J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

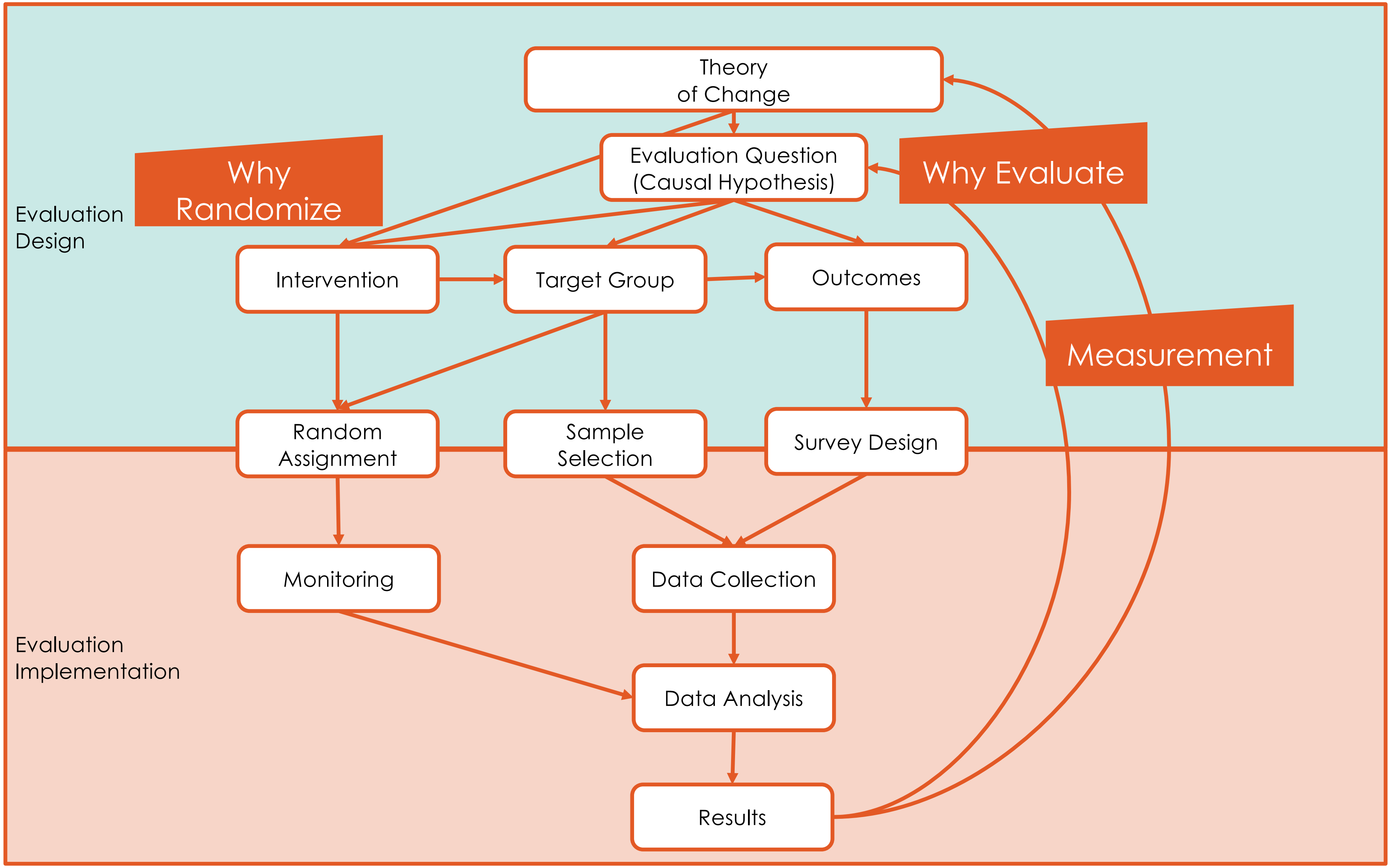
Why Randomize?

J-PAL

Course Overview

1. What is Evaluation?
2. Outcomes, Impact, and Indicators
3. Why Randomize?
4. How to Randomize
5. Sampling and Sample Size
6. Threats and Analysis
7. Start to Finish
8. Generalizability

Randomized Evaluation Process



Evaluation Design

Evaluation Implementation

Theory of Change

Evaluation Question (Causal Hypothesis)

Why Randomize

Why Evaluate

Measurement

Intervention

Target Group

Outcomes

Random Assignment

Sample Selection

Survey Design

Monitoring

Data Collection

Data Analysis

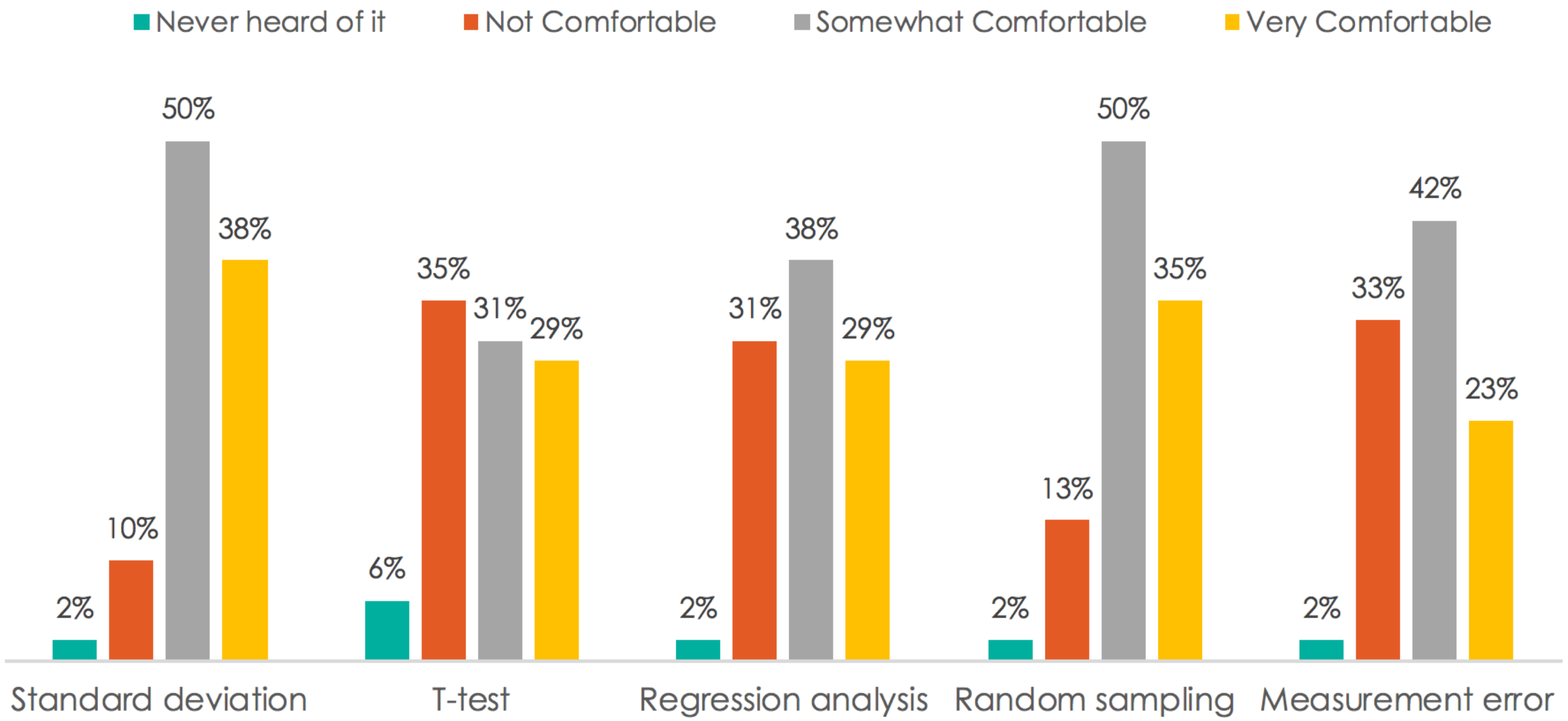
Results

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Why Randomize?

