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. Randomized Evaluation from Start to Finish
7. Threats & Analysis
8. Ethical Considerations
9. Generalizing & Applying Evidence

Learning objectives

- Learn what causal 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

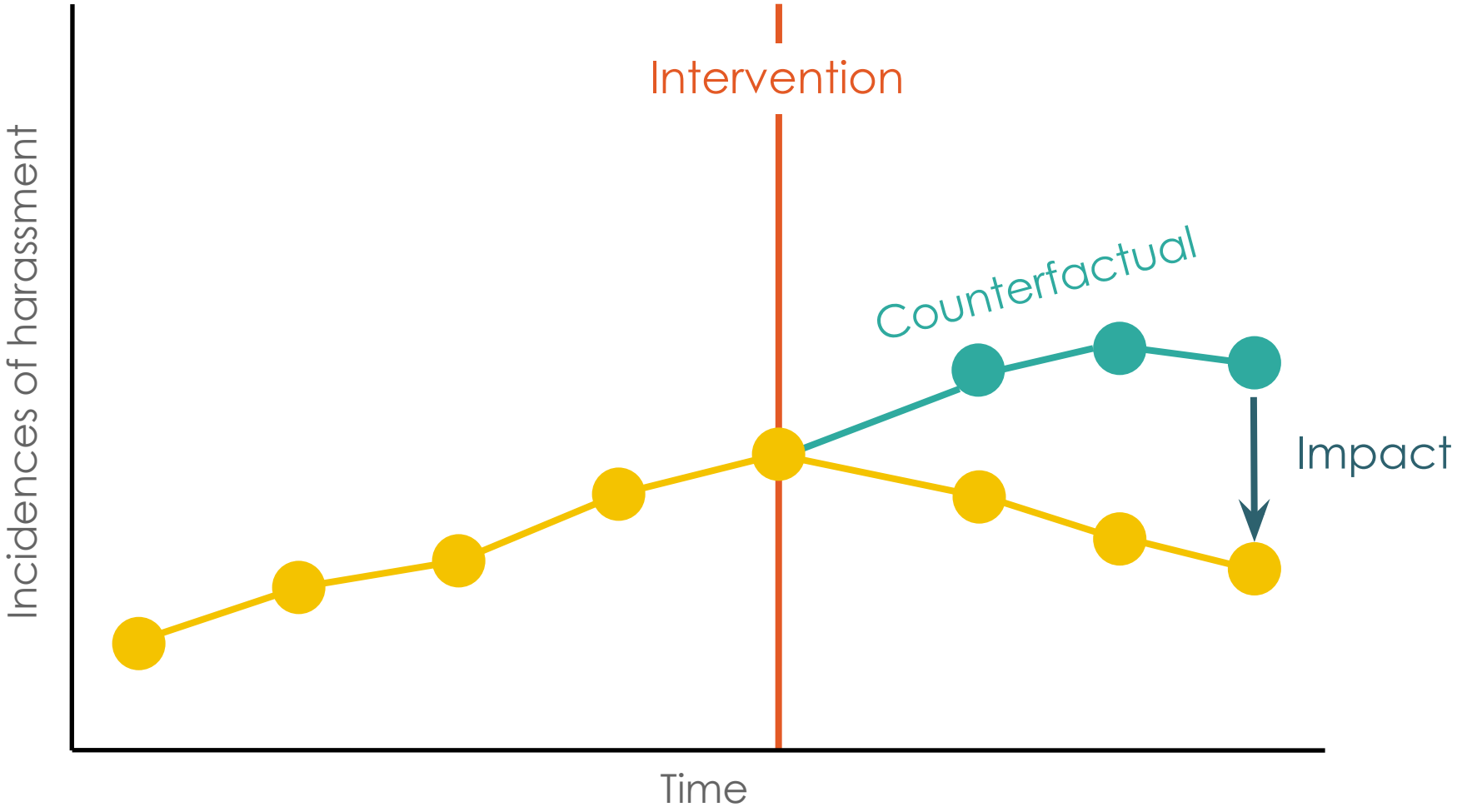
III. When to randomize

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")

Example: What is the impact of increased police patrols on violence against women in public spaces?



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

Impact evaluation methods

- Impact evaluation methods answer *cause-and-effect* questions: What is the effect of [program, policy, intervention] on [outcomes]?
- Different impact evaluation methods use different comparison groups to estimate the counterfactual

Non-experimental methods

- Pre-post comparison
- Simple difference
- Statistical matching
- Difference-in-differences
- Triple difference
- …and others (e.g. RDD)

Experimental method

- Randomized evaluations

The impact evaluation method we choose matters!

