Why Randomize?

Dr. Sharon Barnhardt

CESS Nuffield – FLAME University

4 July 2017

Course Overview

1. What is Evaluation?
2. Theory of Change
3. Outcome, Impact, and Indicators
4. **Why Randomize?**
5. How to Randomize
6. Sampling and Sample Size
7. Threats and Analysis
8. Research to Policy
9. Project from Start to Finish

Session Overview

I.   Basic vocabulary for impact evaluation

II.   Randomized evaluation

III. Other methods of impact evaluation

IV.   Conclusions

Components of Programme Evaluation

- Needs Assessment

- Programme Theory Assessment

- Process Evaluation

- **Impact Evaluation**

- Cost Effectiveness

- What is the problem?

- How, in theory, does the Programme fix the problem?

- Does the Programme work as planned?

- **Were its goals achieved? The magnitude?**

- Given magnitude and cost, how does it compare to alternatives?

BASIC VOCABULARY FOR IMPACT EVALUATION

# Example: Immunization Incentives

- **The Problem:**
  - Despite availability of free immunization, full coverage rates among children remains extremely low in many developing countries

- **Intervention**
  - Reliable, monthly immunization camps set up in villages in Udaipur
  - Small incentives offered to mothers conditional on having child immunized; larger incentive when immunization course completed

Which one of these would make a good question for impact evaluation?

A. What percentage of 3 year old children in Rajasthan were not fully immunized?

B. What is the correlation between regular immunization camps and immunization rates?

C. Does holding regular immunization camps and providing incentives to parents improve immunization rates of children?

81%

8%

**12%**

A.

B.

C.
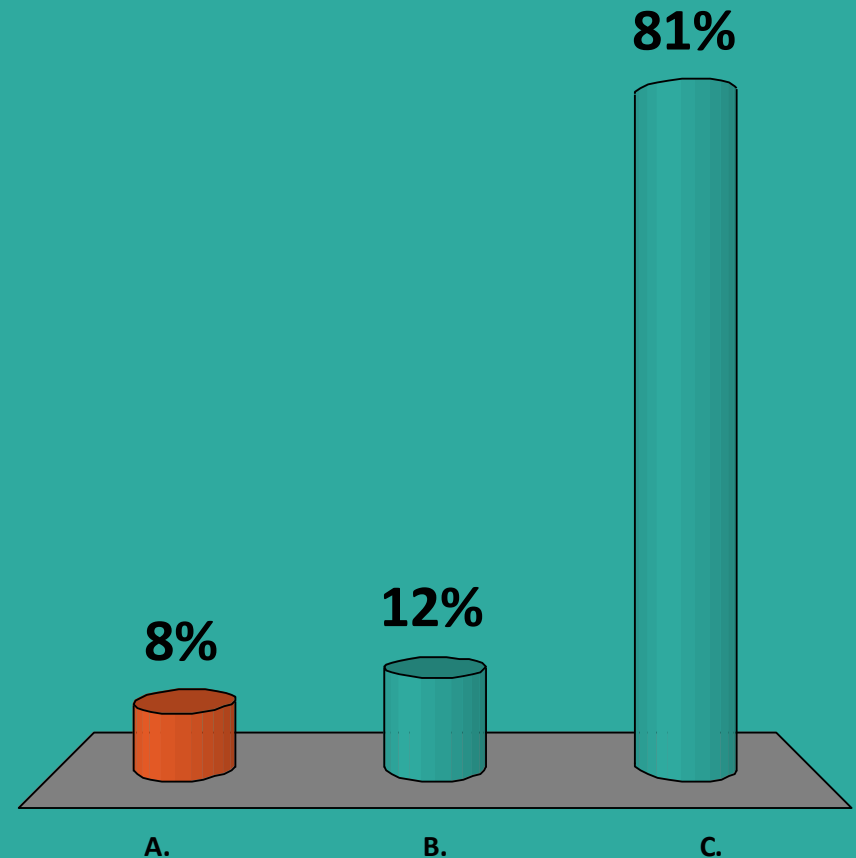
Causal Inference

Cause and effect language is used everyday in a lot of contexts, but it means something very specific in impact evaluation.

- We can think of causality as:
    - The singular effect of a program on an outcome of interest
    - Independent of any other intervening factors,

- Our goal is to estimate the size of this effect accurately and with confidence
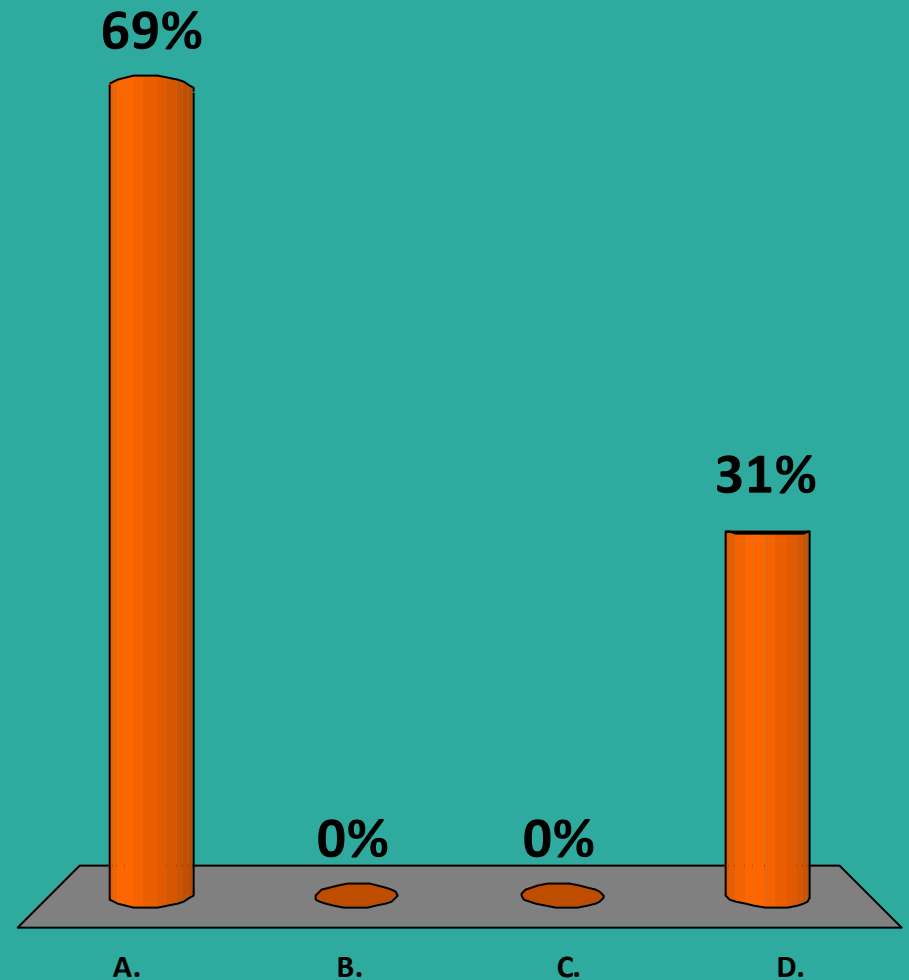
How to measure impact?