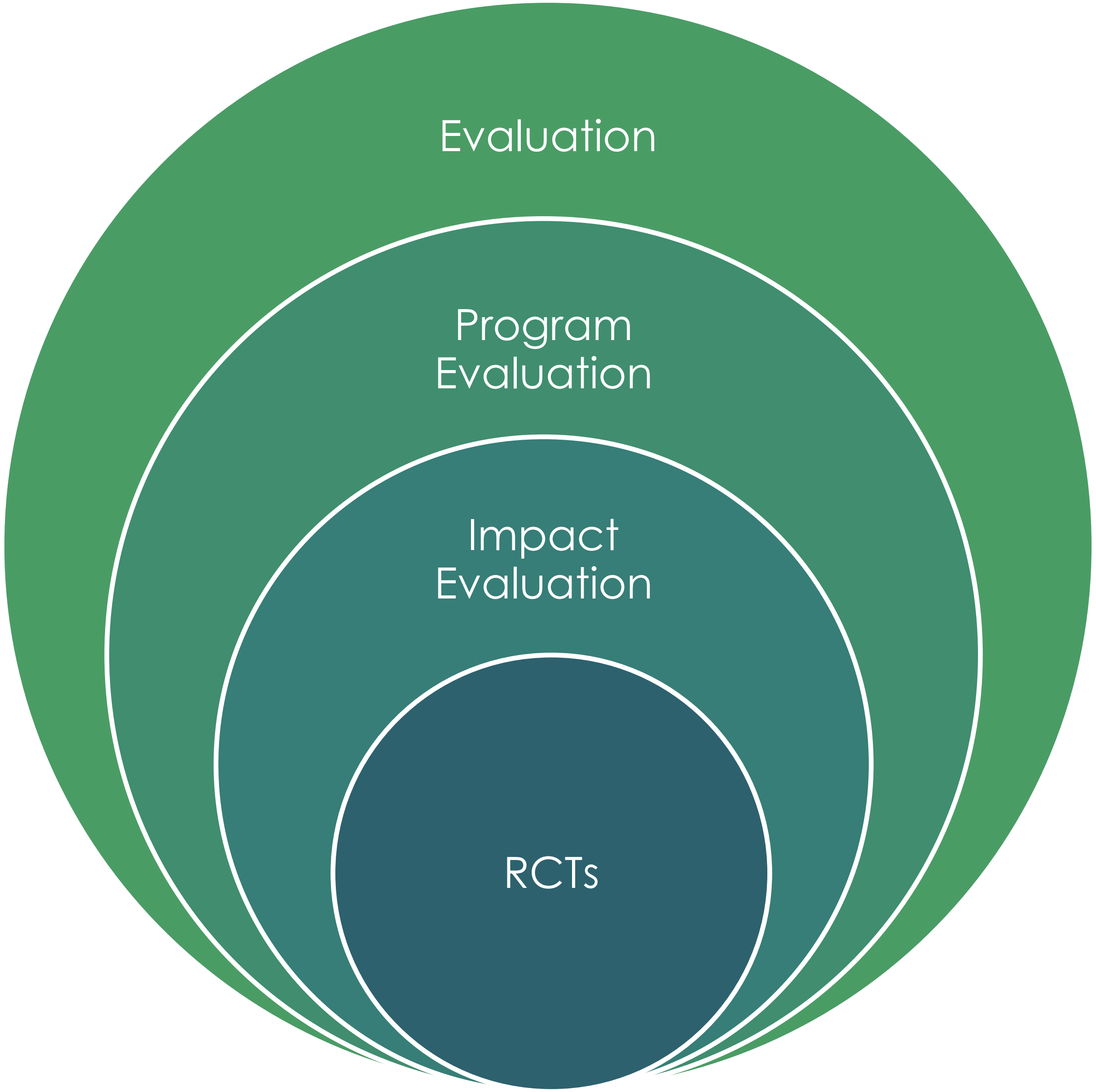
Dan Levy
Harvard Kennedy School

Your background



Participants at Evaluating Social Programs 2019

What is Impact Evaluation?



Methodologically, randomized controlled trials (RCTs) are the best approach to estimate the effect of a program

A. Strongly Disagree

B. Disagree

C. Neutral

D. Agree

E. Strongly Agree

Session Overview

I. Background

II. What is an RCT?

III. Why randomize?

IV. Conclusions

Session Overview

I. Background

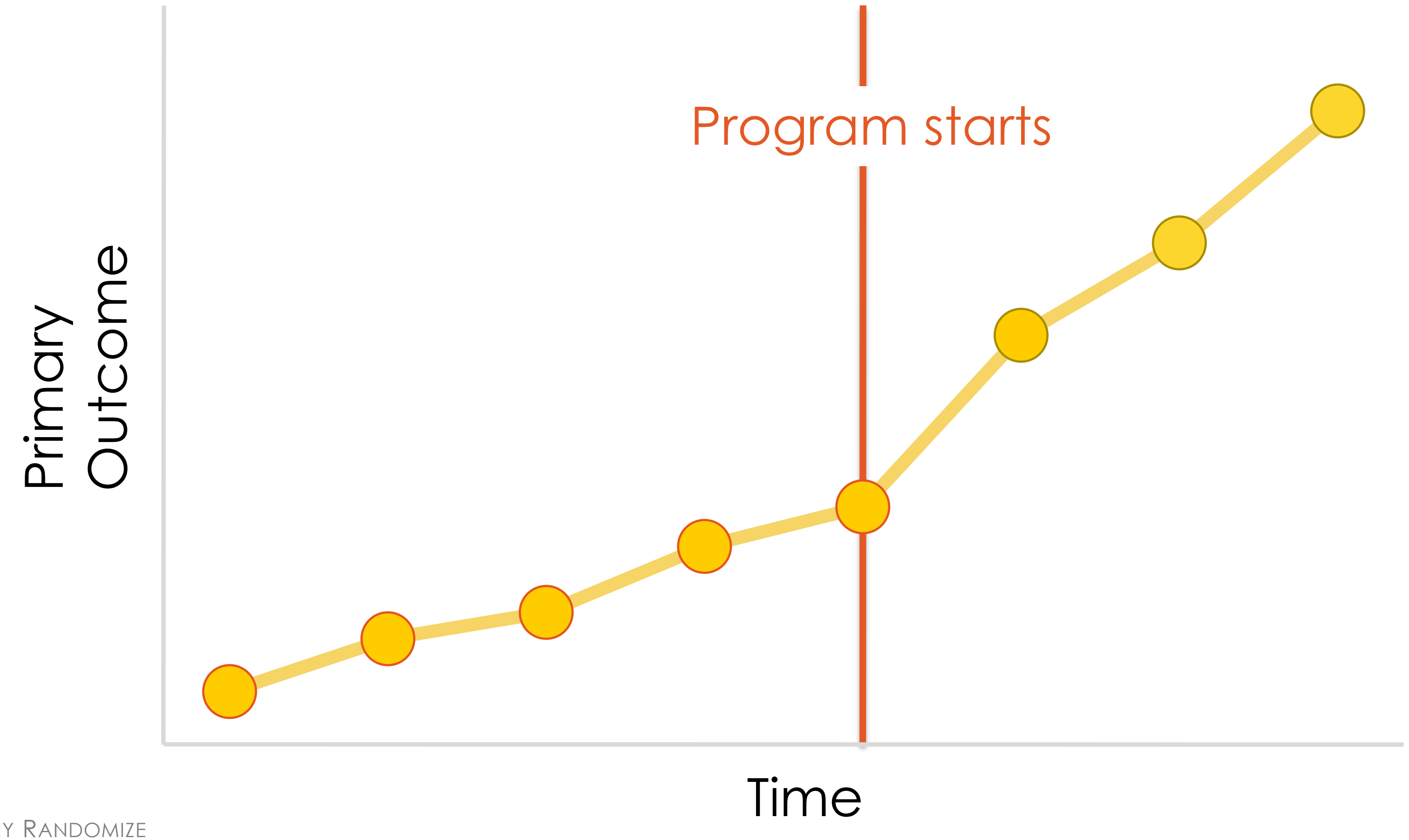
II. What is an RCT?