- Different impact evaluation methods may be more or less appropriate under different circumstances
  - Different methods can yield very different estimates of causal impact
- As we will see, each method relies on different 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 effects compare under different delivery mechanisms?
  - How to accurately target beneficiaries? How to increase program take-up?
- How cost-effective is a program?
  - How does it compare to other programs designed to accomplish similar goals?

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

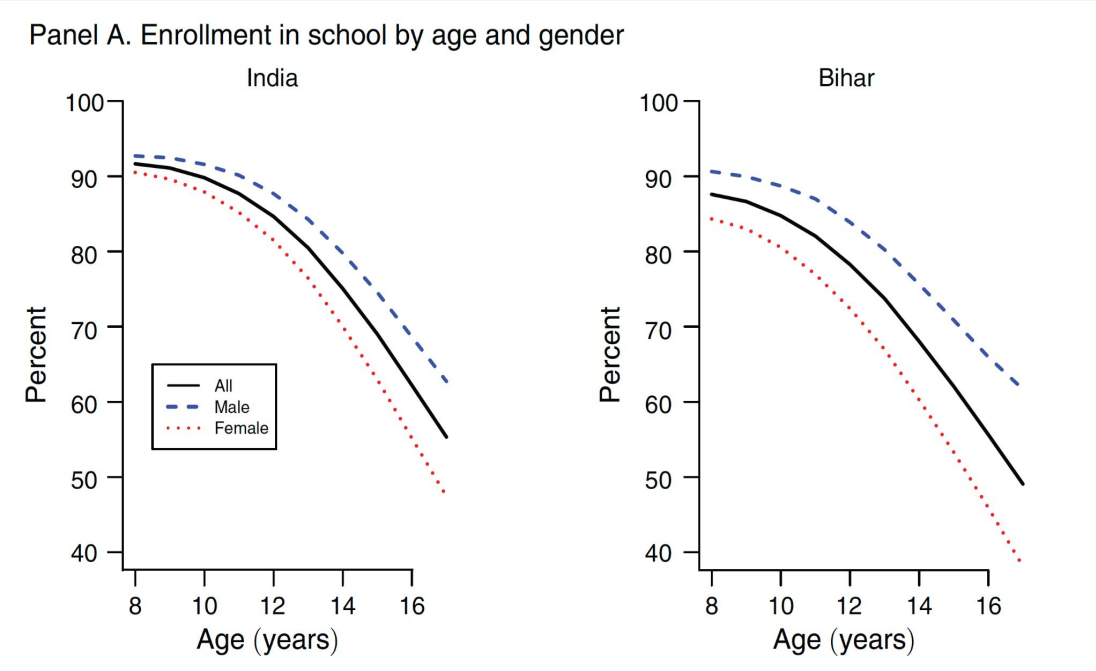
  II. Experimental method

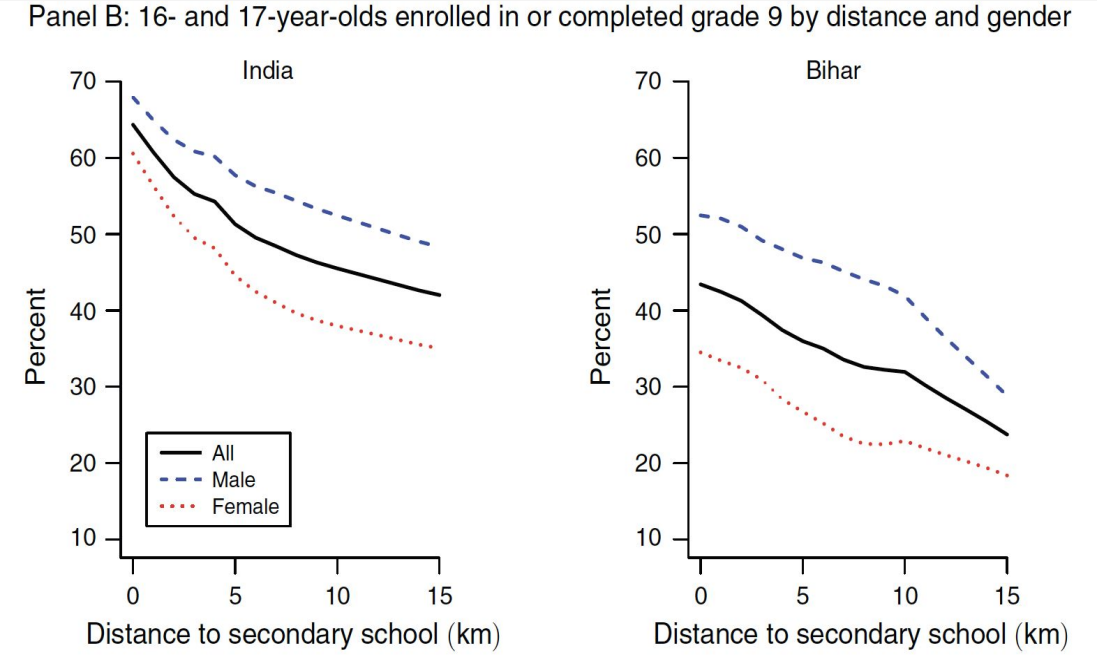
III. When to randomize

Providing bicycles to girls to increase school enrollment

- **Motivation:** Reducing gender gaps in school enrollment has been a global policy priority over the past decade.
