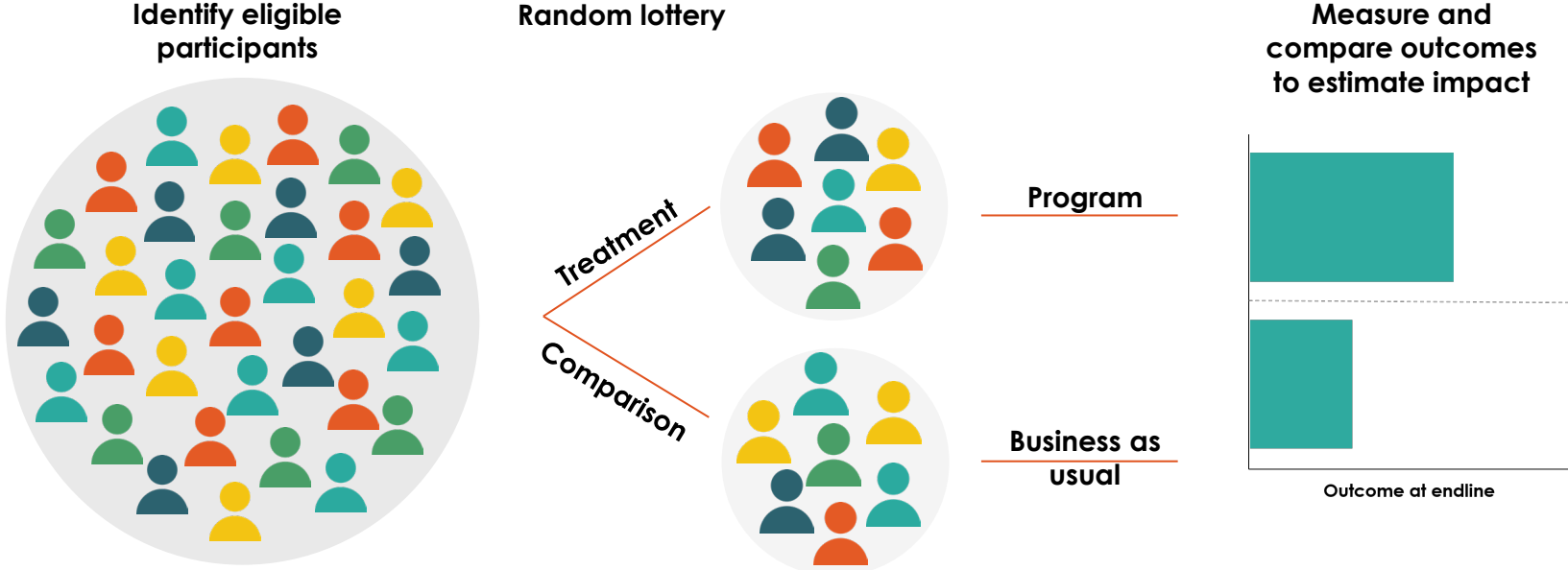
I. Non/Quasi-experimental methods

### II. Experimental method

III. When to randomize

IV. Conclusion

# Randomized evaluations use random assignment to mimic the counterfactual and estimate a program's impact



Is this a good comparison group? Are we confident that differences between the groups resulted from the program?

---

**Probably yes!** If properly designed and conducted, randomized evaluations provide a very credible estimate of the counterfactual.

# Why randomize?

**Key advantage of randomized evaluations (or RCTs):** Due to random assignment, members of the treatment and comparison groups **do not differ systematically at the outset of the evaluation**. Thus, the difference that subsequently arises between them can be attributed to the program, rather than to other factors.

Treatment



Source: freepik

Comparison



# Key steps in conducting a randomized evaluation

1. **Design** the study carefully
1. **Randomly** assign units to treatment or control
1. Collect **baseline** data
1. **Verify** that assignment looks random
1. **Monitor** process so that integrity of evaluation is not compromised

**Optional:** We do not need data to determine a comparison group ex post. Due to random assignment, we know that groups on average are very similar. In practice, baseline data collections are still common for different reasons.

# Key steps in conducting a randomized evaluation (continued)

6. **Collect follow-up data** for both the treatment and control groups
6. Estimate program **impacts** by comparing mean outcomes of treatment group vs. mean outcomes of the control group
6. Assess whether program impacts are **statistically** significant and **practically** significant

# Lecture Overview

- I. What is impact?
- II. Why randomize: A case study
  - I. Non/Quasi-experimental methods
  - II. Experimental method

## III. When to randomize

- IV. Conclusion

# The impact evaluation method we choose matters!

- There are many ways to mimic the counterfactual, i.e., to estimate a program's **causal impact**
  - Different impact evaluation methods can yield very different estimates of impact
  - Different methods may be more or less appropriate under different circumstances
- If properly designed and conducted, randomized evaluations provide a very credible method to estimate the impact of a program
- The credibility of other impact evaluation methods will hinge on a number of assumptions – whether these hold will depend on the evaluation at hand

# Discussion: What is the most convincing argument you have heard **against** RCTs?

- A. Cost considerations
- B. Ethical concerns
- C. Complexity
- D. Lack of generalizability or external validity
- E. Identifying impacts only without mechanisms (i.e., it functions as a "black box")

# When to do a randomized evaluation?

- When you have an important question about impact, and the answer to this question will drive policy or programmatic decisions
- When resources are being invested in a new program, but you do not yet have evidence about its impacts
- Timing—not too early and not too late
- Time, expertise, and money to do it right

# When NOT to do a randomized evaluation

- When the program is premature and still requires considerable “tinkering” to work well
- When the project is on too small a scale to randomize into two “representative groups”
- If a positive impact has been proven using rigorous methodology and resources are sufficient to cover everyone
- If the program has already begun and you are not expanding elsewhere or considering program alterations



**Thank you**