III. Why randomize?

IV. Conclusions

I - BACKGROUND

What is the impact of this program?



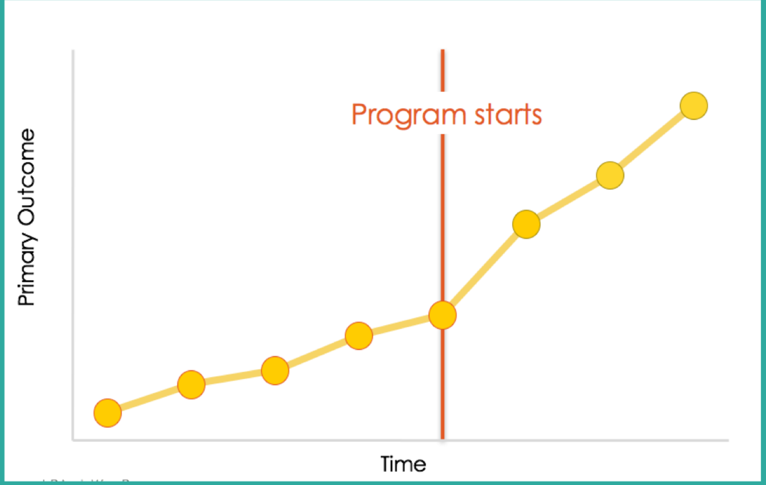
What is the impact of this program?

A. Positive

B. Negative

C. Zero

D. Not enough info



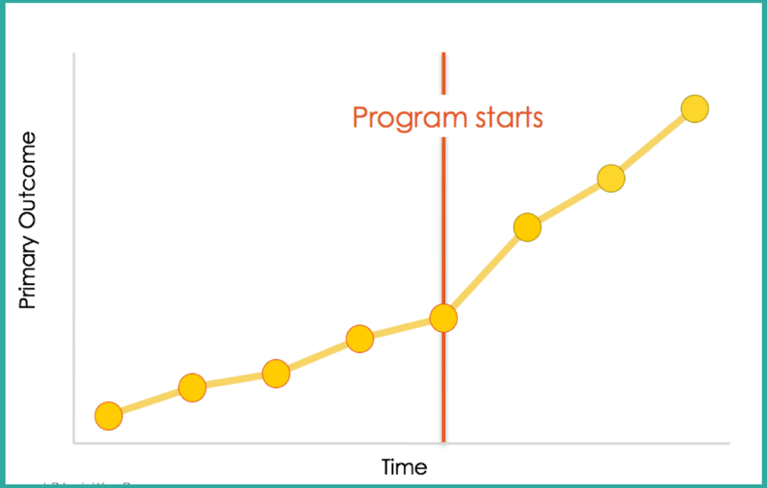
What is the impact of this program?