- **Challenge:** Girls in Bihar, India enroll in school at lower rates than boys, and the gap increases around ages 14. While 90% of villages in Bihar have a primary school, fewer than 12% have a secondary school, making transit a salient barrier for enrollment (DLHS, 2008).
- **Intervention:** The Government of Bihar provided girls entering secondary school in 2006 with a one-time transfer to purchase a bicycle to reduce the daily cost of travelling to school.

Setting up the context: What does the data say?





Thought experiment: Designing an impact evaluation

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

Does providing free bicycles increase girls' enrollment in secondary school?

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

Study: "Cycling to School: Increasing Secondary School Enrollment for Girls in India" (Muralidharan and Prakash, 2017)

Photo: Associated Press

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

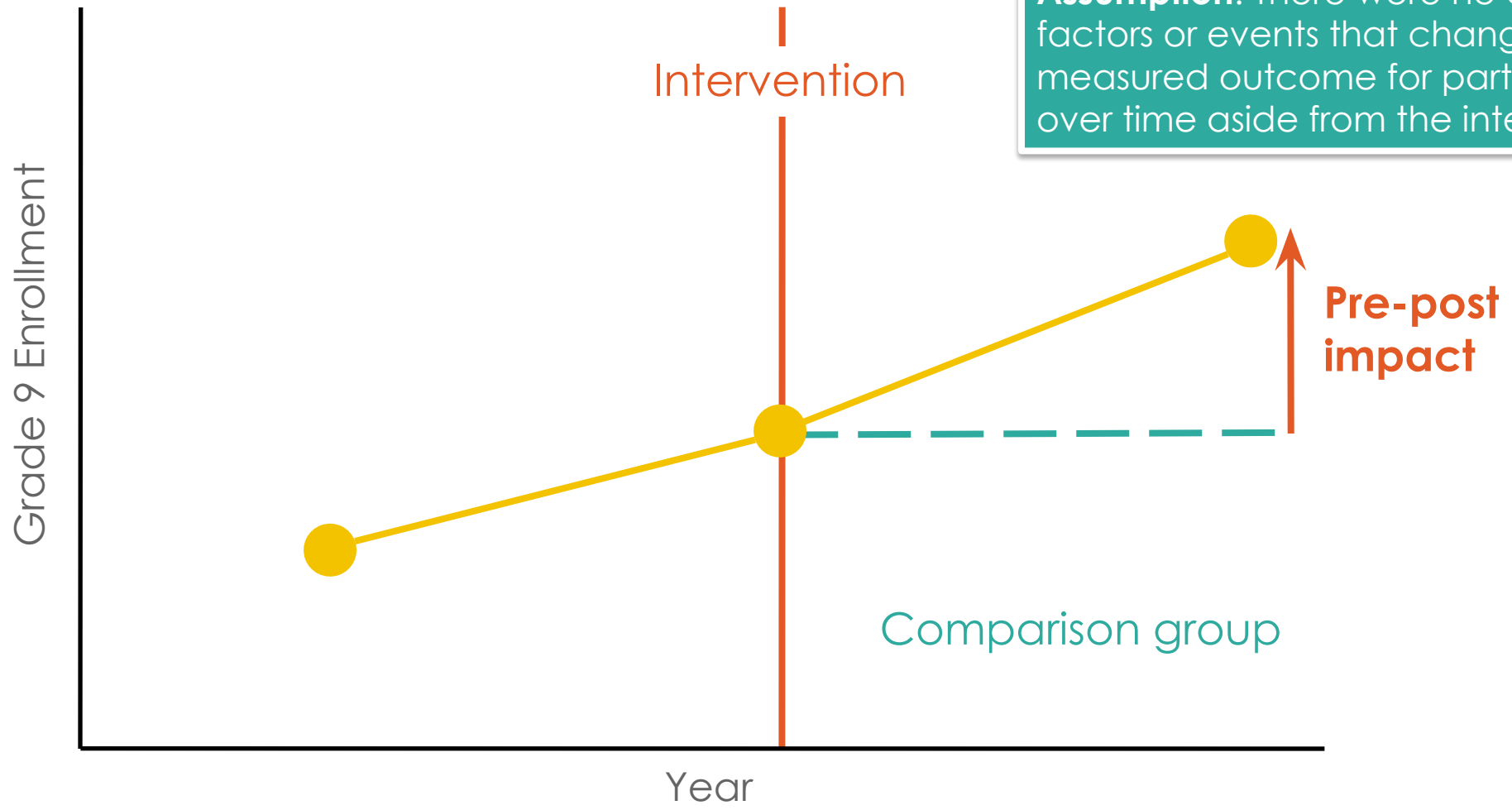
Let's look at different non-experimental methods of estimating the impact of providing free bicycles on girls' secondary school enrollment:

