J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Why & When to Randomize

Course overview

1. Why Evaluate
2. Theory of Change & Measurement
3. Why & When to Randomize
4. How to Randomize
5. Sample Size & Power
6. Ethical Considerations for Randomized Evaluations
7. Threats & Analysis
8. Randomized Evaluation from Start to Finish
9. Applying & Using Evidence
10. The Generalizability Framework

Learning objectives

- Learn what impact is and how different impact evaluation methods aim to measure it.
- Understand the concept of the *counterfactual* and be able to critically discuss the credibility of the counterfactual for a given evaluation.
- Be able to assess when or when not to randomize.

Lecture overview

I. What is impact?

II. Why randomize: A thought experiment

    I. Non-experimental methods

    II. Experimental method

III. When to randomize

Lecture overview

I. **What is impact?**

II. Why randomize: A thought experiment

   I. Non-experimental methods

   II. Experimental method

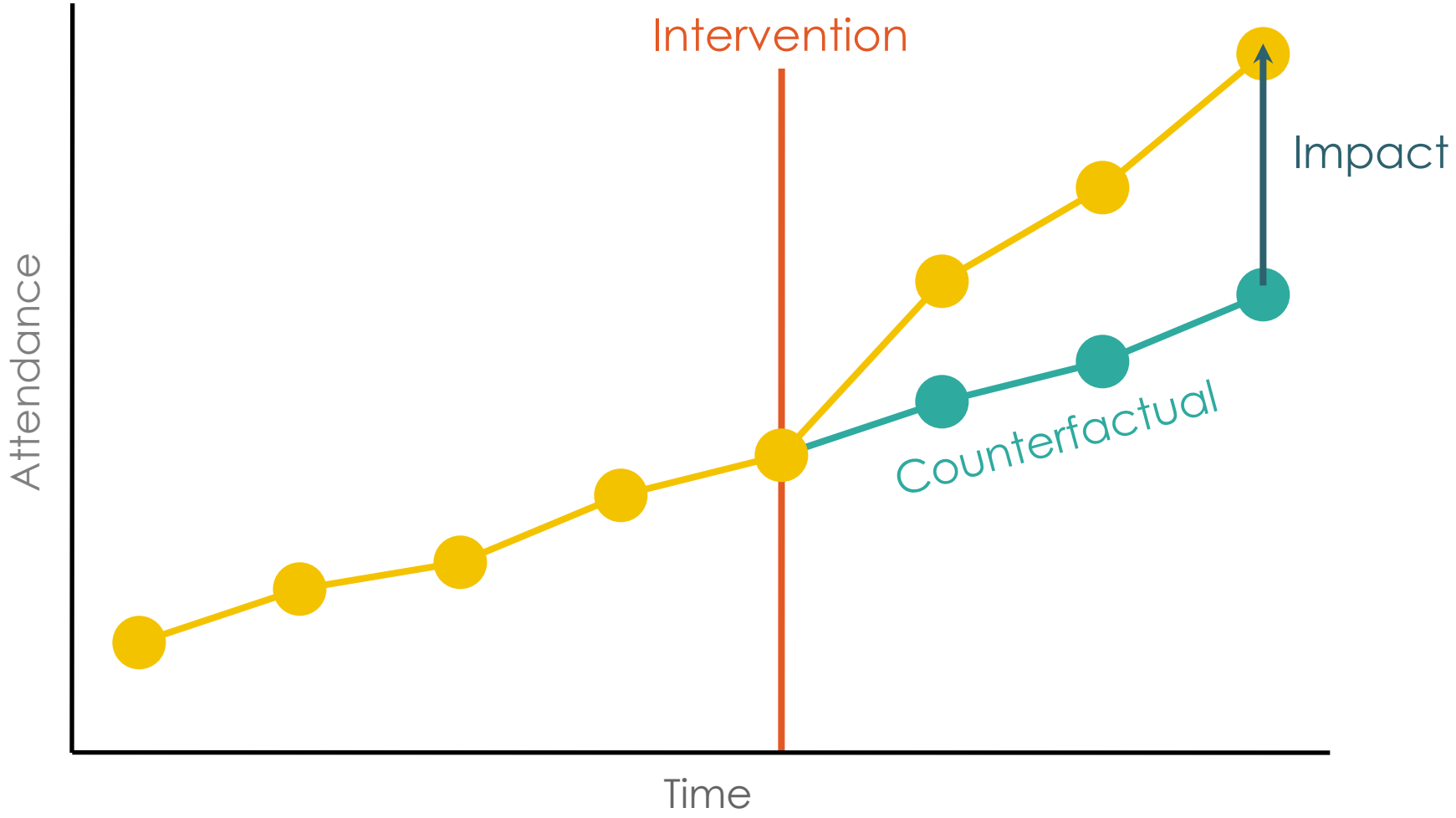
III. When to randomize

Impact: Definition

The causal 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")

What is the impact of text message reminders on attendance for adult learners?



Impact: How can we measure it?

In order to assess the causal 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 or that receives the **status quo**
- 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

Goal: Determine a comparison group that does not differ systematically from the treatment group at the outset of the program/evaluation, so that differences that subsequently arise between them can be attributed to the program rather than to other factors.

Treatment

Source: freepik

Comparison

Overview of impact evaluation methods

Non-experimental methods

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

Experimental method: Randomized evaluations

Also known as:

- Randomized controlled trials (RCTs)
- Randomized (controlled) experiments
- Field experiments
- Social experiments

The impact evaluation method we choose matters!

- Impact evaluation methods answer *cause-and-effect* questions: What is the effect of [program, policy, intervention] on [outcomes]?