A. Positive

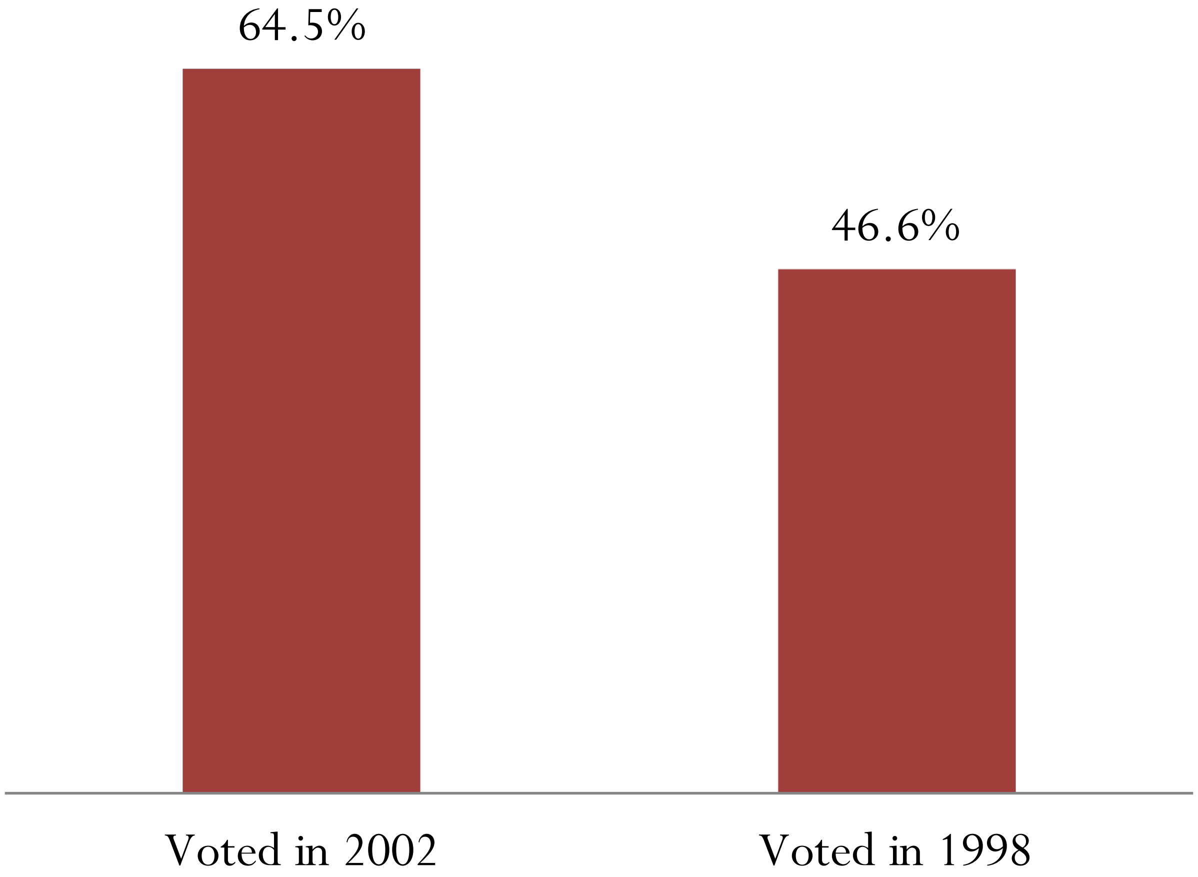
B. Negative

C. Zero

D. Not enough info

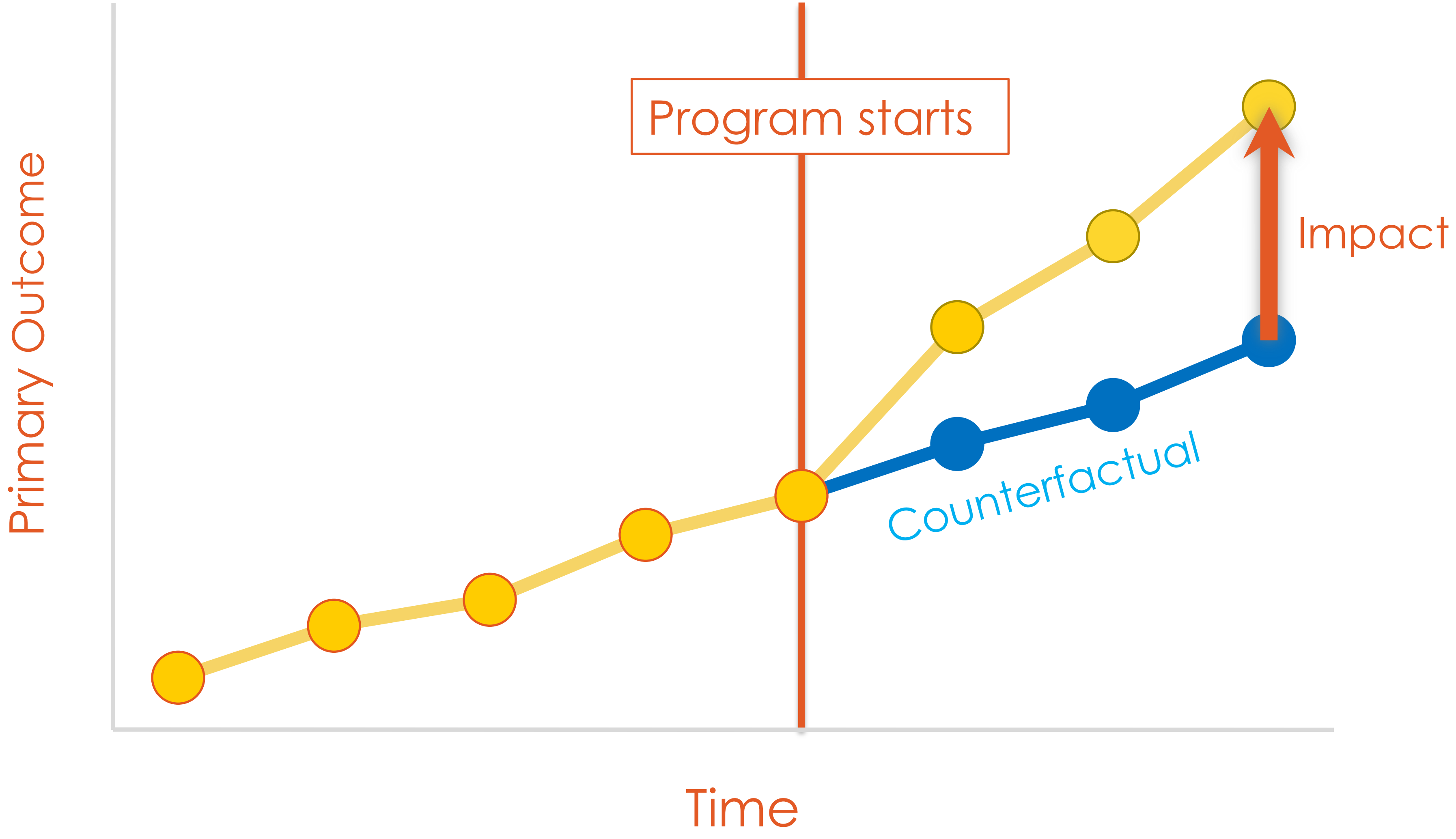


Vote 2002 Campaign: Huge Success?



"Before vs. After" is rarely a good method for assessing impact.

What is the impact of this program?



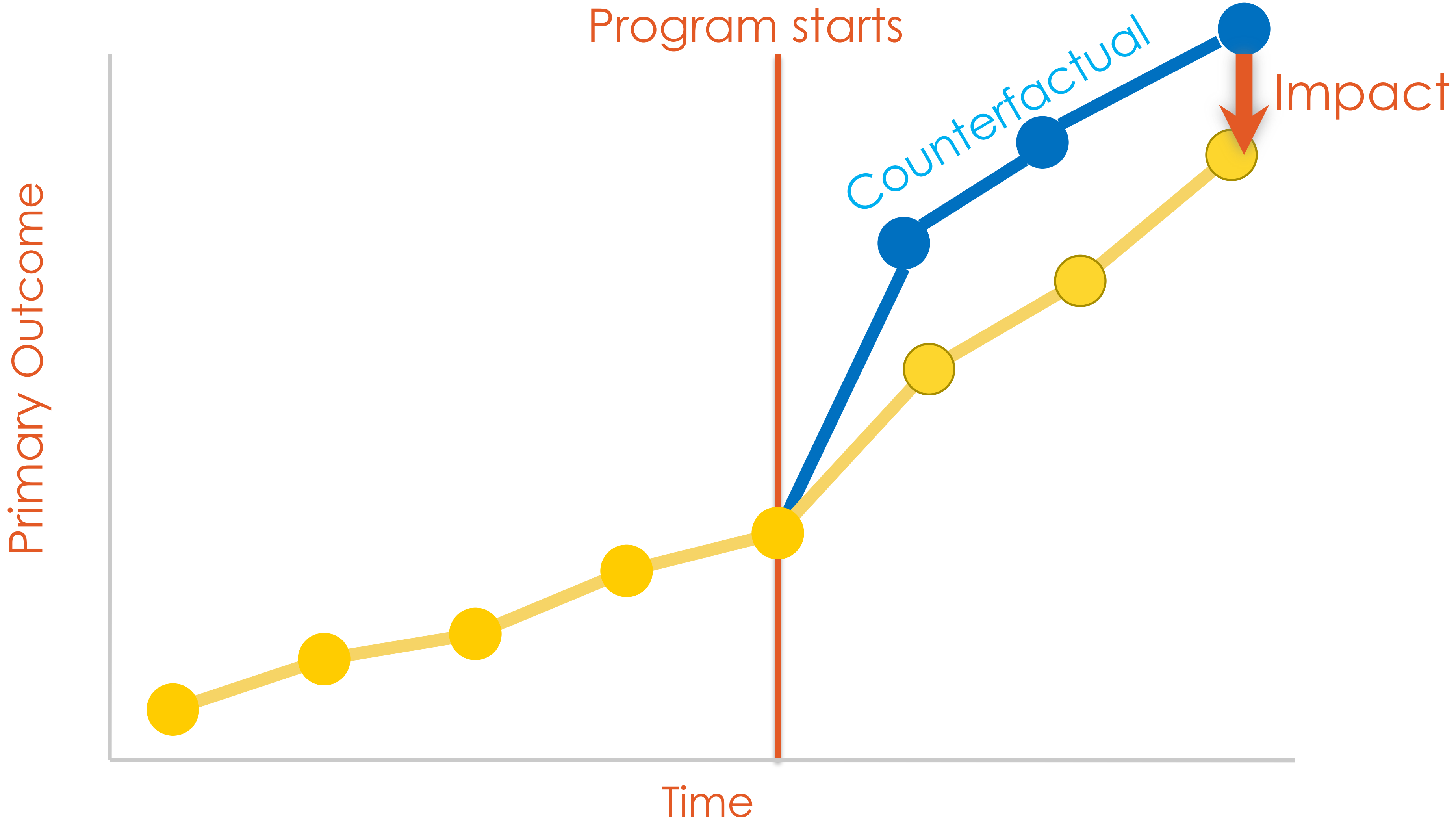
How to measure impact?