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

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 bicycle program? Are we confident that differences between the groups resulted from the program?

Probably not!

- Relies on the **very strong assumption** that secondary school enrollment would not have changed over time in the absence of the program.
- Other things likely influence these outcomes over time (improvements in law and order, roads, etc.)

Method 2: Simple difference

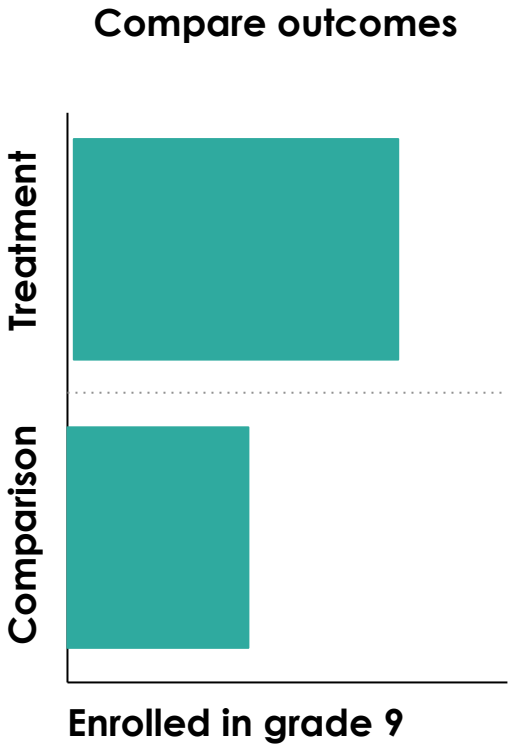
Compare participants with non-participants

Girls in cohort 2 (eligible for the program)

Girls in cohort 1 (not eligible for the program)

Bicycle program

Continue under business as usual



Assumption: There are no differences between participants and non-participants except for program participation; the year that girls began secondary school is unrelated to factors that affect outcomes.

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

Probably not!

- Hard to disentangle whether the changes in outcomes are due to the program or to some other circumstances that differed between cohorts (e.g., changes in public spending on education)

- Among eligible girls, those who choose to take up the program may be more motivated (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)