- *Impact* (also called "causal effect") is defined as a comparison between:

  1. The outcome some time after the program has been introduced

  2. The outcome at that same point in time had the program not been introduced (the *"counterfactual"*)
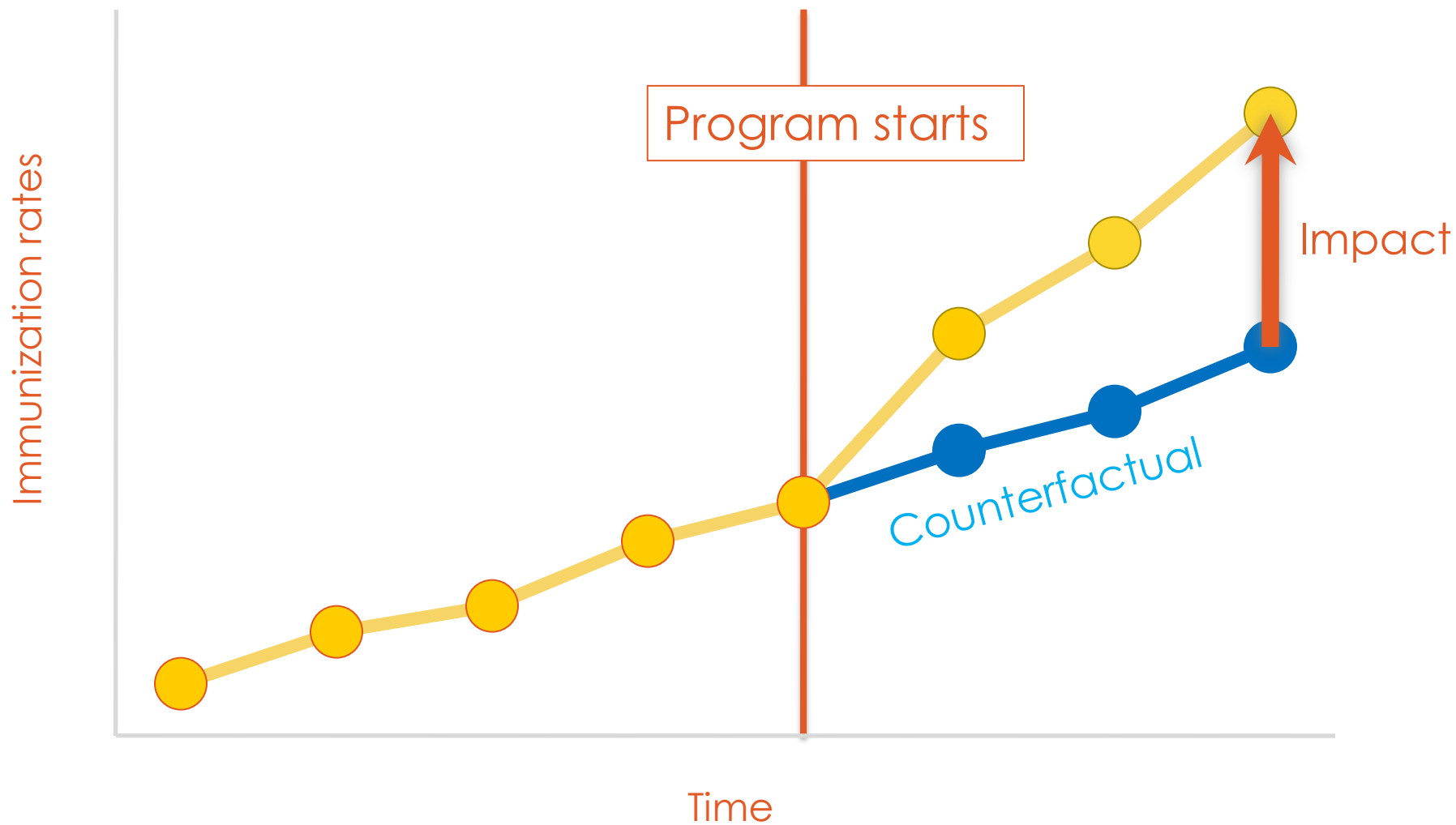
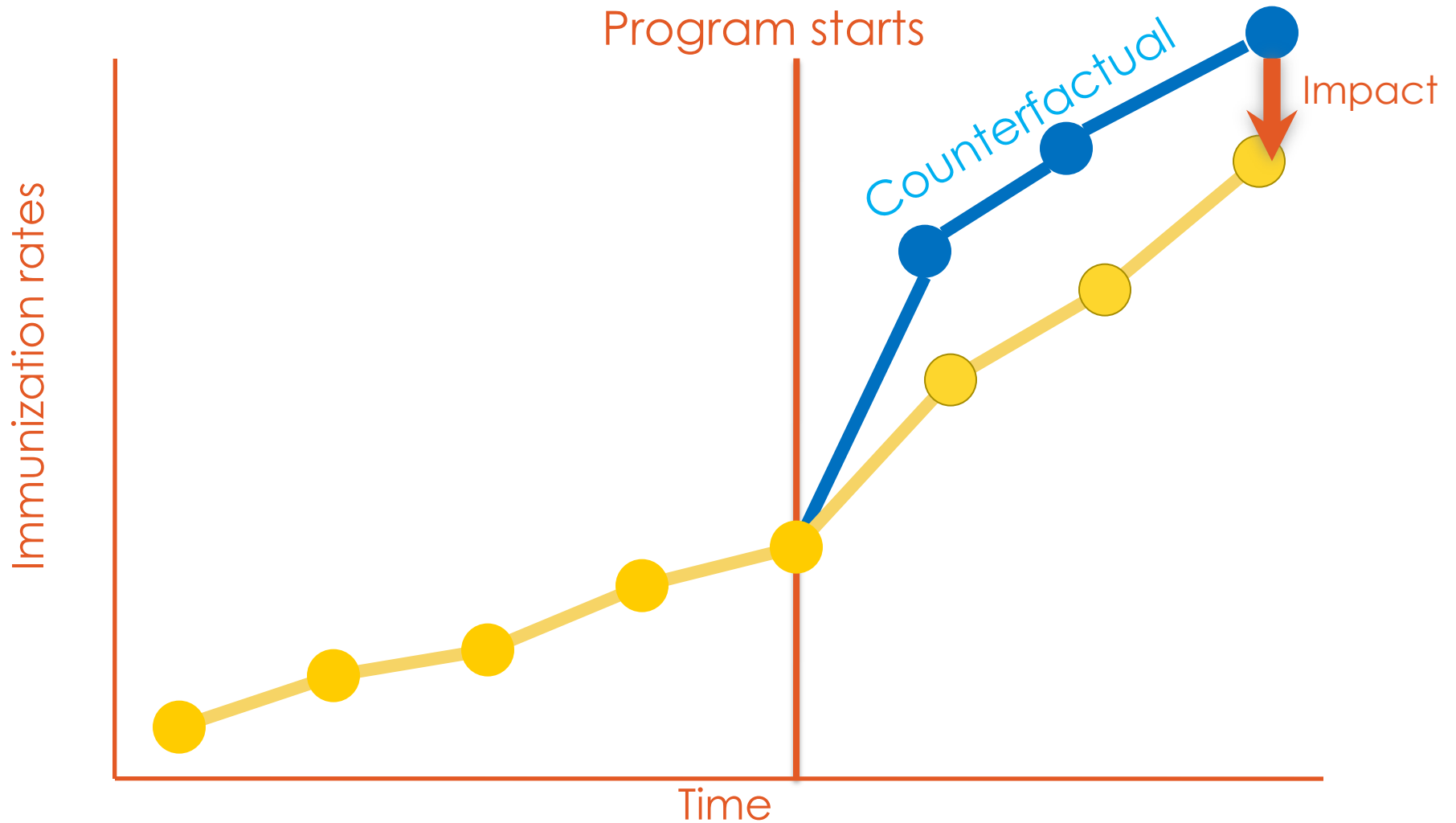# What is the impact of this program?

