Value And Benefits Of Evaluation

Juliette Keeley & Rohit Naimpally
J-PAL North America
I. A message from our sponsors

II. Setting the stage

III. Why evaluate

IV. Why randomize

V. Why it matters
I. A message from our sponsors

II. Setting the stage

III. Why evaluate

IV. Why randomize

V. Why it matters
**J-PAL’s Mission** is to ensure that policy is driven by evidence and research is translated into action.

**Evaluations**

J-PAL researchers conduct randomized evaluations to test and improve the effectiveness of programs and policies aimed at reducing poverty.

**Capacity Building**

Through training courses, evidence workshops, and research projects, J-PAL equips policymakers and practitioners with the expertise to carry out their own rigorous evaluations.

**Policy Outreach**

J-PAL affiliates and staff analyze and disseminate research results and build partnerships with policymakers to ensure policy is driven by evidence and effective programs are scaled up.

www.povertyactionlab.org
860 ongoing and completed projects in 80 countries
300 million+ lives touched by the scale up of proven programs

J-PAL NORTH AMERICA
Massachusetts Institute of Technology (MIT)
Cambridge, USA

J-PAL GLOBAL
Massachusetts Institute of Technology (MIT)
Cambridge, USA

J-PAL EUROPE
Paris School of Economics
Paris, France

J-PAL SOUTH ASIA
Institute for Financial Management Research (IFMR)
Chennai and New Delhi, India

J-PAL LATIN AMERICA
Pontifica Universidad Catolica de Chile
Santiago, Chile

J-PAL AFRICA
University of Cape Town
Cape Town, South Africa

J-PAL SOUTHEAST ASIA
University of Indonesia
Jakarta, Indonesia

J-PAL Research Projects by Sector
- Agriculture
- Crime, Violence, & Conflict
- Education
- Environment & Energy
- Finance
- Health
- Labor Markets
- Political Economy & Governance
I. A message from our sponsors

II. Setting the stage

III. Why evaluate

IV. Why randomize

V. Why it matters
Problem

High rates of violence in American cities

- Homicide and violent crime rates are far higher for disadvantaged youth, especially young men of color, than their peers.


Source: Gun Violence in Chicago, 2016, University of Chicago Crime Lab
What are the causes of violent crime?
Proposed solution

Cognitive behavioral therapy (CBT)

CBT encourages individuals to examine their automatic thought processes and responses.
Clear, credible results

During the program year...

Participants attended, on average, 13 of the one-hour sessions

44% reduction in violent crime arrests
- 18 arrests per 100 youth → 8 arrests per 100 youth

31% reduction in overall arrests
- 52 arrests per 100 youth → 36 arrests per 100 youth

Why? Follow-up experiment suggests the program helps students to slow automatic responses
I. A message from our sponsors

II. Setting the stage

III. Why evaluate

IV. Why randomize

V. Why it matters
What is evaluation?
Program evaluation

Evaluation

Program Evaluation

Impact Evaluation
# Components of program evaluation

<table>
<thead>
<tr>
<th>Component</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Needs Assessment</td>
<td>What is the problem?</td>
</tr>
<tr>
<td>Theory of Change</td>
<td>How, in theory, does the program fix the problem?</td>
</tr>
<tr>
<td>Process Evaluation</td>
<td>Does the program work as planned?</td>
</tr>
<tr>
<td>Impact Evaluation</td>
<td>What is the causal effect of the program on outcomes?</td>
</tr>
<tr>
<td>Cost-Effectiveness Analysis</td>
<td>Given the cost, how does it compare to alternatives?</td>
</tr>
</tbody>
</table>
Implementation failure vs. theory failure

Successful intervention

- Inputs
- Activities
- Outputs
- Outcomes
- Goal

Implementation failure

- Inputs
- Activities
- Outputs
- Outcomes
- Goal

Theory failure

- Inputs
- Activities
- Outputs
- Outcomes
- Goal
How to measure impact?

**Impact** is defined as a comparison between:

- **What actually happened** and
- **What would have happened**, had the program not been introduced (i.e., the “counterfactual”)

J-PAL | VALUE AND BENEFITS OF EVALUATION
What is the impact of this program?

- Test scores vs. Time
- Program starts
- Impact
- Counterfactual
Impact: What is it?