Method 3: Statistical Matching

Try to identify similar individuals among non-participants

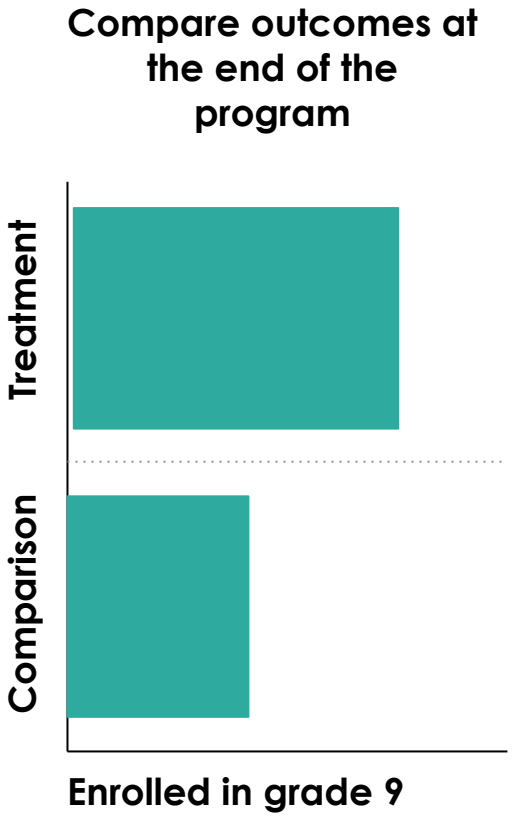
Participants in the program

Non-participant pool

Bicycle program

Continue under business as usual

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 bicycle program? Are we confident that differences between the groups resulted from the program?

Maybe, if individuals in the group of non-participants:

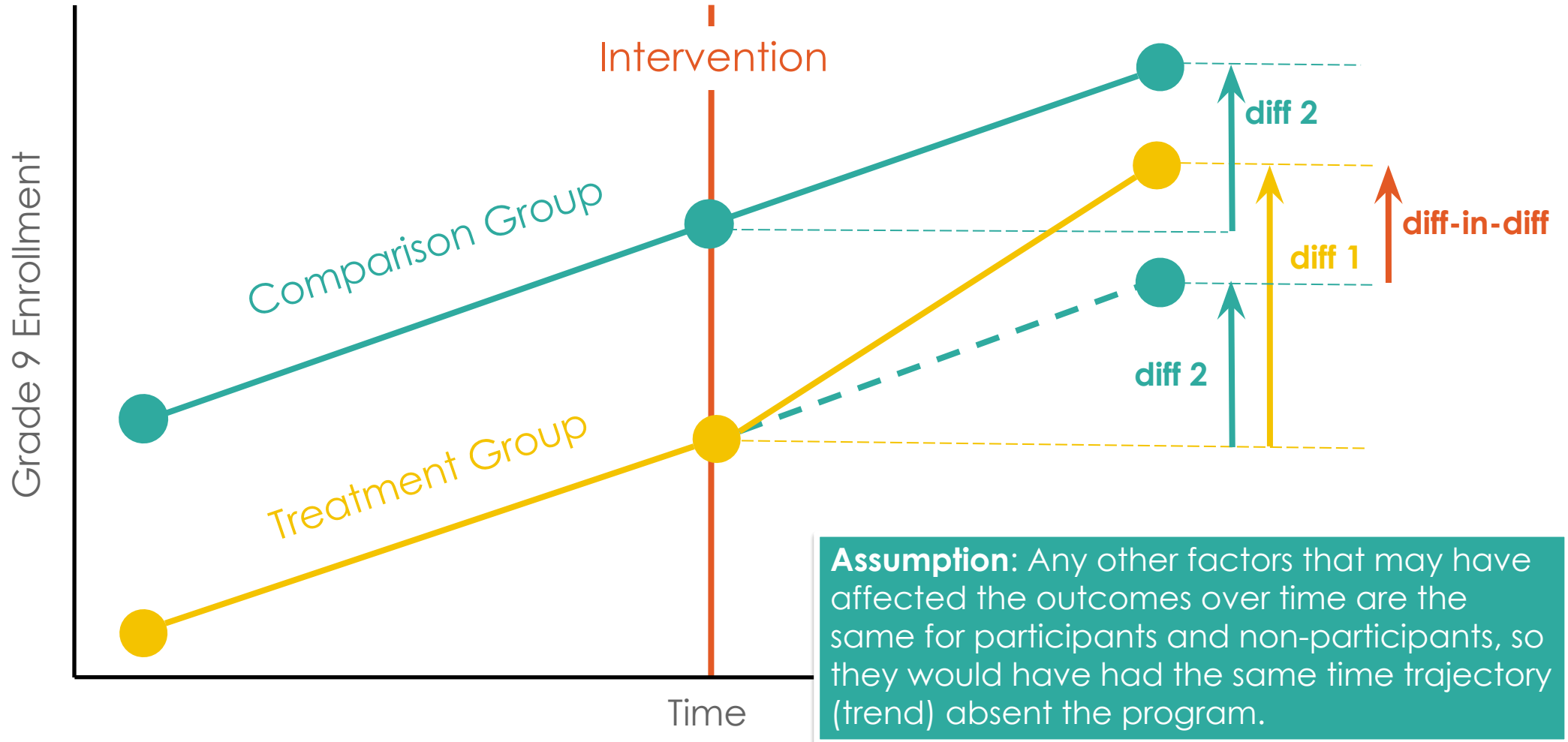
- Are very similar to our participants across observable characteristics (something we can verify)
- Are also similar across so-called non-observable characteristics (something we cannot test)

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)

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 outcomes over time are the same for participants and non-participants, so they would have had the same time trajectory (trend) absent the program.