Immunization rates

Program starts

Time
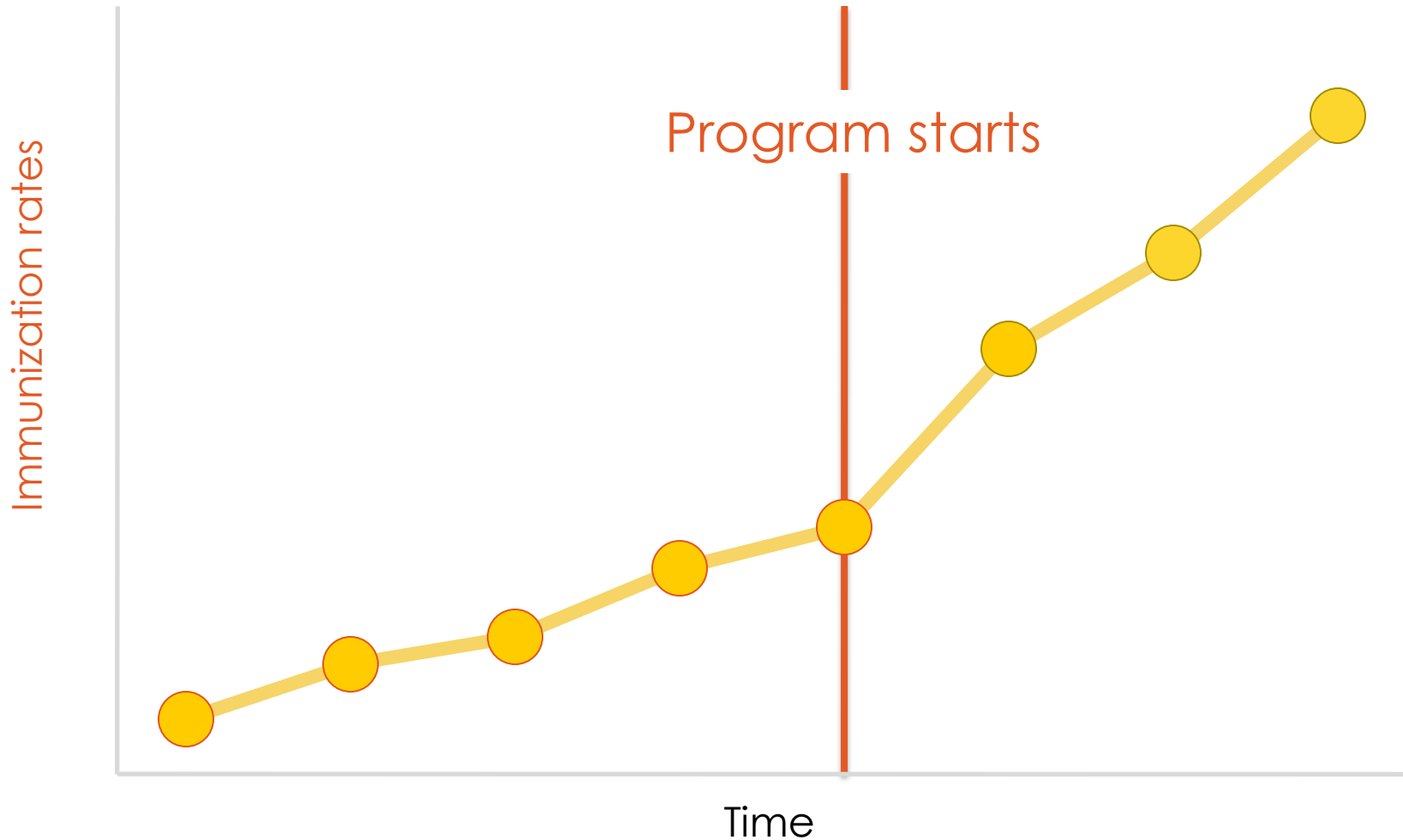
What is the impact of this program?

A. Positive

B. Negative

C. Zero

D. Not enough info

69% A.

0% B.

0% C.

31% D.

What is the impact of this program?

Program starts

Immunization rates

Impact

Counterfactual

Time

Impact: What is it?

Program starts

Counterfactual

Impact

Immunization rates

Time

Impact: What is it?
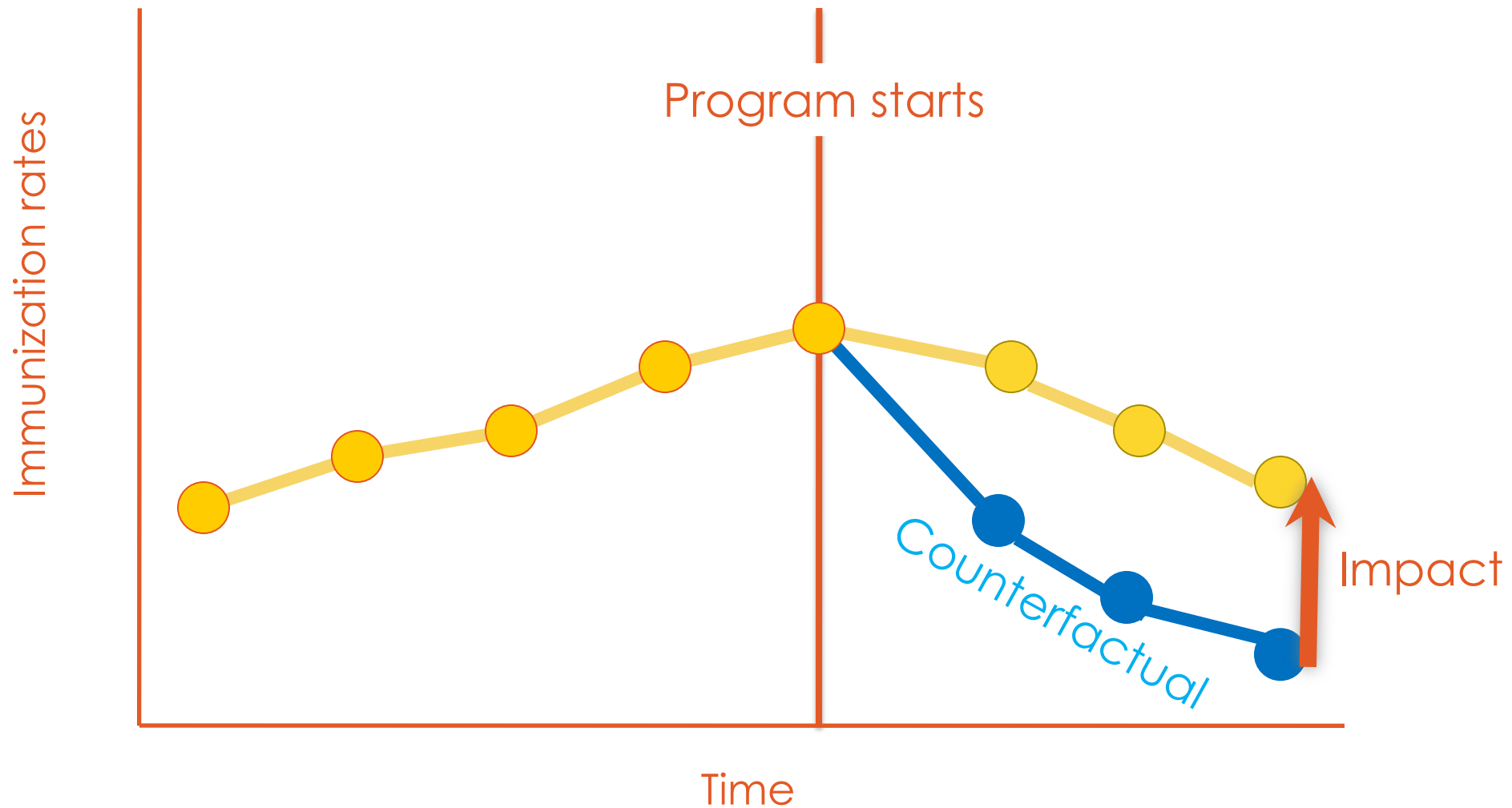
Counterfactual

The **counterfactual** represents the state of the world that program participants would have experienced in the absence of the program (i.e. had they not participated in the program)

Problem: Counterfactual cannot be observed

Solution: We need to "mimic" or construct the counterfactual

Constructing the counterfactual

- 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

Selecting the comparison group

- Idea: Select a group that is exactly like the group of participants in all ways except one: their exposure to the program being evaluated

- Goal: To be able to attribute differences in outcomes between the group of participants and the comparison group to the program (and not to other factors)

- An impact evaluation is only as good as the comparison group it uses to mimic the counterfactual

Impact evaluation methods

1. Randomized Experiments

Use random assignment of the program to create a comparison group which mimics the counterfactual.