Impact is defined as a comparison between:

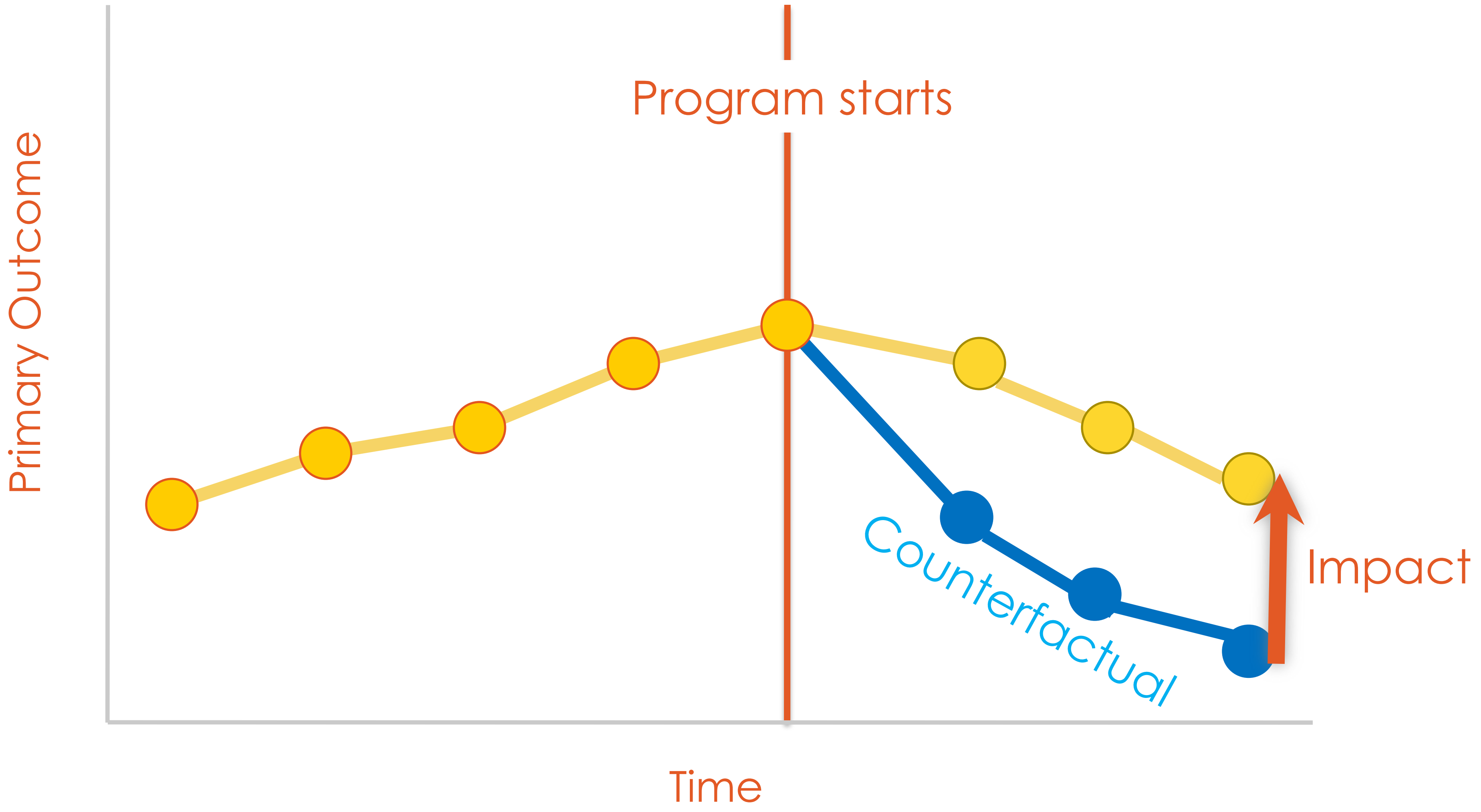
1. the outcome some time after the program has been introduced (the *"factual"*)

2. the outcome at that same point in time had the program not been introduced (the *"counterfactual"*)

Impact: What is it?



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

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: Comparability

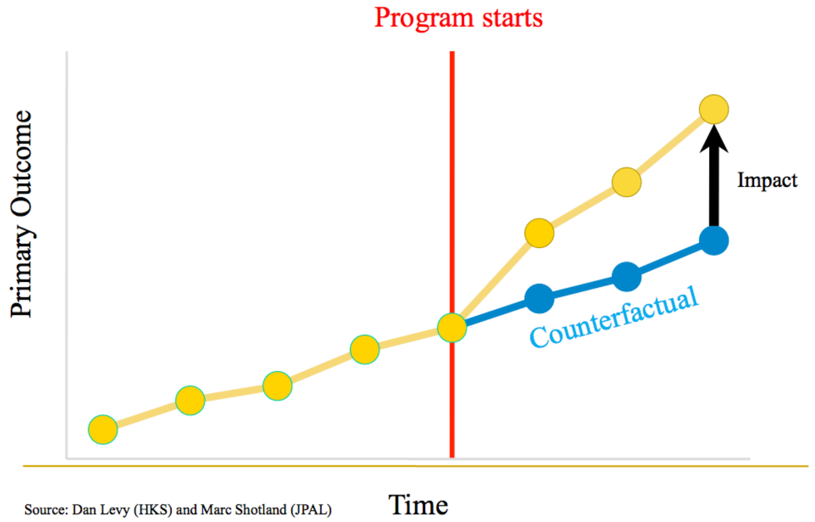
Treatment

Comparison

- Goal: Attribution

3 Key Ideas about Impact

1 - Counterfactual



Source: Dan Levy (HKS) and Marc Shotland (JPAL)

2 – Comparison group mimics the counterfactual

Treatment

Comparison

3 - Goal of Impact Evaluations: Attribution

Impact evaluation methods

1. **Randomized Controlled Trials (RCTs)**

   Also known as:

   - Random Assignment Studies
   - Randomized Field Trials
   - Social Experiments
   - Randomized Trials
   - Randomized Experiments
   - Randomized Controlled Experiments

Impact evaluation methods

2. **Non- or Quasi-Experimental Methods**
   - Pre-Post
   - Simple Difference
   - Differences-in-Differences
   - Multivariate Regression
   - Statistical Matching
   - Interrupted Time Series
   - Instrumental Variables
   - Regression Discontinuity

Session Overview

I. Background

II. What is an RCT?

III. Why randomize?

IV. Conclusions

II – WHAT IS AN RCT?

The basics

Start with simple case:

- Take a sample of program applicants
- Assign them to either:
  - *Randomly* as Treatment Group – are offered treatment
  - Control Group – are not offered the treatment (during the evaluation period)

Key advantage of randomized evaluations

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

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

Treatment

Comparison

Evaluation of "Women as Policymakers": Treatment vs. Control villages at baseline

Variables	Treatment Group	Control Group	Difference
Female Literacy Rate	0.35	0.34	0.01 (0.01)
Number of Public Health Facilities	0.06	0.08	-0.02 (0.02)
Tap Water	0.05	0.03	0.02 (0.02)
Number of Primary Schools	0.95	0.91	0.04 (0.08)
Number of High Schools	0.09	0.10	-0.01 (0.02)

Standard Errors in parentheses. Statistics displayed for West Bengal

*/**/***: Statistically significant at the 10% / 5% / 1% level

Source: Chattopadhyay and Duflo (2004)

Some variations on the basics

- Assigning to multiple treatment groups

- Assigning of units other than individuals or households
  - Health Centers
  - Schools
  - Local Governments
  - Villages

Key Steps in Conducting a Randomized Evaluation

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

Key Steps in Conducting a Randomized Evaluation (contd.)

6. Collect follow-up 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

Session Overview

I. Background

II. What is an RCT?

III. Why randomize?

IV. Conclusions

III – WHY RANDOMIZE?

Why Randomize?

Conceptual Argument

Empirical Argument

Why Randomize?

Conceptual Argument

Empirical Argument

Why Randomize? - Conceptual Argument

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

Why “most credible”?

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

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

Treatment

Comparison

Why Randomize?