Is this a good comparison group to estimate the impact of the bicycle 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 outcomes 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).

Estimating the impact of the bicycle program with triple differences method

Simple difference (across cohort)	**Difference in differences** (across cohort & gender)	**Triple difference** (across cohort, gender, & state)
Compare girls who are not eligible for the program (cohort 1) vs. girls who are eligible (cohort 2) *However, other changes may impact enrollment across cohorts.*	Compare girls across cohorts 1 and 2 vs. boys across cohorts 1 and 2 *However, changes in enrollment over time may vary by gender.*	Compare girls across cohorts and boys across cohorts in Bihar vs. girls across cohorts and boys across cohorts in Jharkhand *We still rely on assumptions (i.e., parallel trends across states), but believe these are more likely to hold in this case.*

Estimating the impact of the bicycle program with triple differences method

Source: International Growth Centre

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 thought experiment

   I. Non-experimental methods

   II. Experimental method

III. When to randomize

Methods as tools

Pre-post

Simple difference

Difference-in-differences

Regression

Randomized evaluation

Observational; retrospective; non-experimental

Experimental; prospective

Where J-PAL focuses

Images: the Noun Project

Example: Wheels of Change evaluation in Zambia

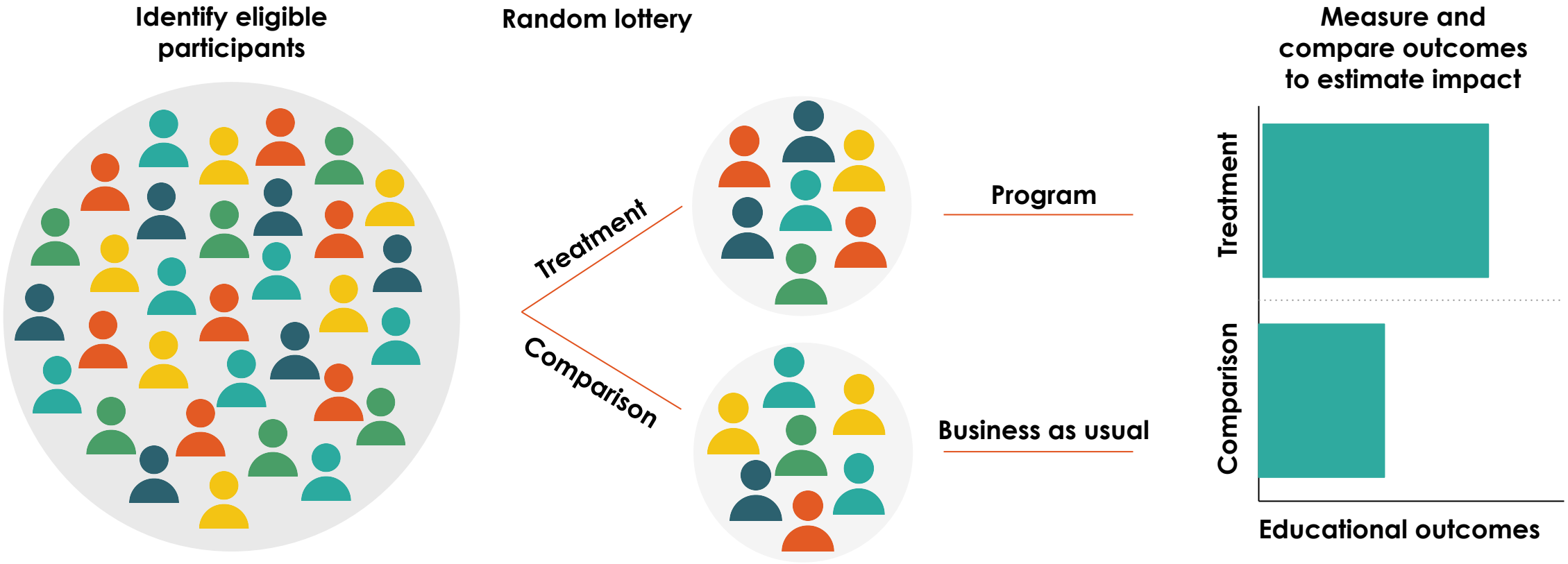
Photo: World Bicycle Relief

- We just saw how triple differences was used to estimate the impact of providing bicycles to girls in secondary school in India
- Now, let's look at a similar program in Zambia
  - Girls' enrollment drops more than boys' in secondary school
  - Average travel time to school is 110 minutes
  - 35% of girls experience sexual harassment during their commute

Why was an RCT the best choice in this context?