Also known as:

- Random Assignment Studies
- Randomized Field Trials
- Social Experiments
- Randomized Controlled Trials (RCTs)
- Randomized Controlled Experiments

Impact evaluation methods

2. Non- or Quasi-Experimental Methods

Argue that a certain excluded group mimics the counterfactual

a. Pre-Post

b. Simple Difference

c. Differences-in-Differences

d. Multivariate Regression

e. Statistical Matching

f. Interrupted Time Series

g. Instrumental Variables

h. Regression Discontinuity

Example: Balsakhi Program

Balsakhi Program: Background

- Problem:
  - Many children in 3rd and 4th standard were not even at the 1st standard level of competency
  - Class sizes were large
  - Social distance between teacher and many of the students was large

- Proposed solution:
  - Hire local women (balsakhis) from the community and train them to teach basic competencies (reading, numeracy) to lowest performing students
  - Implemented by Pratham, an NGO from India
  - In Vadodara, the balsakhi program was run in government primary schools in 2002-2003
  - Teachers decided which children would get the balsakhi

Balsakhi: Outcomes

- Children were tested at the beginning of the school year (Pretest) and at the end of the year (Post-test)

- QUESTION: How can we estimate the impact of the balsakhi program on test scores?

RANDOMIZED EVALUATIONS

Randomized Evaluation

- Suppose we evaluated the balsakhi program using a randomized evaluation

- QUESTION #1: What would this entail? How would we do it?

- QUESTION #2: What would be the advantage of using this method to evaluate the impact of the balsakhi program?

The basics

- Take a sample of program applicants

- Randomly assign them to either:
  - Treatment Group – is offered the program
  - Control Group – not allowed to receive the program (during the evaluation period)

- The two groups will, on average, have the same observable and unobservable characteristics
  - since assignment is purely by chance
  - provided we have a large enough number of units

- Impact = Difference in outcomes between the treatment and control groups after the program

Treatment and control before the program: **Balsakhi -Vadodara**

		Treatment	Control	Difference
Year 1 (grades 3 & 4)	Math	-0.007	0.000	-0.007 (0.059)
	Language	0.025	0.000	0.025 (0.061)
Year 2 (grades 3 & 4)	Math	0.046	0.000	0.046 (0.053)
	Language	0.055	0.000	0.055 (0.058)

Key advantage of experiments

Because members of the groups (treatment and control) do not differ systematically at the outset of the experiment,

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

If properly designed and conducted, randomized experiments provide the most credible method to estimate the impact of a program

28

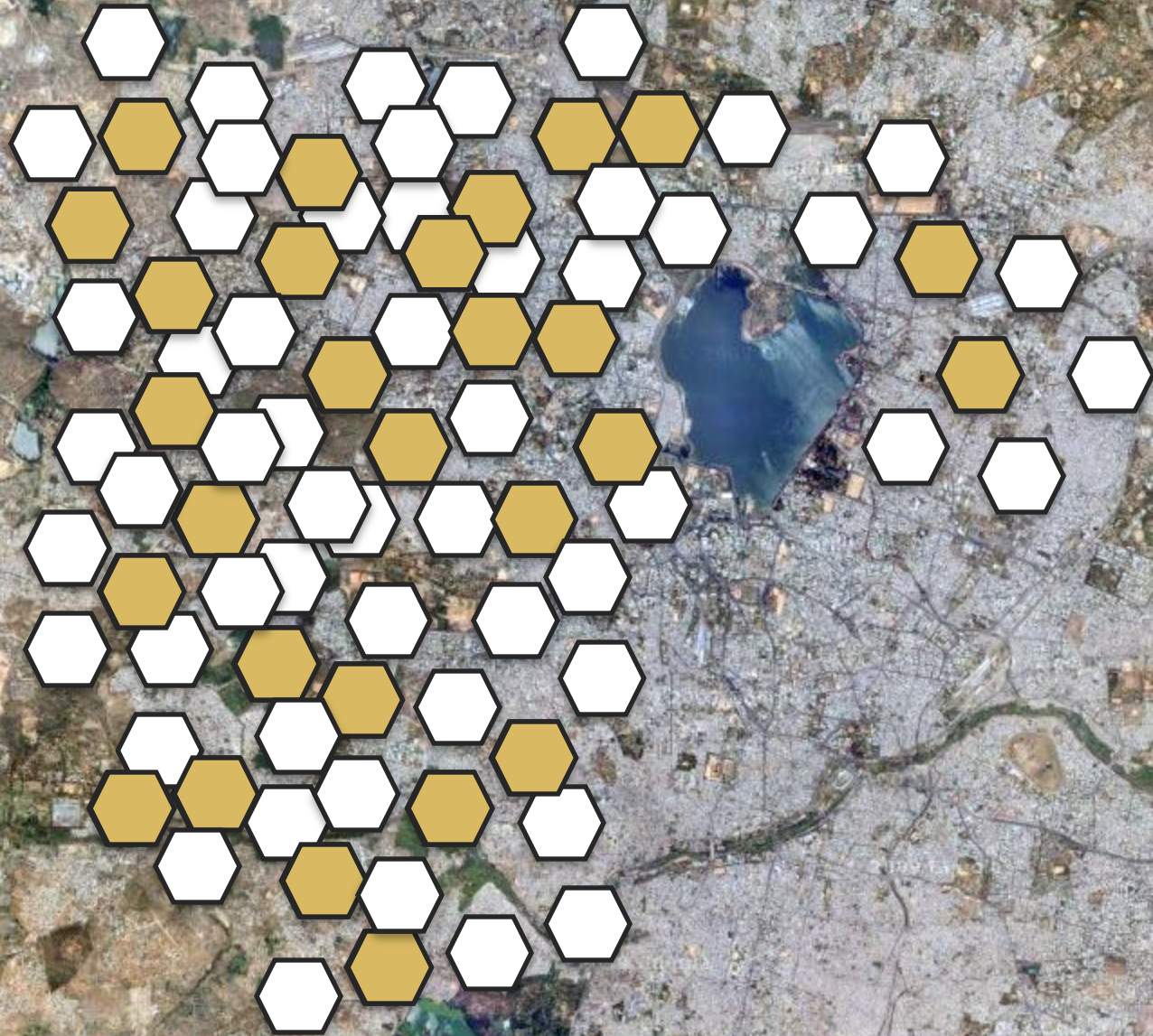# Testing Assumptions: Randomized Evaluations

- What is the main assumption of randomized evaluation that must hold for it to give the true impact of the program?

  – No randomization failure: that randomization generates two statistically identical groups

- How can you test whether this assumption is true?