Conceptual Argument

Empirical Argument

Example #1 – Pratham’s Read India program

Example #1 – Pratham's Read India program

Method	Impact
(1) Pre-Post	0.60*
(2) Simple Difference	-0.90*
(3) Difference-in-Differences	0.31*
(4)Regression	0.06

*: Statistically significant at the 5% level

Example #1 – Pratham's Read India program

Method	Impact
(1) Pre-Post	0.60*
(2) Simple Difference	-0.90*
(3) Difference-in-Differences	0.31*
(4)Regression	0.06
(5) Randomized Evaluation	**0.88***

*: Statistically significant at the 5% level

Example #2 – A voting campaign in the USA

Courtesy of Flickr user theocean

Example #2 – A voting campaign in the USA

Method	Estimated Impact
(1) Pre-Post	17.9 pp*
(2) Simple Difference	10.8 pp*
(3) Difference-in-Differences	1.9 pp*
(4)Multiple Regression	4.6 pp*
(5) Matching	2.8 pp*

pp= percentage points; *: Statistically significant at the 5% level

Example #2 – A voting campaign in the USA

Method	Estimated Impact
(1) Pre-Post	17.9 pp*
(2) Simple Difference	10.8 pp*
(3) Difference-in-Differences	1.9 pp*
(4)Multiple Regression	4.6 pp*
(5) Matching	2.8 pp*
(6) Randomized Evaluation	**0.4 pp**

pp= percentage points; *: Statistically significant at the 5% level

Example #2 – A voting campaign in the USA

Method	Estimated Impact
(1) Pre-Post	17.9 pp*
(2) Simple Difference	10.8 pp*
(3) Difference-in-Differences	1.9 pp*
(4)Multiple Regression	4.6 pp*
(5) Matching	2.8 pp*
(6) Randomized Evaluation	**0.4 pp**

pp= percentage points; *: Statistically significant at the 5% level

Bottom Line: Which method we use matters

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

A. Too expensive
B. Not ethical
C. Too difficult to design/implement
D. Not externally valid (Not generalizable)
E. Can tell us *what the impact is* impact, but not *why* or *how* it occurred (i.e. it is a black box)

Methodologically, randomized controlled trials (RCTs) are the best approach to estimate the effect of a program

A. Strongly Disagree

B. Disagree

C. Neutral

D. Agree

E. Strongly Agree

IV – CONCLUSIONS

Conclusions – Why Randomize?

- There are many ways to estimate a program’s impact

- This course argues in favor of one: RCTs
  - **Conceptual argument:** If properly designed and conducted, RCTs provide the most credible method to estimate the impact of a program
  - **Empirical argument:** Different methods can generate different impact estimates

THANK YOU!

References, Reuse, and Citation

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Why Randomize? Backup Slides

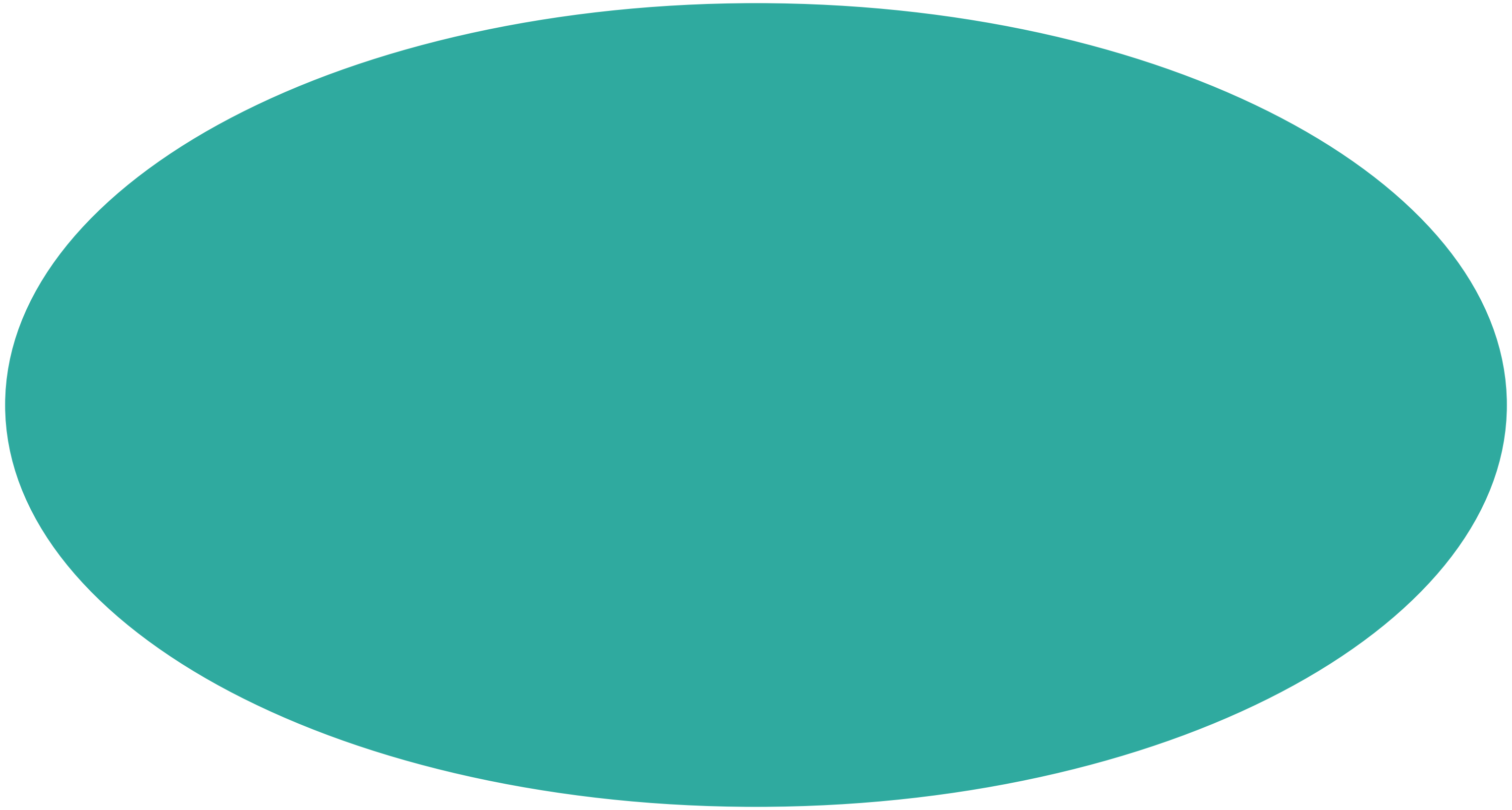
Dan Levy

Harvard Kennedy School

Program: "Get Out the Vote"