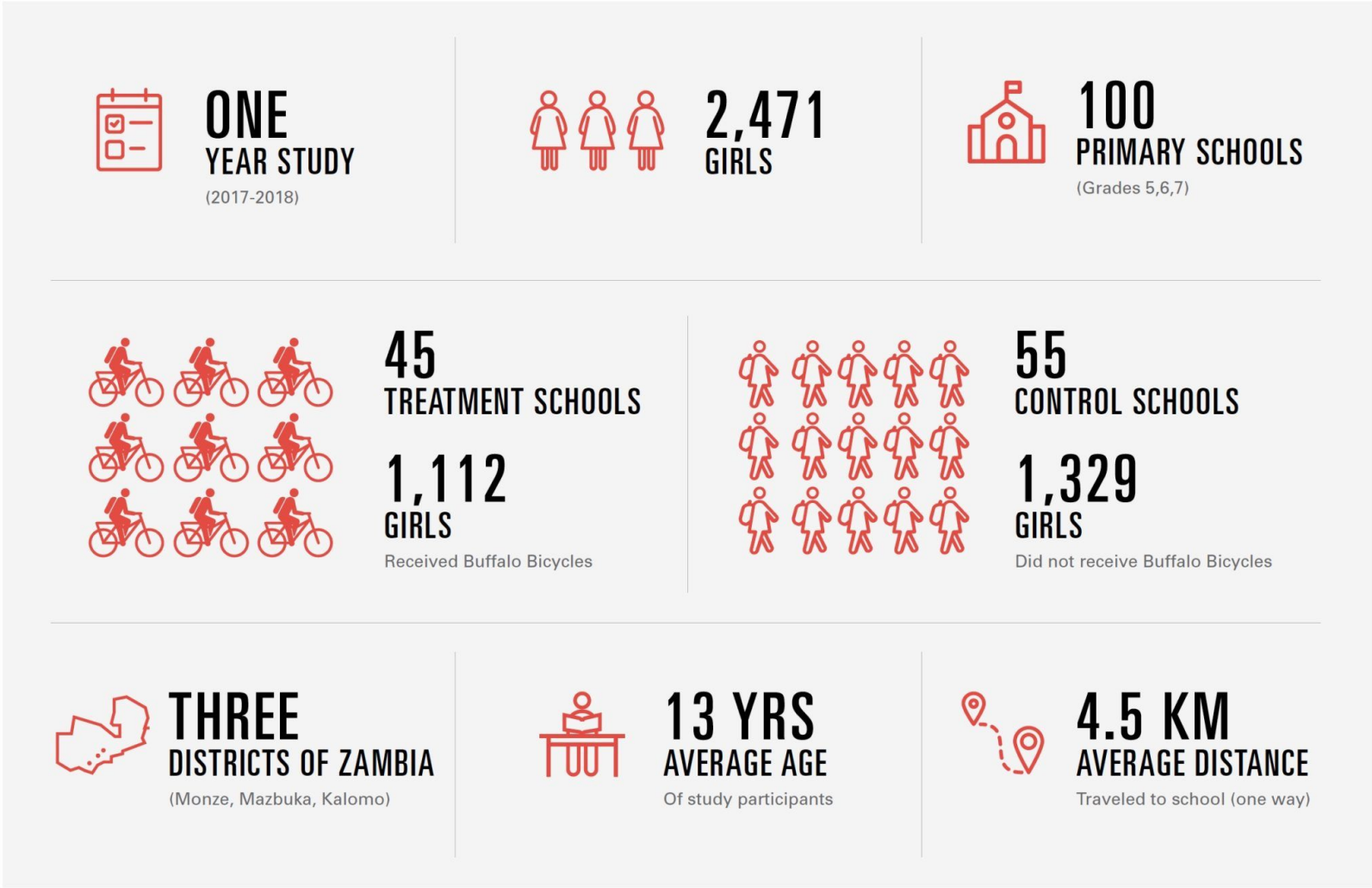
Study: "Wheels of change: Transforming girls' lives with bicycles" (Fiala et al. 2022)

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).