  – Balance test – compare their characteristics at baseline (beginning of the program)

Basic set-up of a Randomized Evaluation
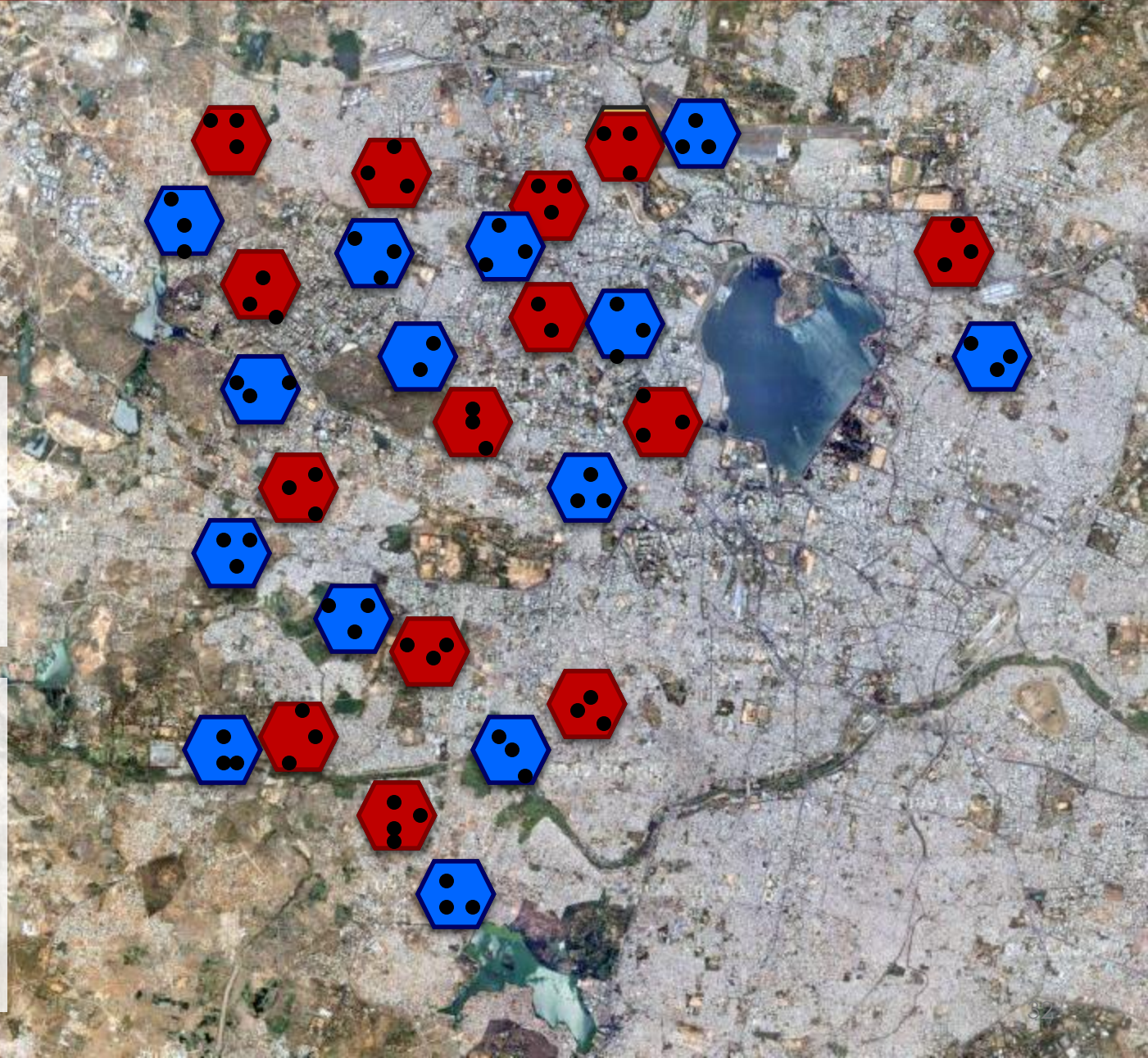
Randomly *sample* from area of interest
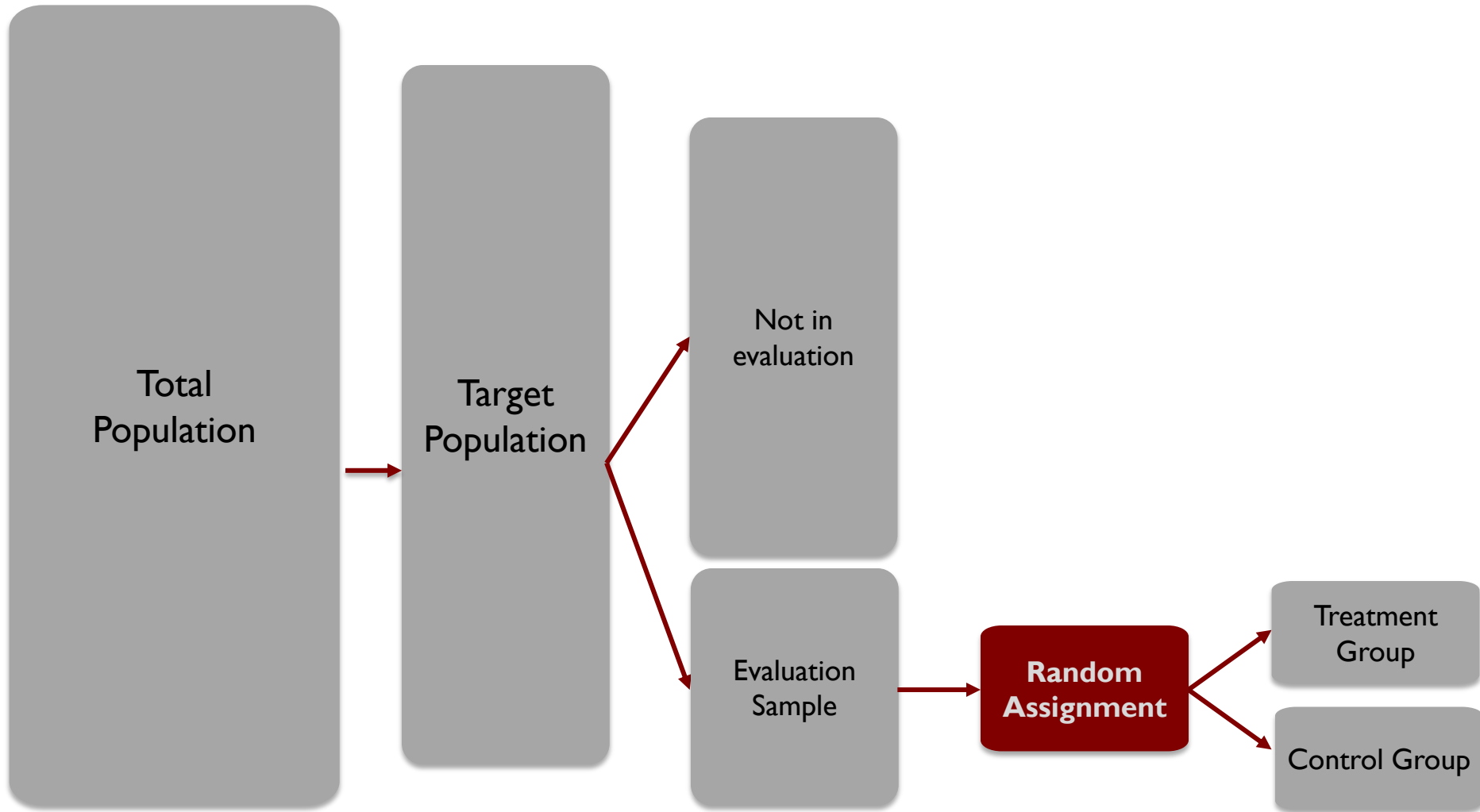
Random Sampling and Random Assignment

Randomly *sample* from area of interest

Randomly *assign* to **treatment** and **control**

Randomly *sample* from both treatment and control

NON AND QUASI-EXPERIMENTAL METHODS

Non or Quasi-Experimental Methods

- Let us look at other methods of estimating impact using the data from the schools that got a balsakhi

  1. Pre – Post (Before vs. After)
  2. Simple difference
  3. Difference-in-difference

- Other methods can be effective if the specific conditions needed for that method's assumption to hold exist

- Limitation: Conditions needed for them to be valid do not always apply

1 - Pre-post (Before vs. After)