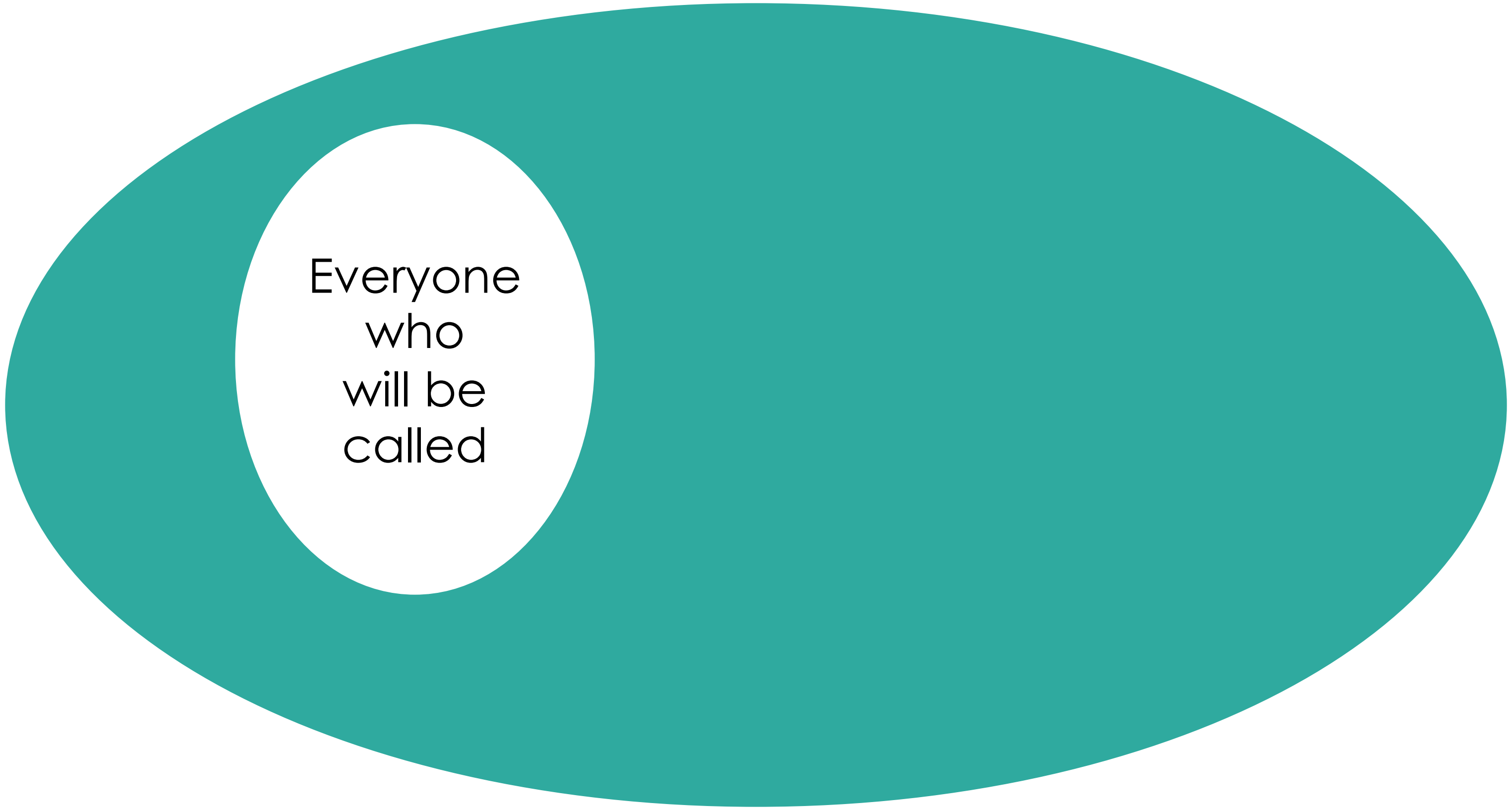
- Low voter turnout is seen as a problem in many countries in the world
- Some countries have looked for ways to increase voter turnout
- "Get Out the Vote" Program
  - Compiled a list of all the 100,000 individuals who could vote in an election
  - Call a sample individuals in this list
  - In this phone call, responder is encouraged to vote

Program: “Get Out the Vote”



Everyone eligible to vote (100,000)

Program: “Get Out the Vote”



Everyone eligible to vote (100,000)

Everyone who will be called

Program: “Get Out the Vote”(Contd.)

Key Question: What is the **impact** of the “Get Out the Vote” program on the voter turnout rate?

Methodological Question: How should we estimate the impact of the program?

Resources available for the evaluation

- List of all the persons eligible to vote with information on:
  - Income
  - Education
  - Sex
  - Age
  - Whether person voted in the last election

- Money to make up to 8,000 calls that could be used to:
  - Implement the program (i.e. call before the election encouraging person to vote)
  - Collect data (i.e. call people after the election to ask whether they voted or not)

- List of 2,000 people who came to a political rally one month before the election
  - You already called them and encouraged them to vote
  - These calls count as part of your 8,000 calls

Which design would you choose?

A. Design 1
B. Design 2
C. Design 3
D. Design 4
E. Design 5

Methodologically, randomized trials are the best approach to estimate the effect of a program

A. Strongly Disagree

B. Disagree

C. Neutral

D. Agree

E. Strongly Agree

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

A. Too expensive

B. Not ethical

C. Too difficult to design/implement

D. Not externally valid (Not generalizable)

E. Can tell us *what the impact is* impact, but not *why* or *how* it occurred (i.e. it is a black box)

What do you want to do?

A. Example

B. Objections to RCTs

Example #3 – Balsakhi Program

Balsakhi Program: Background

- Implemented by Pratham, an NGO from India
- Program provided tutors (Balsakhi) to help at-risk children with school work
- 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?

Methods to estimate impacts

- Let’s look at different ways of estimating the impacts using the data from the schools that got a balsakhi
    1. Pre – Post (Before vs. After)
    2. Simple difference
    3. Difference-in-difference
    4. Other non-experimental methods
    5. Randomized Evaluation

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)

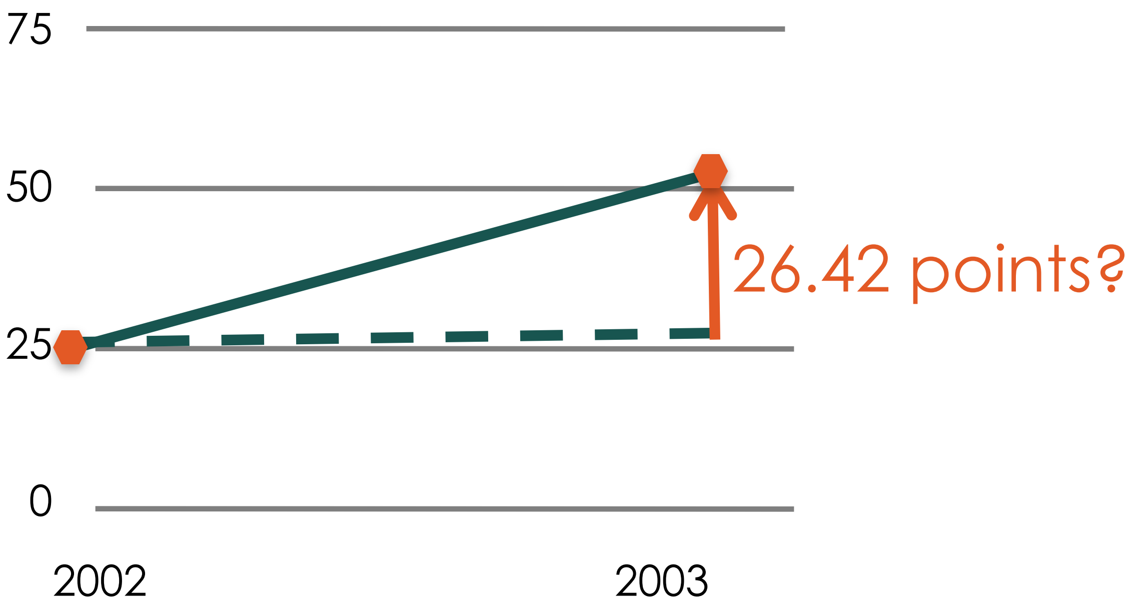
Average post-test score for children with a balsakhi	51.22
Average pretest score for children with a balsakhi	24.80
Difference	26.42

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

What would have happened without balsakhi?

Method 1: Before vs. After

Impact = 26.42 points?