Program starts

Test scores

Time

Counterfactual

Impact
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 to the program (and not to other factors).
Methods as tools

- Pre-post
- Simple Difference
- Difference-in-Difference
- Regressions
- Randomized evaluation
I. A message from our sponsors

II. Setting the stage

III. Why evaluate

IV. Why randomize

V. Why it matters
Evaluating a voting campaign
Program “Get Out the Vote”

**Problem:** Low voter turnout

**Proposed solution:** “Get Out the Vote” program

- Compile a list of registered voters with listed phone numbers
- Call a sample of individuals in this list
- In this phone call, responder is encouraged to vote
Program “Get Out the Vote”

What is the impact of the “Get Out the Vote” program on voter turnout rate?
1. Pre-Post

Compare the “before” voter turnout of individuals who were called... to the voter turnout of the same group of individuals after the call.
1. Pre-Post

Among individuals who were called...

<table>
<thead>
<tr>
<th>Average <strong>pre-call</strong> voter turnout (1998)</th>
<th>46.6 percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average <strong>post-call</strong> voter turnout (2002)</td>
<td>64.5 percent</td>
</tr>
<tr>
<td>Difference</td>
<td>+17.9 percentage points</td>
</tr>
</tbody>
</table>

Under what conditions can this difference (17.9 pp) be interpreted as the impact of the call?
1. Pre-Post

What would have happened without the call?

Impact = 17.9?

Counterfactual

1998  2002

17.9 pp?
2. Simple Difference

Compare the voter turnout of individuals who were called...

to the voter turnout of individuals who were not called.

After
2. Simple Difference

After election day in 2002...

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average voter turnout for individuals who were <strong>called</strong></td>
<td>64.5 percent</td>
</tr>
<tr>
<td>Average voter turnout for individuals who were <strong>not called</strong></td>
<td>53.6 percent</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td><strong>+10.8 percentage points</strong></td>
</tr>
</tbody>
</table>

Under what conditions can this difference (10.8 pp) be interpreted as the impact of the call?
2. Simple Difference

What would have happened without the call?

Impact = 10.8 pp?

Counterfactual

1998  2002
3. Difference-in-Differences

Compare the **change** in voter turnout of individuals who **were called**...

to the **change** in voter turnout of individuals who **were not called**.

Before | After
---|---

Called | Not called

User by Mert Güler from the Noun Project
Telephone by ProSymbols from the Noun Project
### 3. Difference-in-Differences

<table>
<thead>
<tr>
<th></th>
<th>1998</th>
<th>2002</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average voter turnout for individuals who were <strong>called</strong></td>
<td>46.6 %</td>
<td>64.5 %</td>
<td>17.9 pp</td>
</tr>
<tr>
<td>Average voter turnout for individuals who were <strong>not called</strong></td>
<td>37.6 %</td>
<td>53.6 %</td>
<td>16.0 pp</td>
</tr>
<tr>
<td>Difference-in-Differences</td>
<td>17.9-16.0=</td>
<td>+1.9 pp</td>
<td></td>
</tr>
</tbody>
</table>

Under what conditions can this difference in the differences (1.9 pp) be interpreted as the impact of the call?
3. Difference-in-Differences

What would have happened without the call?

Impact = 1.9 pp?
Other non-experimental methods

1. Regression
2. Matching
3. Instrumental Variables
4. Regression Discontinuity
### Evaluating a voting campaign

<table>
<thead>
<tr>
<th>Method</th>
<th>Impact (Vote %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Pre-Post</td>
<td>17.9 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>1.9 pp*</td>
</tr>
<tr>
<td>(4) Multivariate Regression</td>
<td>4.6 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td>(6) Randomized Experiment</td>
<td></td>
</tr>
</tbody>
</table>

* Statistically significant at the 5% level
Randomized evaluation

Population is split into 2 groups by random lot

INTERVENTION

Outcomes for both groups are measured

CONTROL
# 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>17.9 pp</td>
</tr>
<tr>
<td>(2) Simple Difference</td>
<td>10.8 pp*</td>
</tr>
<tr>
<td>(3) Difference-in-Differences</td>
<td>1.9 pp*</td>
</tr>
<tr>
<td>(4) Multivariate Regression</td>
<td>4.6 pp*</td>
</tr>
<tr>
<td>(5) Matching</td>
<td>2.8 pp*</td>
</tr>
<tr>
<td><strong>(6) Randomized Evaluation</strong></td>
<td><strong>0.4 pp</strong></td>
</tr>
</tbody>
</table>

* Statistically significant at the 5% level
Why use randomization?