- Different impact evaluation methods estimate the counterfactual in different ways
  - Different methods can yield very different estimates of causal impact
  - Different methods may be more or less appropriate under different circumstances
- As we will see, these methods rely on different underlying assumptions to be able to construct a credible estimate of the counterfactual
  - Whether these assumptions hold will depend on the evaluation at hand

What types of cause-and-effect questions can randomized evaluations help to answer?

- How effective is a given program?
  - Who benefits most?
- How do different versions of a program compare to one another?
  - Which components work or do not work? How do these function together?
- How do program impacts compare under different delivery mechanisms?
  - How to accurately target beneficiaries or respondents?
  - How to increase program take-up?
- How cost-effective is a program?
  - How does it compare to other programs designed to accomplish similar goals?
- How accurate is a measurement tool for a given outcome?
  - How do measured outcomes compare under different versions of a survey?

Discussion question

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 thought experiment

   I. Non-experimental methods

   II. Experimental method

III. When to randomize

An email-based social support program to reduce frontline worker burnout

- **Challenge:** Many frontline workers report experiencing emotional exhaustion and other characteristics of burnout, contributing to poor health outcomes for these public service employees and high turnover rates for their employers.

- **Intervention:** A six-week email campaign to 911 dispatchers across nine cities in the United States encouraging participants to share their experiences and engage with other dispatchers through an anonymous online platform.

Study: "Reducing Burnout and Resignations among Frontline Workers: A Field Experiment" (Linos, Ruffini, & Wilcoxen 2022)

Thought experiment: Designing an impact evaluation

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

Does this email-based social support program reduce frontline worker burnout?

- How can we identify a good comparison group to estimate the counterfactual?