- Look at average change in test scores over the school year for the balsakhi children
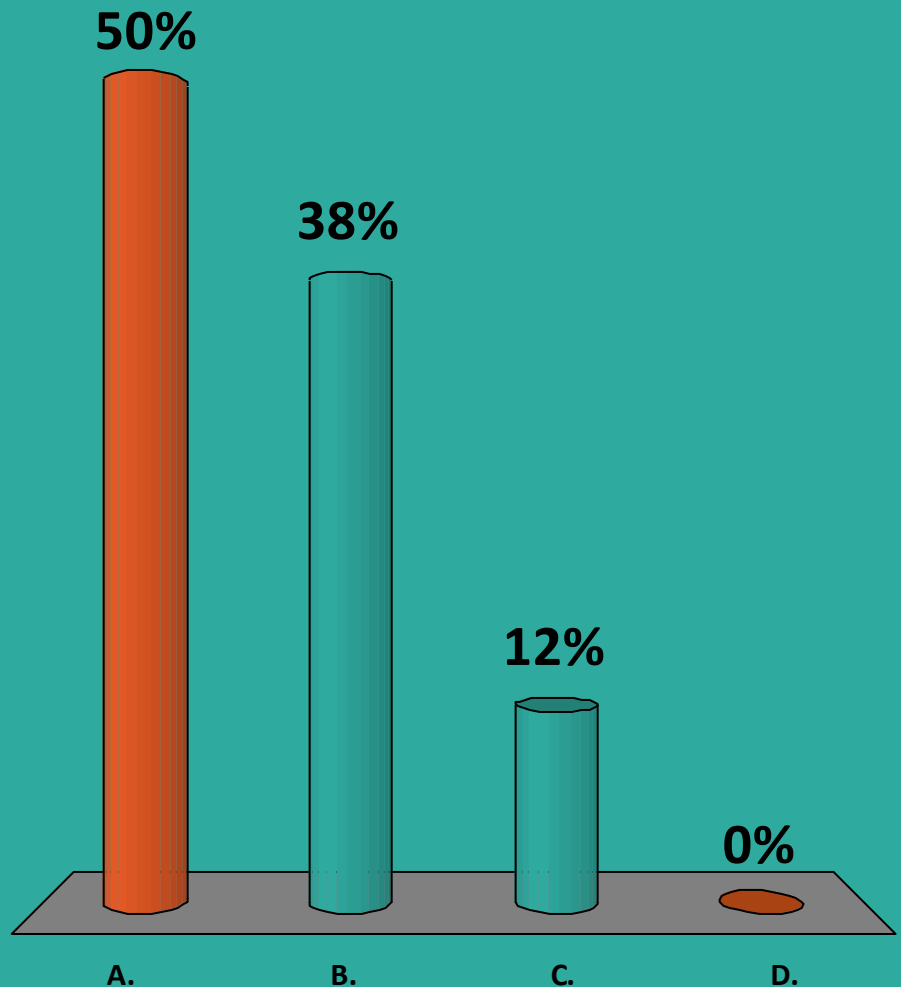
1 - Pre-post (Before vs. After)

Average <u>post-test</u> score for children with a balsakhi	51.22
Average <u>pretest</u> score for children with a balsakhi	24.80
Difference	26.42

QUESTION: Under what conditions can this difference (26.42) be interpreted as the impact of the balsakhi program?

# Which of the following represents the counterfactual in this case:

A. Balsakhi students before participating in the program

B. The non-Balsakhi students in the same schools

C. Students from other schools in Vadodara where the Balsakhi program is not being implemented

D. None of the above

What would have happened without Balsakhi?

Method 1: Before vs. After

Impact = 26.42 points?
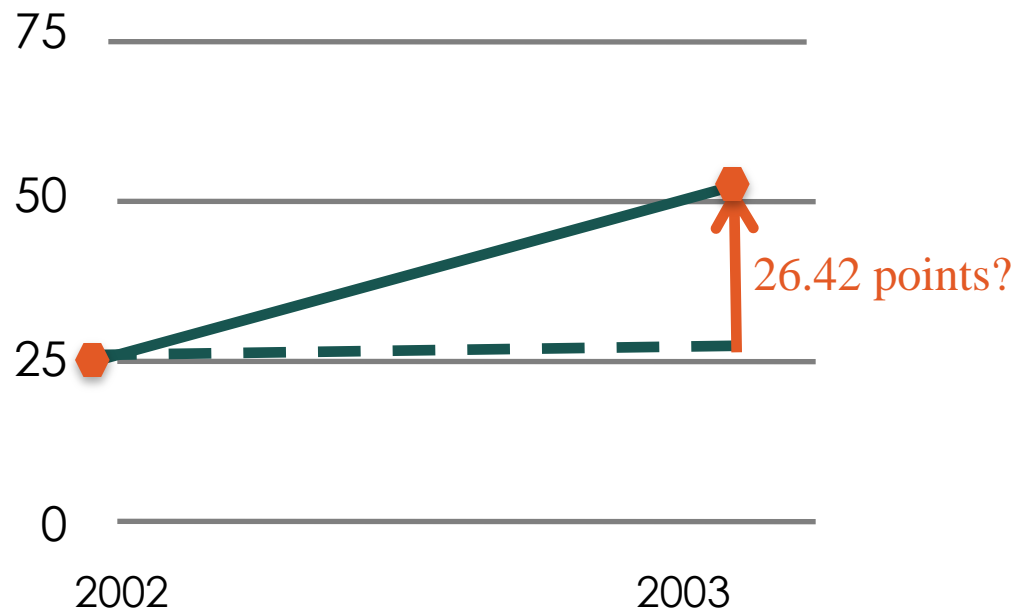
2 - Simple difference

Compare test scores of…

With test scores of…

Children who got balsakhi

Children who did not get balsakhi

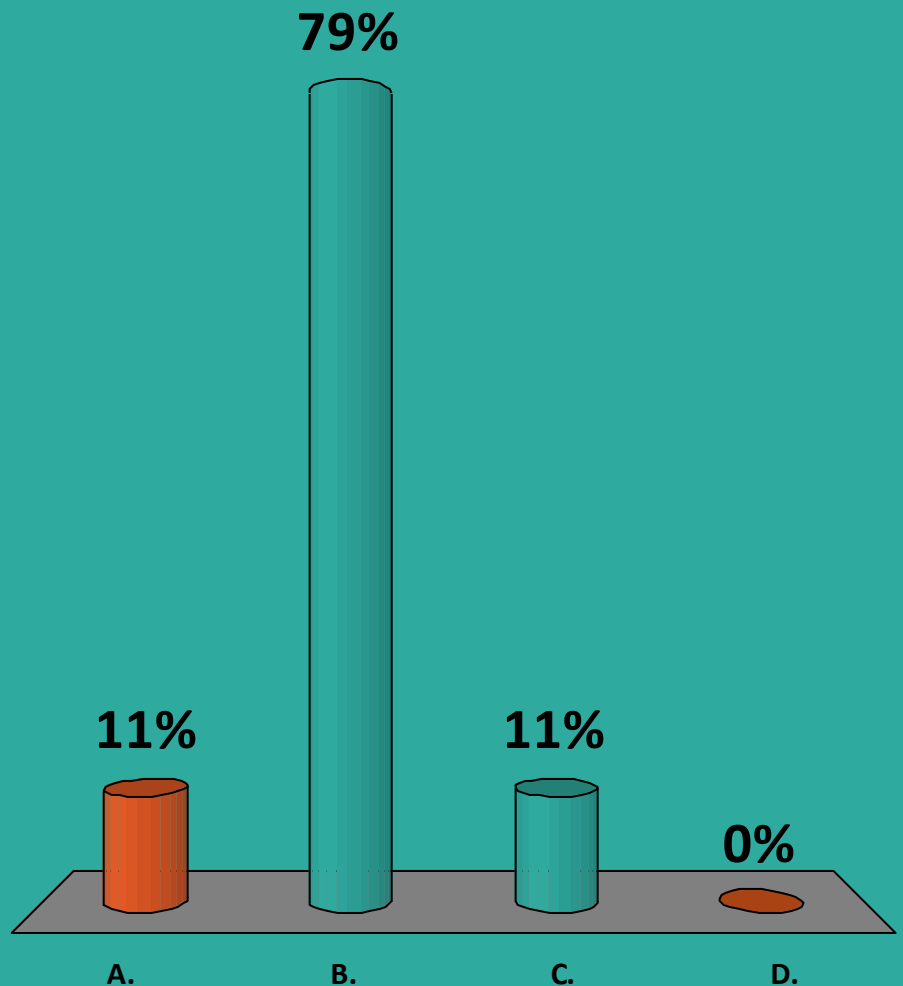# 2 - Simple difference

Average score for children with a balsakhi	**51.22**
Average score for children without a balsakhi	56.27
Difference	-5.05

