Why Randomize?

Dan Levy
Harvard Kennedy School
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. Generalizability
8. Project from Start to Finish
Why Randomize?

Dan Levy
Harvard Kennedy School
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 a randomized experiment?

III. Why randomize?

IV. Conclusions
Session Overview

I. Background

II. What is a randomized experiment?

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
“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?

Program starts

Counterfactual

Impact

Outcome

Time
Impact: What is it?

- Time
- Primary Outcome
- Program starts
- Counterfactual
- Impact
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

• Goal: 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
   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
Session Overview

I. Background

II. What is a randomized experiment?

III. Why randomize?

IV. Conclusions
II – WHAT IS A RANDOMIZED EXPERIMENT?
The basics

Start with simple case:

- Take a sample of program applicants
- Randomly assign them to either:
  - Treatment Group – is offered treatment
  - Control Group – not allowed to receive treatment (during the evaluation period)
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.
Evaluation of “Women as Policymakers”: Treatment vs. Control villages at baseline

<table>
<thead>
<tr>
<th>Variables</th>
<th>Treatment Group</th>
<th>Control Group</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Literacy Rate</td>
<td>0.35</td>
<td>0.34</td>
<td>0.01 (0.01)</td>
</tr>
<tr>
<td>Number of Public Health Facilities</td>
<td>0.06</td>
<td>0.08</td>
<td>-0.02 (0.02)</td>
</tr>
<tr>
<td>Tap Water</td>
<td>0.05</td>
<td>0.03</td>
<td>0.02 (0.02)</td>
</tr>
<tr>
<td>Number of Primary Schools</td>
<td>0.95</td>
<td>0.91</td>
<td>0.04 (0.08)</td>
</tr>
<tr>
<td>Number of High Schools</td>
<td>0.09</td>
<td>0.10</td>
<td>-0.01 (0.02)</td>
</tr>
</tbody>
</table>

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 an experiment

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 experiments is not compromised
Key Steps in conducting an experiment (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
EXAMPLE
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 that 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

![Bar chart showing the percentage of votes for each design]

- A: 29%
- B: 6%
- C: 12%
- D: 53%
- E: 0%
Session Overview

I. Background

II. What is a randomized experiment?

III. Why randomize?

IV. Conclusions
III – WHY RANDOMIZE?
Why Randomize?- Conceptual Argument

If properly designed and conducted, randomized experiments 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 experiment,

any difference that subsequently arises between them can be attributed to the program rather than to other factors.
Example #1 - Pratham’s Read India program
Example #1 - Pratham’s Read India program

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>0.60*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-0.90*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>0.31*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>0.06</td>
</tr>
<tr>
<td>(5) Randomized Experiment</td>
<td></td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
Example #1 – Pratham’s Read India program

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>0.60*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-0.90*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>0.31*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td>0.88*</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
Example #2: A voting campaign in the USA

Courtesy of Flickuser theocean
A voting campaign in the USA

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact (Vote %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>-7.2 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>3.8 pp*</td>
</tr>
<tr>
<td>(4) Multiple Regression</td>
<td>6.1 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td></td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
## A voting campaign in the USA

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact (Vote %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>-7.2 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>3.8 pp*</td>
</tr>
<tr>
<td>(4) Multiple Regression</td>
<td>6.1 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td>0.4 pp</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
## A voting campaign in the USA

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact (Vote %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>-7.2 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>3.8 pp*</td>
</tr>
<tr>
<td>(4) Multiple Regression</td>
<td>6.1 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td><strong>0.4 pp</strong></td>
</tr>
</tbody>
</table>

*: 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: If properly designed and conducted, randomized experiments provide the most credible method to estimate the impact of a program
  - Empirical argument: Different methods can generate different impact estimates
What is the most convincing argument you have heard against RCTs? Enter your top 3 choices.

A. Too expensive
B. Takes too long
C. Not ethical
D. Too difficult to design/implement
E. Not externally valid (Not generalizable)
F. Less practical to implement than other methods and not much better
G. Can tell us what the impact is, but not why or how it occurred (i.e. it is a black box)
THANK YOU!
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
Why Randomize? Backup Slides

Dan Levy
Harvard Kennedy School
What is the most convincing argument you have heard against RCTs? Enter your top 3 choices.