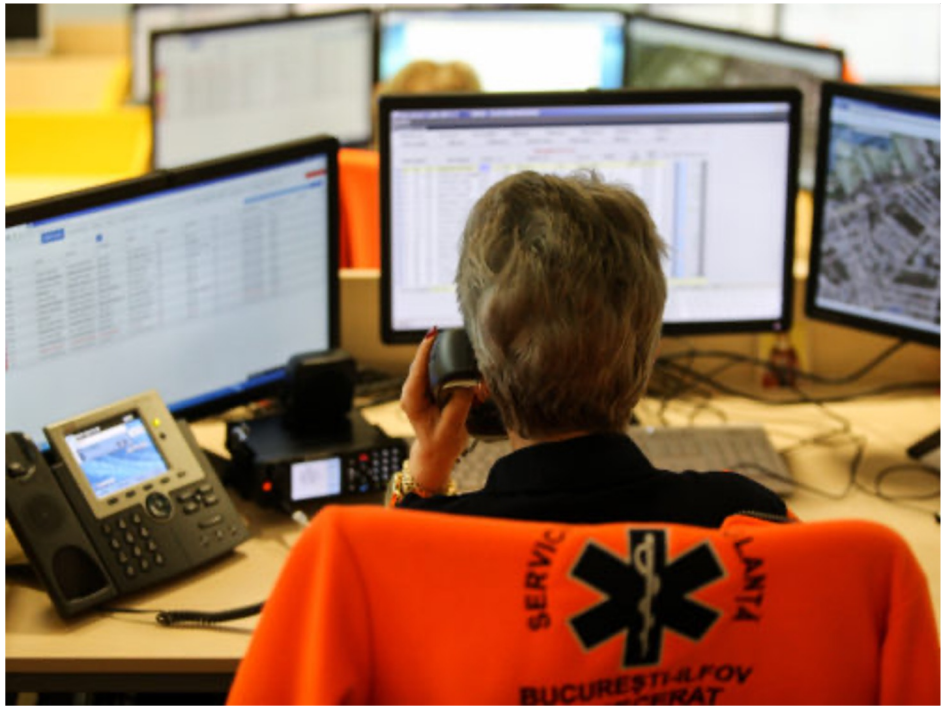


Photo: The People Lab

Lecture overview

I. What is impact?

II. Why randomize: A thought experiment

   I. **Non-experimental methods**

   II. Experimental method

III. When to randomize

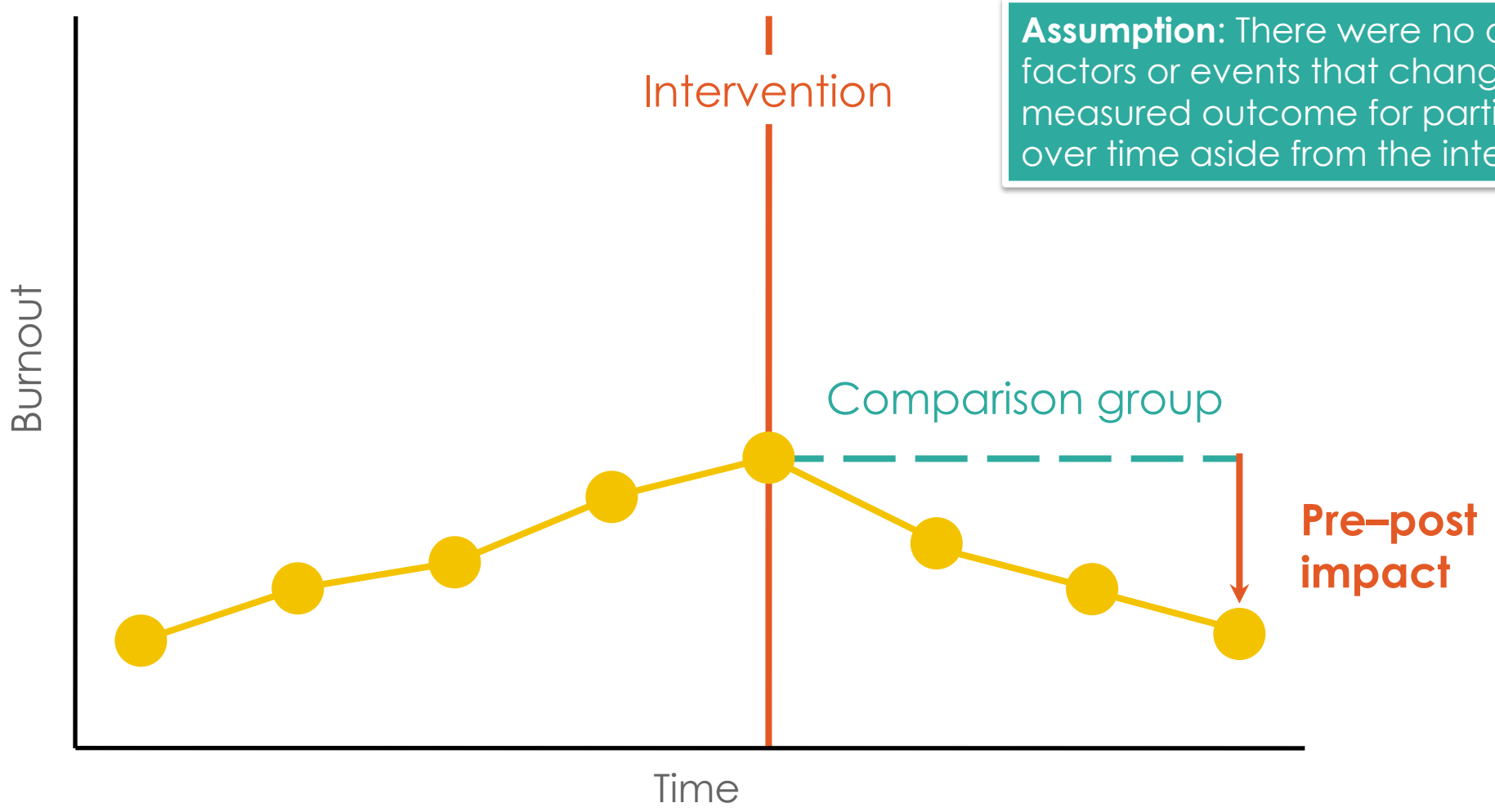
Non-experimental impact evaluation methods