Statistically, random assignment makes it very likely that we are making an apples-to-apples comparison.
I. A message from our sponsors

II. Setting the stage

III. Why evaluate

IV. Why randomize

V. Why it matters
So what was the policy impact of the BAM evaluation?
Policy attention

Chicago Mayor **Rahm Emmanuel** increased funding for BAM across the city after the first study.

Former President **Barack Obama** made federal funding available to expand BAM and mentioned BAM at the launch of the “My Brother’s Keeper” initiative.
Questions?

Feel free to reach us at:

jkeeley@mit.edu and rohit_n@mit.edu
Appendix: When are RCTs a good option?
Is an RCT the right evaluation method?

<table>
<thead>
<tr>
<th></th>
<th>No RCT</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing evidence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size and capacity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maturity of the program</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feasibility</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Is an RCT the right evaluation method?**

<table>
<thead>
<tr>
<th></th>
<th><strong>No RCT</strong></th>
<th><strong>RCT</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing evidence</strong></td>
<td>• Rigorous evidence exists already</td>
<td>• No existing rigorous evidence</td>
</tr>
<tr>
<td><strong>Sample size and capacity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Maturity of the program</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Feasibility</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Is an RCT the right evaluation method?

<table>
<thead>
<tr>
<th></th>
<th>No RCT</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing evidence</strong></td>
<td>• Rigorous evidence exists already</td>
<td>• No existing rigorous evidence</td>
</tr>
<tr>
<td><strong>Sample size and capacity</strong></td>
<td>• Program is already rolled out (no room for expansion)</td>
<td>• Program can be expanded (capacity)</td>
</tr>
<tr>
<td></td>
<td>• All eligible individuals are already being served</td>
<td>• Not all eligible individuals are served</td>
</tr>
<tr>
<td></td>
<td>• Program is too small for detectable impact</td>
<td>• Program size and take-up are sufficiently high (power)</td>
</tr>
<tr>
<td><strong>Maturity of the program</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Feasibility</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Is an RCT the right evaluation method?

<table>
<thead>
<tr>
<th></th>
<th>No RCT</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing evidence</strong></td>
<td>• Rigorous evidence exists already</td>
<td>• No existing rigorous evidence</td>
</tr>
<tr>
<td><strong>Sample size and capacity</strong></td>
<td>• Program is already rolled out (no room for expansion)</td>
<td>• Program can be expanded (capacity)</td>
</tr>
<tr>
<td></td>
<td>• All eligible individuals are already being served</td>
<td>• Not all eligible individuals are served</td>
</tr>
<tr>
<td></td>
<td>• Program is too small for detectable impact</td>
<td>• Program size and take-up are sufficiently high (power)</td>
</tr>
<tr>
<td><strong>Maturity of the program</strong></td>
<td>• Program is too young or underdeveloped</td>
<td>• Program is well established, but has not been tested</td>
</tr>
<tr>
<td><strong>Feasibility</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is an RCT the right evaluation method?

<table>
<thead>
<tr>
<th></th>
<th>No RCT</th>
<th>RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing evidence</strong></td>
<td>• Rigorous evidence exists already</td>
<td>• No existing rigorous evidence</td>
</tr>
<tr>
<td><strong>Sample size and capacity</strong></td>
<td>• Program is already rolled out (no room for expansion)</td>
<td>• Program can be expanded (capacity)</td>
</tr>
<tr>
<td></td>
<td>• All eligible individuals are already being served</td>
<td>• Not all eligible individuals are served</td>
</tr>
<tr>
<td></td>
<td>• Program is too small for detectable impact</td>
<td>• Program size and take-up are sufficiently high (power)</td>
</tr>
<tr>
<td><strong>Maturity of the program</strong></td>
<td>• Program is too young or underdeveloped</td>
<td>• Program is well established, but has not been tested</td>
</tr>
<tr>
<td><strong>Feasibility</strong></td>
<td>• No good data available</td>
<td>• Access to (reliable) data</td>
</tr>
<tr>
<td></td>
<td>• Ethical concerns</td>
<td>• Value exceeds expected costs</td>
</tr>
</tbody>
</table>
When to consider randomization

• New programs
• Program or service is over-subscribed
  – If you can’t serve everyone, what is the optimal way to allocate spots?
• Program expansion
  – e.g. moving into a new location, service area, or target population
• Adding a new component/service
  – Can you roll out the new component to some people and not others?
• Program thresholds/cutoffs
  – Those just below the cutoff could be randomly admitted