A. Too expensive
B. Takes too long
C. Not ethical
D. Too difficult to design/implement
E. Not externally valid (Not generalizable)
F. Less practical to implement than other methods and not much better
G. Can tell us what the impact is, 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 Experiment
1 - Pre-post (Before vs. After)

- Look at average change in test scores over the school year for the balsakhi children
QUESTION: Under what conditions can this difference (26.42) be interpreted as the impact of the balsakhi program?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average post-test score for children with a balsakhi</td>
<td>51.22</td>
</tr>
<tr>
<td>Average pretest score for children with a balsakhi</td>
<td>24.80</td>
</tr>
<tr>
<td>Difference</td>
<td>26.42</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Average score for children with a balsakhi</th>
<th>51.22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average score for children without a balsakhi</td>
<td>56.27</td>
</tr>
<tr>
<td>Difference</td>
<td>-5.05</td>
</tr>
</tbody>
</table>

**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...

With gains in test scores of...

Children who got balsakhi

Children who did not get balsakhi
3 – Difference-in-difference

<table>
<thead>
<tr>
<th>Average score for children with a balsakhi</th>
<th>Pretest</th>
<th>Post-test</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.80</td>
<td>51.22</td>
<td>26.42</td>
</tr>
</tbody>
</table>

**Question:** Under what conditions can this difference (-5.05) be interpreted as the impact of the balsakhi program?
### 3 – Difference-in-difference

<table>
<thead>
<tr>
<th>Average score for children with a balsakhi</th>
<th>Pretest</th>
<th>Post-test</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.80</td>
<td>51.22</td>
<td>26.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Average score for children without a balsakhi</th>
<th>Pretest</th>
<th>Post-test</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>36.67</td>
<td>56.27</td>
<td>19.60</td>
</tr>
</tbody>
</table>
### 3 – Difference-in-difference

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th>Post-test</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average score for children with a balsakhi</strong></td>
<td>24.80</td>
<td>51.22</td>
<td>26.42</td>
</tr>
<tr>
<td><strong>Average score for children without a balsakhi</strong></td>
<td>36.67</td>
<td>56.27</td>
<td>19.60</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td></td>
<td></td>
<td><strong>6.82</strong></td>
</tr>
</tbody>
</table>
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 Experiment

• Suppose we evaluated the balsakhi program using a randomized experiment

• 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?

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-post</td>
<td>26.42*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-5.05*</td>
</tr>
<tr>
<td>(3) Difference-in-Difference</td>
<td>6.82*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>1.92</td>
</tr>
</tbody>
</table>

*: 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>26.42*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-5.05*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>6.82*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>1.92</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td><strong>5.87</strong>*</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
# Impact of Balsakhi - Summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>26.42*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-5.05*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>6.82*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>1.92</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td><strong>5.87</strong>*</td>
</tr>
</tbody>
</table>

*: 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>0.60*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-0.90*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>0.31*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td></td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
Example #2 - Pratham’s Read India program

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>0.60*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-0.90*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>0.31*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>0.06</td>
</tr>
<tr>
<td>(5) Randomized Experiment</td>
<td>0.88*</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
Example #3: A voting campaign in the USA
## A voting campaign in the USA

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact (Vote %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>-7.2 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>3.8 pp*</td>
</tr>
<tr>
<td>(4) Multiple Regression</td>
<td>6.1 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td>(5) Randomized Experiment</td>
<td></td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
### A voting campaign in the USA

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact (Vote %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>-7.2 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>3.8 pp*</td>
</tr>
<tr>
<td>(4) Multiple Regression</td>
<td>6.1 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td>(5) Randomized Experiment</td>
<td>0.4 pp</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level
THANK YOU!
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

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>26.42*</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>-5.05*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>6.82*</td>
</tr>
<tr>
<td>(4) Regression</td>
<td>1.92</td>
</tr>
<tr>
<td><strong>(5) Randomized Experiment</strong></td>
<td>5.87*</td>
</tr>
</tbody>
</table>

*: Statistically significant at the 5% level