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

1. Pre–post (before vs. after)
2. Simple difference
3. Matching
4. Difference-in-differences

Method 1: Pre–Post

Compare participants before and after the intervention



Assumption: There were no other factors or events that changed the measured outcome for participants over time aside from the intervention.

Is this a good comparison group to estimate the impact of the social support program? Are we confident that differences between the groups resulted from the program?

Probably not!

- Relies on the **very strong assumption** that burnout would not have changed over time in the absence of the program.
- Other things likely influence these outcomes over time.

Discussion question

Can you think of a scenario where you might use a pre-post method?

Method 2: Simple difference

Compare participants with non-participants

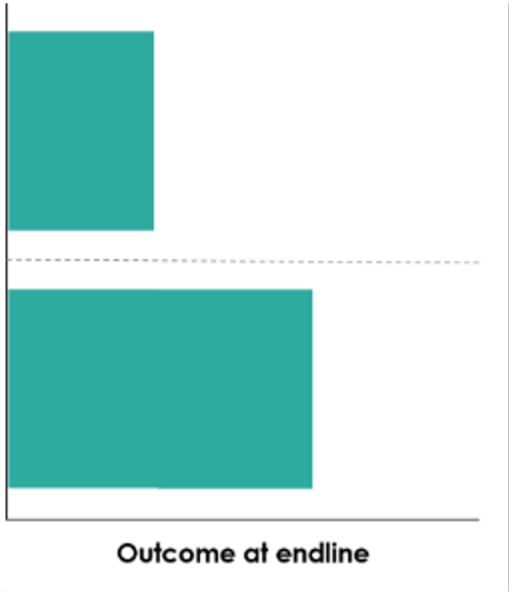
Frontline workers who participate in the program

Frontline workers who choose not to participate

Social support program

Continue under business as usual

Compare outcomes at the end of the program



Assumption: There are no differences between participants and non-participants except for program participation; the decision to participate in the program is unrelated to factors that affect outcomes.

Is this a good comparison group to estimate the impact of the social support program? Are we confident that differences between the groups resulted from the program?

Probably not!

- Individuals who participate might be better informed, more motivated, etc. (risk of "**selection bias**", since those who "select in" to a program may differ from those who do not in terms of their pre-program outcomes)
- Hard to disentangle whether the changes in outcomes are due to the program or due to some other aspects of the workers
    - E.g., potentially hard-to-capture underlying personality traits or other so-called "unobservable" factors

Discussion question

Can you think of a scenario where you might use a simple difference method?