QUESTION: Under what conditions can this difference (-5.05) be interpreted as the impact of the balsakhi program?

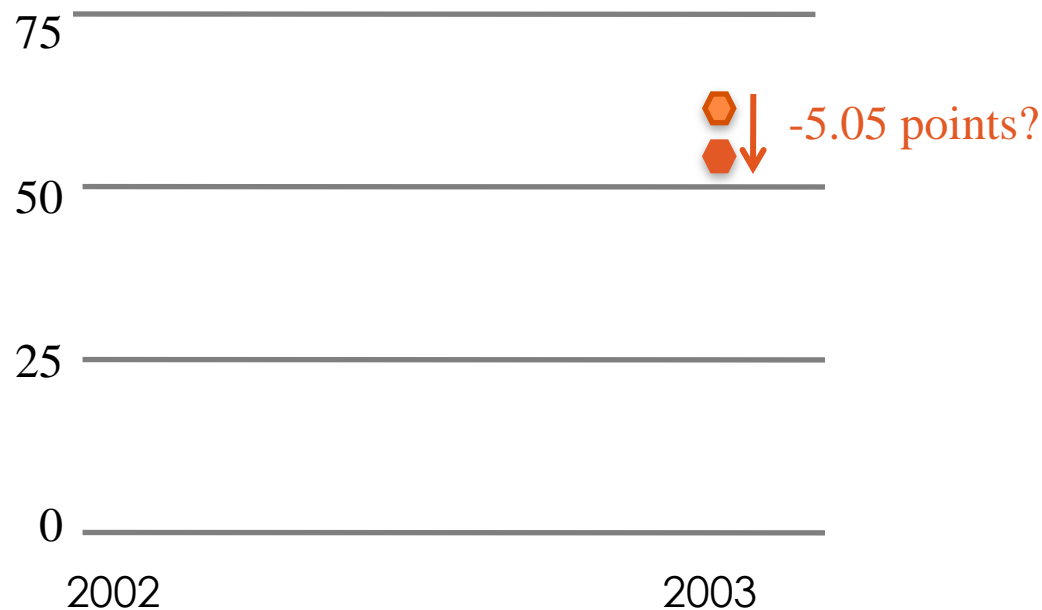# Which of the following represents the counterfactual in this case:

A. Balsakhi students before participating in the program

B. The non-Balsakhi students in the same schools

C. Students from other schools in Vadodara where the Balsakhi program is not being implemented

D. None of the above

79%

11%

11%

0%

A.   B.   C.   D.

What would have happened without balsakhi?

Method 2: Simple Comparison

Impact = -5.05 points?

Selection Bias

Population

Non-participants

Participants

Baseline **Intervention** **Endline**
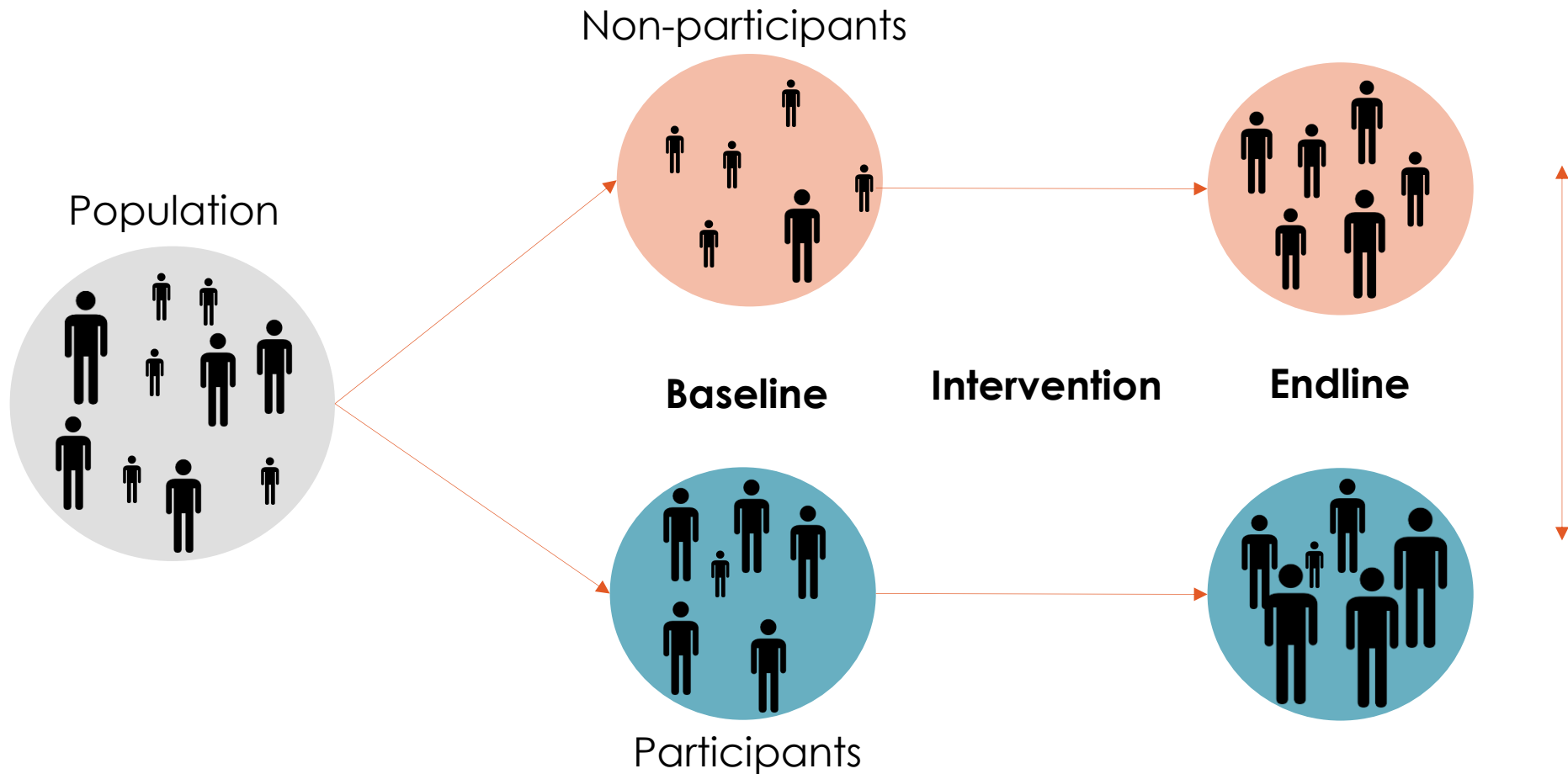
Is this difference due to the program?

Or pre-existing differences?