2 - Simple difference

Compare test scores of…

Children who got balsakhi

With test scores of…

Children who did not get balsakhi

2 - Simple difference

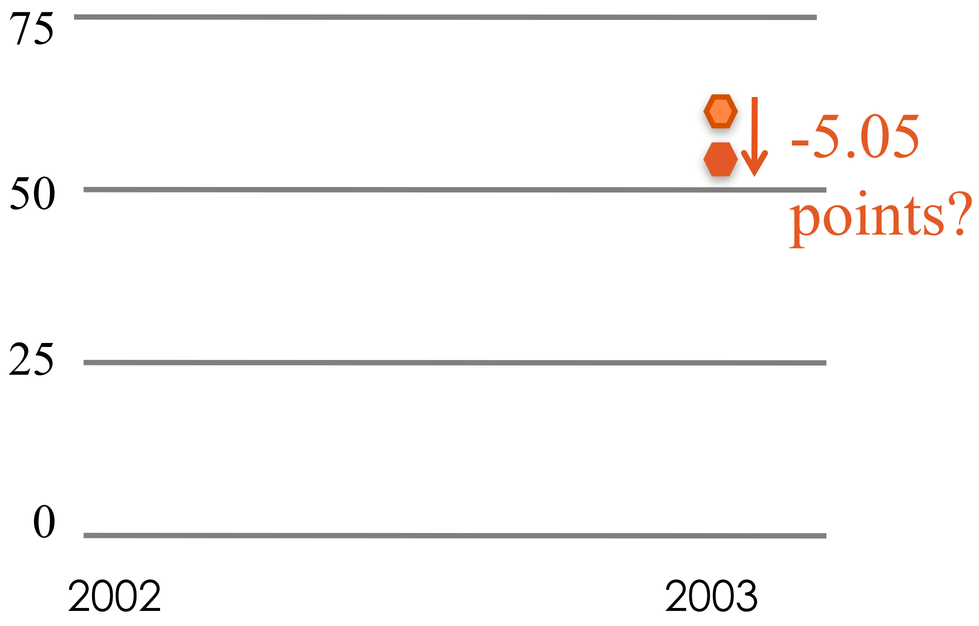
Average score for children with a balsakhi	**51.22**
Average score for children without a balsakhi	56.27
Difference	-5.05

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

What would have happened without balsakhi?

Method 2: Simple Comparison

Impact = -5.05 points?



3 - Difference-in-Differences

Compare gains in test scores of…

Children who got balsakhi

With gains in test scores of…

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

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

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

4 - Other Methods

- There are more sophisticated non-experimental methods to estimate program impacts:
  - Regression
  - Matching
  - Instrumental Variables
  - Regression Discontinuity

- These methods rely on being able to “mimic” the counterfactual under certain assumptions
- Problem: Assumptions are not testable

5 - 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?

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

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 Evaluation	**5.87***

*: Statistically significant at the 5% level

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 Evaluation	**5.87***

*: Statistically significant at the 5% level

Bottom Line: Which method we use matters!

Example #2 – Pratham’s Read India program

Example #2 – Pratham’s Read India program

Method	Impact
(1) Pre-Post	0.60*
(2) Simple Difference	-0.90*
(3) Difference-in-Differences	0.31*
(4) Regression	0.06
(5) Randomized Evaluation	

*: Statistically significant at the 5% level

Example #2 – Pratham’s Read India program

Method	Impact
(1) Pre-Post	0.60*
(2) Simple Difference	-0.90*
(3) Difference-in-Differences	0.31*
(4) Regression	0.06
(5) Randomized Evaluation	**0.88***

*: Statistically significant at the 5% level

Example #3 – A voting campaign in the USA

Courtesy of Flickr user theocean

A voting campaign in the USA

Method	**Impact (Vote %)**
(1) Pre-Post	-7.2 pp
(2) Simple Difference	10.8 pp*
(3) Difference-in-Differences	3.8 pp*
(4)Multiple Regression	6.1 pp*
(5) Matching	2.8 pp*
(5) Randomized Evaluation	**0.4 pp**

*: Statistically significant at the 5% level

What is the impact of this program?

A. Positive
B. Negative
C. Zero
D. Not enough info

What is the impact of this program?

A. Positive

B. Negative

C. Zero

D. I don’t know

E. Who knows?

Example #3 – Balsakhi Program

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 Evaluation	**5.87***

*: Statistically significant at the 5% level

THANK YOU!

Marshmallow Test

Selecting the comparison group

- Idea: Comparability

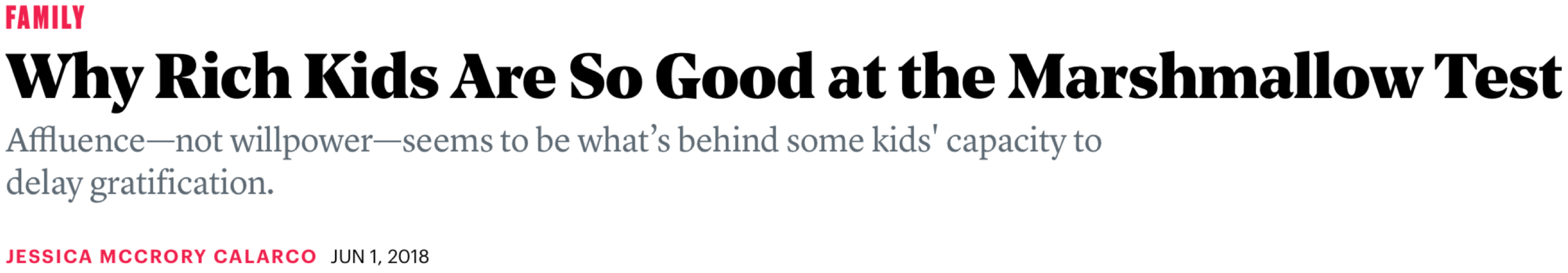
Treatment

Comparison

- Goal: Attribution

Marshmallow Test

Marshmallow Study Revisited

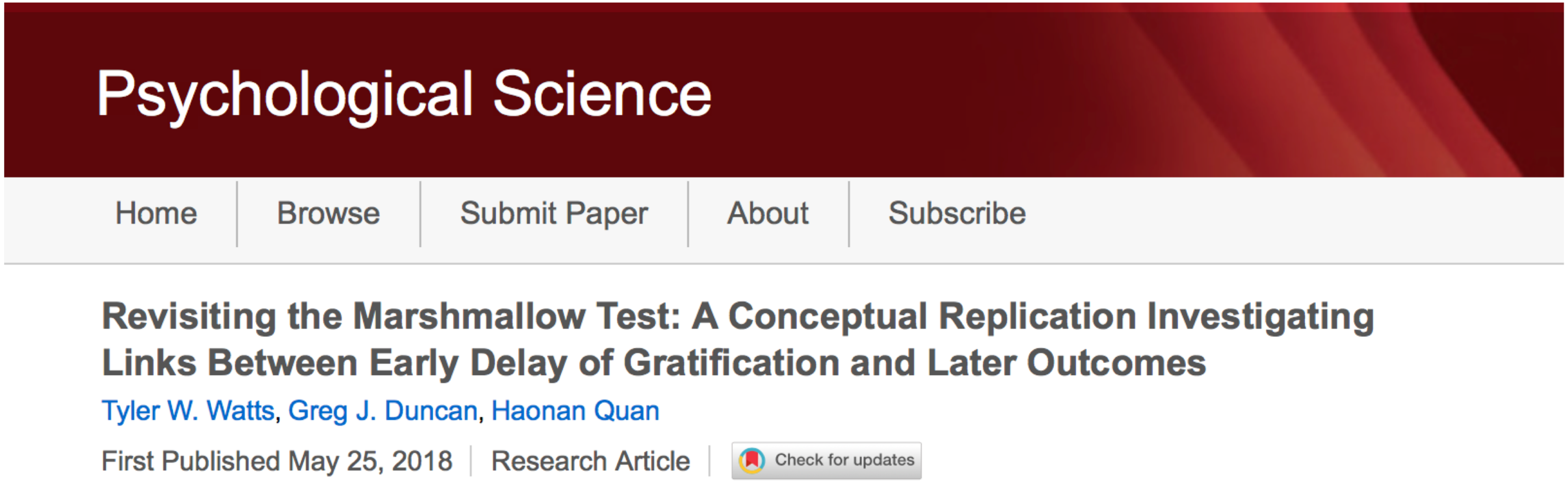


FAMILY

Why Rich Kids Are So Good at the Marshmallow Test

Affluence—not willpower—seems to be what's behind some kids' capacity to delay gratification.

JESSICA MCCRORY CALARCO JUN 1, 2018



Revisiting the Marshmallow Test: A Conceptual Replication Investigating Links Between Early Delay of Gratification and Later Outcomes

Tyler W. Watts, Greg J. Duncan, Haonan Quan

First Published May 25, 2018 | Research Article