Method 3: Matching

Try to identify similar individuals among non-participants

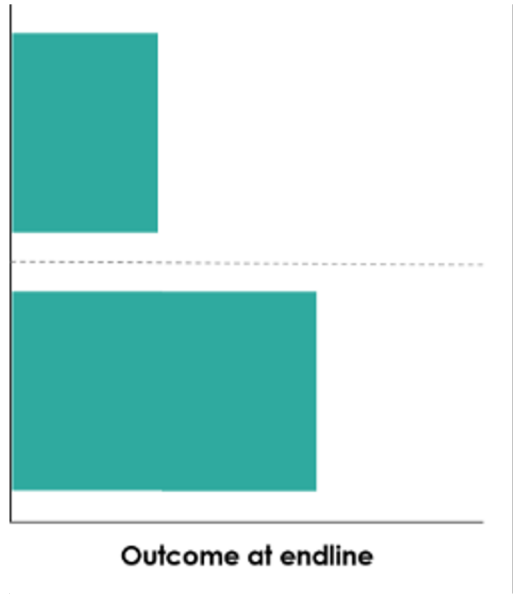
Participants in the program

Non-participant pool

Social support program

Continue under business as usual

Compare outcomes at the end of the program



Assumption: There are no differences between participants and non-participants with the same matching variables that affect the measured outcome.

Is this a good comparison group to estimate the impact of the social support program? Are we confident that differences between the groups resulted from the program?

Maybe, **if** we find individuals in the group of non-participants who:

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

Maybe not, **if** individuals in the group of non-participants:

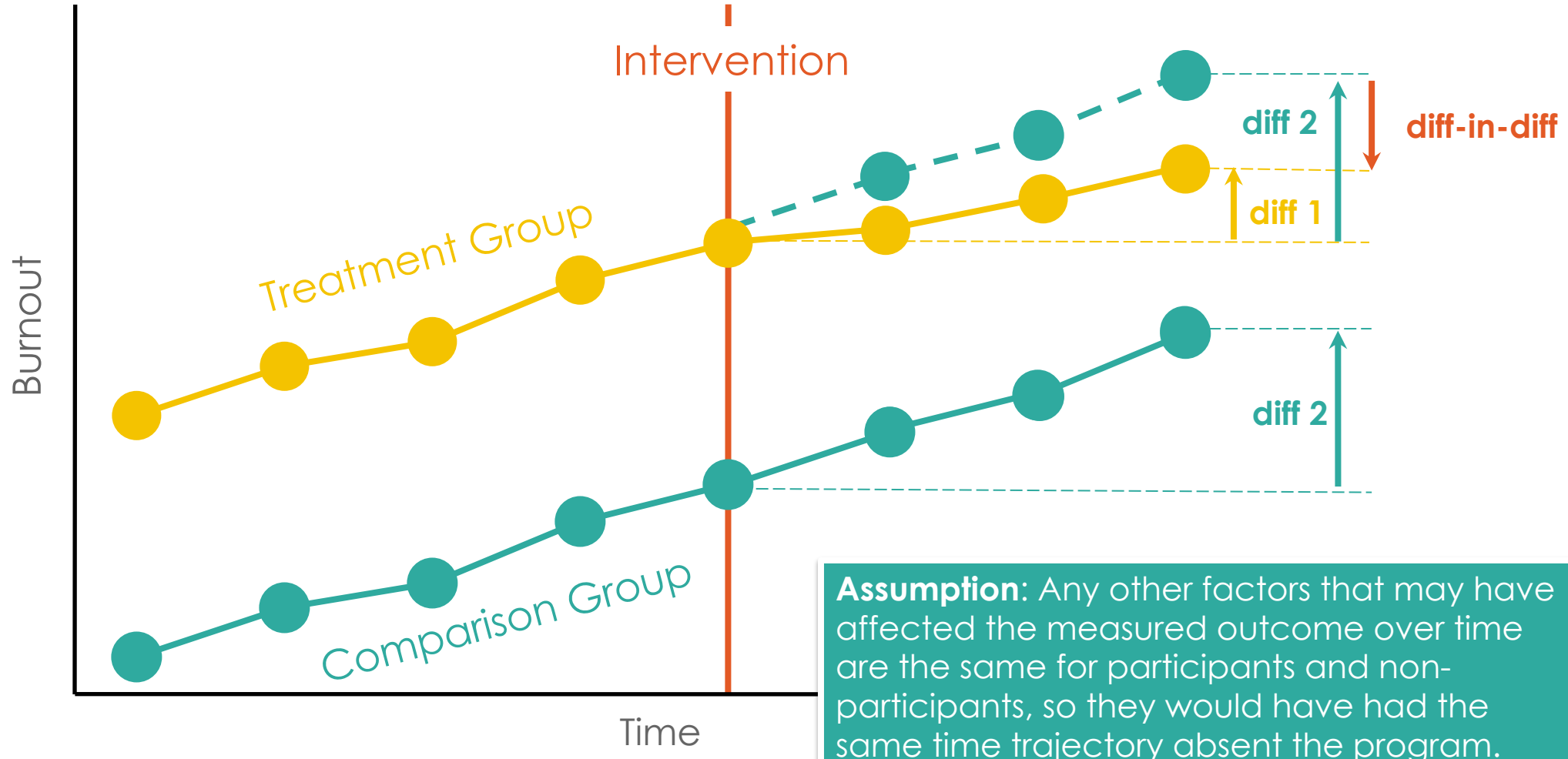
- Are not very similar to our participants across observable characteristics
- Are not similar across characteristics that we cannot observe (something we cannot test)