3 – Difference-in-Differences

Compare <u>gains</u> in test scores of…

With gains in test scores of…

Children who got balsakhi

Children who did not get balsakhi

3 – Difference-in-difference

	Pretest	Post-test	Difference
Average score for children with a balsakhi	24.80	51.22	26.42

3 – Difference-in-difference

	Pretest	Post-test	Difference
Average score for children with a balsakhi	24.80	51.22	26.42
Average score for children without a balsakhi	36.67	56.27	19.60

3 – Difference-in-difference

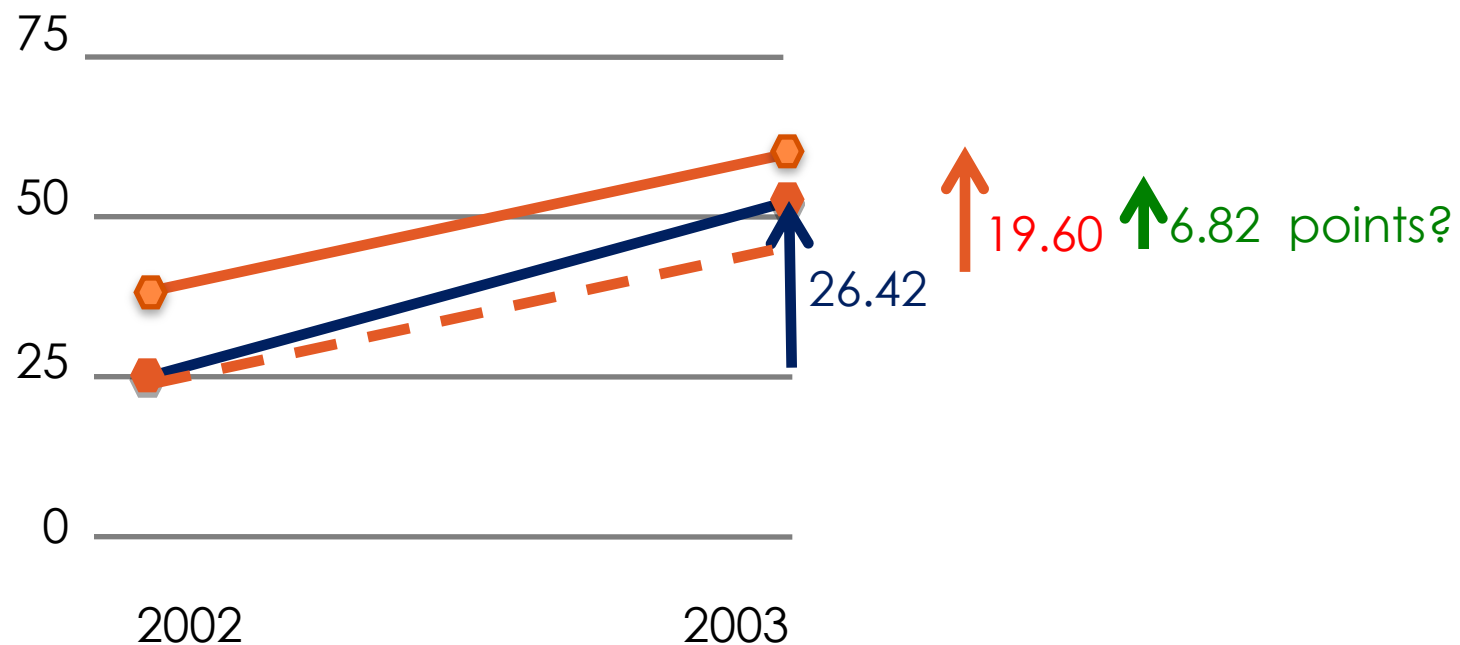|	**Pretest**	**Post-test**	**Difference**
Average score for children with a balsakhi	24.80	51.22	26.42
Average score for children without a balsakhi	36.67	56.27	19.60
Difference			6.82

QUESTION: Under what conditions can this difference (6.82) be interpreted as the impact of the balsakhi program?

What would have happened without balsakhi?

- Method 3: Difference-in-differences

4 – Multivariate Regression

A regression creates a model that estimates the relationship between a dependent and independent variable.

$$\text{Literacy} = a + b_1(\text{Tutor}) + \mathbf{b_2(\text{Controls})} + \text{Error}$$

We can control for observable factors like:

- pretest score
- income
- gender
- age

	Difference
Average score for children with a balsakhi	1.92

We can't control for unobservable characteristics. This bias is called omitted variable bias.

5 – Other Methods

- There are more sophisticated non-experimental methods to estimate programme impacts:

  - Matching
  - Instrumental Variables
  - Regression Discontinuity

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

- Problem: Assumptions are not testable

Which of these methods do you think is closest to the truth?

Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92

*: Statistically significant at the 5% level

A. Pre-Post

B. Simple Difference

C. Difference-in-Differences

D. Regression

E. Don't know

52%

30%

11%

4%

4%

A.   B.   C.   D.   E.
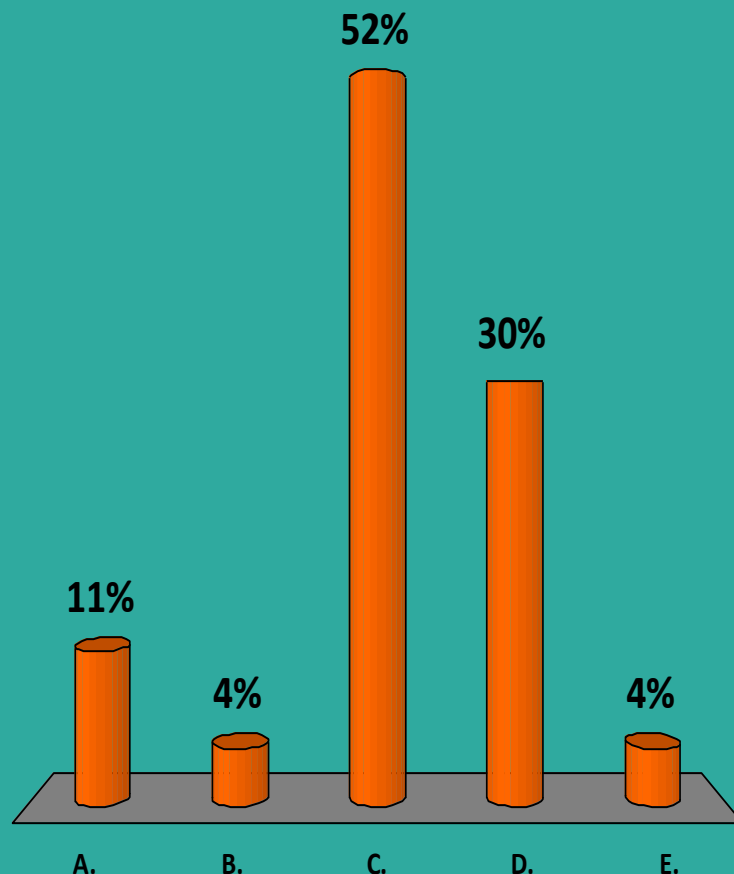
Impact of Balsakhi - Summary

Method	Impact Estimate
(1) Pre-Post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Differences	6.82*
(4) Regression	1.92
(5) Randomized Experiment	**5.87***

*: Statistically significant at the 5% level

Bottom Line: Which method we use matters!

IV – CONCLUSIONS

Conclusions - Why Randomize?

- There are many ways to estimate a program's impact
- This course argues in favor of one: randomized experiments
  - Conceptual argument: <u>If properly designed and conducted</u>, randomized experiments provide the most credible method to estimate the impact of a program
  - Empirical argument: Different methods can generate different impact estimates

sharon.barnhardt@flame.edu.in

Follow us on Twitter @JPAL_SA
Visit our website www.povertyactionlab.org for more resources