Study methodology



Source: World Bicycle Relief

Evaluation findings

Positive impacts on educational outcomes, safety, and empowerment for participating girls led to program scale up in four countries.

- Attendance: reduced absenteeism by 28% (5 days more school per year)
- Time use: reduced commute time by over one hour per day (33%)
- Safety: 22% less likely to be verbally harassed on commute
- Empowerment: positive impacts on feelings of empowerment
  - Follow up to look at how girls' actual decisions compare to self-reporting
  - Observed increased rates of early marriages and pregnancies among girls in the treatment group (not what we had anticipated) => made them attractive in the marriage market
  - Also observed decreased levels of domestic violence among girls who had received a bicycle => still an empowerment story!

Are girls in randomly assigned control schools a good comparison group to estimate the impact of the bicycle 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. Design the study carefully, considering possible challenges
2. Create a list of units to be randomized [and collect baseline data]
3. Randomly assign units to treatment or control and verify this process
4. Monitor processes to ensure the integrity of the evaluation
5. Collect endline data for both the treatment and control groups
6. Estimate impacts by comparing average outcomes across groups
7. Assess whether impacts are statistically and practically significant

Lecture overview

I. What is impact?

II. Why randomize: A thought experiment

    I. Non-experimental methods

    II. Experimental method

III. When to randomize

When to consider a randomized evaluation

- Is the program ready for an evaluation?
  - If it requires further tinkering, first focus on piloting, implementation, and process monitoring
- Is there an element that can be randomized?
  - If the program has already been rolled out and you are not expanding elsewhere or considering changes to the program, randomization may not be feasible or ethical
- Is there genuine uncertainty about effectiveness and cost-effectiveness?
  - Do the anticipated benefits from the evaluation outweigh the potential risks?
- Will the findings be credible and actionable?
  - Is the study design credible? Is the project on a large enough scale to randomize?
  - Will the findings inform concrete policy decisions? What is the timeline for influencing policy?

Considerations for exploring a randomized evaluation

- What is the research question of interest? How will this inform learning and policy or program design and delivery going forward?
  - Align goals across the research participants/community, implementing partner, and researchers

- 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? What is the tradeoff with the cost of not evaluating?
  - How will this contribute to recalibration and the cumulative evidence base?

Discussion question

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

Conclusion

- Each impact evaluation method relies on being able to "mimic" the counterfactual under certain assumptions
  - Need to think critically about the credibility of these assumptions for a given evaluation to estimate a program's impact as accurately as possible

- 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 are not always the best option for a given scenario

Appendix

In what scenario might you consider using a given non-experimental method?

Method	Key Assumption	Example
Pre-post	No other factors changed the measured outcome over time aside from the intervention	Comparing pre- and post-course assessments to measure learning from a one-week impact evaluation training
Simple difference	No differences in outcomes between groups aside from program participation	Measuring the impact of winning the lottery on income–we can assume lottery players are similar *on average*
Statistical matching	No differences in outcome between individuals who are similar on matching variables	Comparing businesses with similar characteristics who did vs. did not apply for a cash transfer program
Difference in differences	The outcome between groups would have followed parallel trends without the program	Comparing wages before and after a minimum wage law in a state that implemented the law vs. one that did not

References

Amaral, Sofia, Girija Borker, Nathan Fiala, Anjani Kumar, Nishith Prakash, and Maria Micaela Sviatschi. 2023. "Sexual Harassment in Public Spaces and Police Patrols: Experimental Evidence from Urban India." National Bureau of Economic Research Working Paper no. w31734.

Fiala, Nathan, Ana Garcia-Hernandez, Kritika Narula, and Nishith Prakash. 2022. "Wheels of change: Transforming girls' lives with bicycles." CESifo Working Paper no. 9865.

J-PAL. 2020. "Street Police Patrolling To Reduce Crime Against Women in Public Spaces in India." Evaluation Summary.

Muralidharan, Karthik, and Nishith Prakash. 2017. "Cycling to school: Increasing secondary school enrollment for girls in India." *American Economic Journal: Applied Economics* 9, no. 3: 321-350.

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. 2024. "Lecture: Why & When to Randomize." Abdul Latif Jameel Poverty Action Lab. Cambridge, MA