Discussion question

Can you think of a scenario where you might use a matching method?

Method 4: Difference-in-differences

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



Assumption: Any other factors that may have affected the measured outcome over time are the same for participants and non-participants, so they would have had the same time trajectory absent the program.

Is this a good comparison group to estimate the impact of the social support program? Are we confident that differences 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 affect both groups in the same way).

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).

Discussion question

Can you think of a scenario where you might use a difference-in-differences method?

Non-experimental methods rely on being able to "mimic" the counterfactual under certain assumptions

The non-experimental methods just discussed rely on assumptions that must hold to create a credible estimate of the 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?

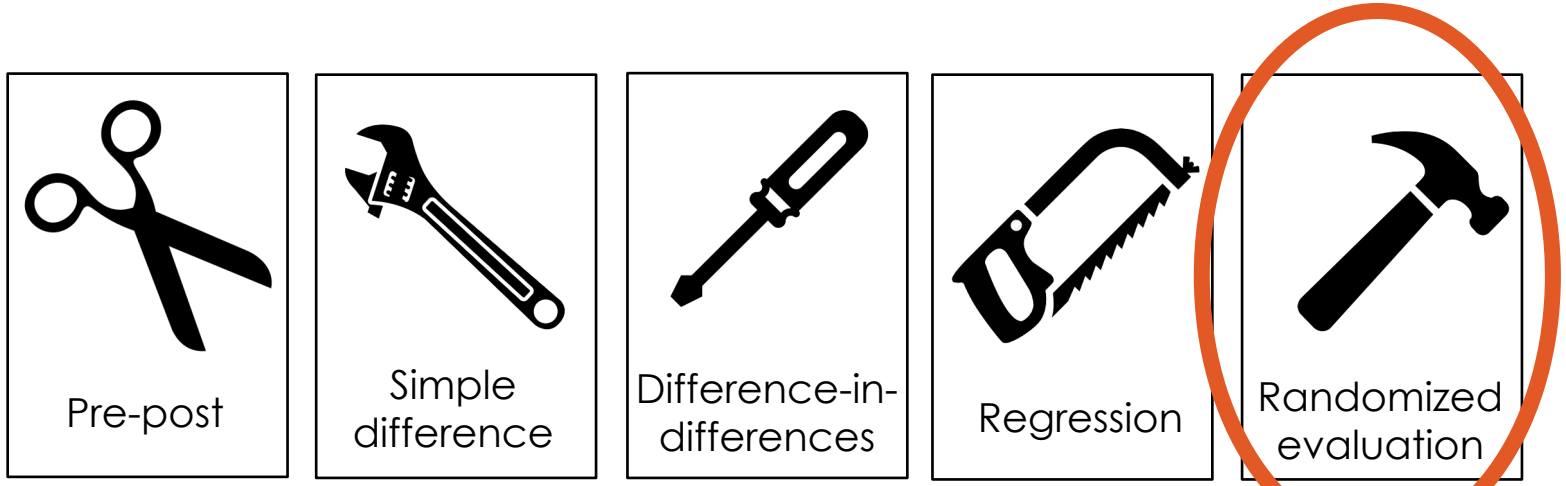
II. Why randomize: A case study

   I. Non-experimental methods

   II. Experimental method

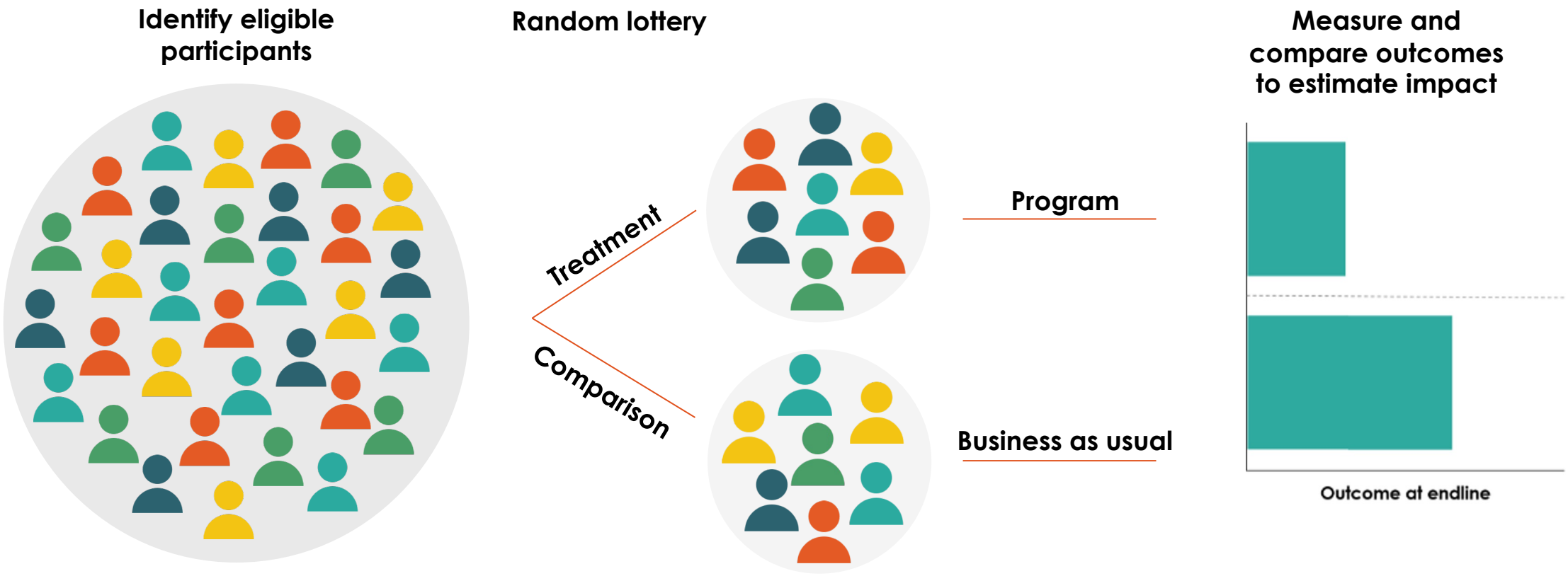
III. When to randomize

Methods as tools



Wrench by Daniel Garrett Hickey from the Noun Project
Screwdriver by Chameleon Design from the Noun Project

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



Assumption: Outcomes are only affected by program participation itself, not by assignment to participate in the program (or the evaluation).

Is this a good comparison group to estimate the impact of the social support program? 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.

Note: This course will cover how to design RCTs, common challenges and strategies to address these, as well as ethical considerations related to RCTs.

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, differences that subsequently arise between them can be attributed to the program, rather than to other factors.

Treatment

Source: freepik

Comparison

Key steps in conducting a randomized evaluation: 1/2

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

Note: We do not necessarily need baseline data it to estimate impact, though in practice this can be valuable for different reasons.

Key steps in conducting a randomized evaluation: 2/2

6. Collect endline data for both the treatment and control groups

7. Estimate program impacts by comparing mean outcomes of treatment group vs. mean outcomes of the control group

8. Assess whether program impacts are statistically significant and practically significant

Discussion question

What is a constraint to using a randomized evaluation to answer an impact evaluation question at your organization?

Lecture overview

I. What is impact?

II. Why randomize: A case study

   I. Non-experimental methods

   II. Experimental method

III. When to randomize

Considerations for when to do a randomized evaluation

- Is the program ready for an evaluation?
  - If it requires further tinkering, sort out implementation and process monitoring first
- Is there an element that can be randomized?
  - If the program has already been rolled out and you are not expanding elsewhere or considering program alterations, randomization may not be feasible or ethical
- Is there genuine uncertainty about the effectiveness and cost-effectiveness?
  - Is there a potential for unintended consequences we may not know about?
  - Do the anticipated benefits to participants outweigh the potential risks?
- Will the findings be credible? That is, is the study design credible?
  - Is the project on a large enough scale to randomize into "representative groups"?
  - Does the evaluation account for or measures spillovers, track compliance, etc.?
- Will the findings be actionable? Will they inform concrete policy decisions?

What do we need to consider when exploring a new randomized evaluation?

- What is the research question of interest? How will this inform learning and policy / program design and delivery going forward?
  - Ensure the goals of the implementing partner and those of the researchers align

- What are potential opportunities for random assignment?
  - Provide information to support partners in making informed decisions about their participation in a randomized evaluation and the proposed study design

- What time or resource constraints do you face?
  - Consider also the cost of not evaluating

Conclusion

- In order to measure a program's impact, we need to estimate the counterfactual as well as possible

- Non-experimental methods rely on being able to "mimic" the counterfactual under certain assumptions
  - Need to think critically about the credibility of these assumptions for a given evaluation

- If properly designed and conducted, randomized evaluations can provide a very credible method to estimate the impact of a program
  - Even so, randomized evaluations are just one of many tools and will not always the best option for a given scenario

References

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *The American Economic Review*, Volume 84, No. 4.

Delius, Antonia, and Olivier Sterck. 2020. "Cash Transfers and Micro-Enterprise Performance: Theory and Quasi-Experimental Evidence from Kenya." Working Paper.

Linos, Elizabeth, Krista Ruffini, and Stephanie Wilcoxen. 2022. "Reducing Burnout and Resignations among Frontline Workers: A Field Experiment." *Journal of Public Administration Research and Theory,* Volume 32, Issue 3.

Sanders, Michael, Elspeth Kirkman, Raj Chande, Michael Luca, Elizabeth Linos, and Xian-Zhi Soon. 2019. "Using Text Reminders to Increase Attendance and Attainment: Evidence from a Field Experiment." Working Paper.

Resources & further reading

- J-PAL Research Resource: Introduction to randomized evaluations
- J-PAL Research Resource: The elements of a randomized evaluation
- J-PAL Research Resource: Assessing viability and building relationships
- J-PAL's table of Impact Evaluation Methods
- J-PAL's Advantages of Randomized Evaluations
- J-PAL's Common Questions and Concerns about Randomized Evaluations
- World Bank blog post: Are we over-investing in baselines?

Reuse and citation

To reference this lecture, please cite as:

J-PAL. "Lecture: Why Randomize." Abdul Latif Jameel Poverty Action Lab. 2023. Cambridge, MA