WELCOME

Social Experiments in Practice:
The Why, When, Where, and How of Experimental Design & Analysis
MEMBER FORUM | WASHINGTON, DC

Session 1: On the Frontier of the “Why” and “When” of Social Experiments: Doing the Right Thing at the Right Time

Social Experiments in Practice:
The Why, When, Where, and How of Experimental Design & Analysis
MEMBER FORUM | WASHINGTON, DC
The Forty-Five Year History

- First used to estimate behavioral response to incentives in a handful of federal research projects
- Expanded to assess effects of:
  - New interventions
  - Existing programs
- Augmented individual-level random assignment with site-level
Lessons from Practice

- Proven to be ethical and feasible in an expanding and diverse group of settings
- But in any *particular case*, this leaves open:
  - Is it ethical?
  - Is it feasible?
- What historical and current practice shows
- Bottom line
  - Boundary between hard and malleable barriers not fixed
  - Area of what’s possible continues to expand

Ethical Barriers

- Normative
  - Is it right?
  - Will an IRB see it that way?
- Pragmatic
  - Can we convince programs to participate?
  - Will the broader group of stakeholders cooperate?
Denying Entitlements

- Entitlement benefits often have intrinsic worth
- Purpose often not just distributional but to affect behavior
- Usually illegal to deny, but sometimes waivers can permit
- A barrier we should not surmount

Alternative Entitlement Rules

- Tests of more generous policies
- Tests that create winners and losers
- Feasibility: implications of informed consent
- Public benefits/services waiver of informed consent
- Conditioning entitlements on stricter requirements
- Public benefits/services exemption
- Difficulty drawing a sharp line
Evaluating Existing Services

- Avoiding reduction in number served
- Fairness of a lottery in allocating a limited resource, (i.e., not an entitlement)
- Provides a strong basis for engaging programs and stakeholders
- Implies over-recruitment
- Most often beyond current level

Implications of Over-Recruitment

- **Ethics**: disappointment of those who otherwise would not have been denied (a relatively small harm)
- **Ethics**: sometimes has beneficial effects, such as greater exposure of eligible individuals to possibility of treatment
- **Feasibility**: some programs may have difficulty doing so (says something about their potential reach)
- **Substance**: may result in a somewhat different population served (can be addressed with baseline identification of those who would have been served)
- Barriers created and overcome by developing practice
Previously Vexing Barriers

- **Saturation programs**: All eligible expected to participate
- **Entry effects**: Community-wide messages expected to affect who would appear for randomization
- Would-be control group members interact with treatment group members, potentially altering relevant outcomes of both

Cluster Random Assignment

- Last decade has proven feasibility
- Large and growing body of school-level and community-level in U.S. and internationally
- Can remove contamination problem
- Can capture entry effects
- More generally captures effects of social interactions that are part of the treatment
Practical, but Often Surmountable, Barriers

- Less efficient statistically and can be more costly
  - Depending on design, can be cheaper
- Can be difficult for sites to accept what they will implement being based on a random draw, and remain committed to it
  - Sometimes can offer delayed treatment for control sites
- Site development may alter outcomes for control group
  - May still have large T/C service differential

Individual Random Assignment as a Fallback

- Often, despite initial claims, treating all eligibles in a site isn’t possible once available resources are considered
- In such cases, random assignment of individuals all of whom can be treated (saturated) or to a control group can work
- However, this won’t work in some others, (e.g., messages are broadcast on television)
Individual Random Assignment as a Fallback

- Programmatically individual-level might be a good strategy
- Integration of services often promoted to improve outcomes, but very little evidence supports it
- Might be wiser to accomplish it initially for a subset of individuals and estimate effects using individual random assignment
- Changing the question

Cost

- Ranges from very little to a great deal
- Surveys are biggest cost drivers
- Administrative data a big saver
- **All else equal**, experimental studies less data-dependent and more efficient than non-experimental
- Random assignment itself a secondary cost in individual-level, but virtually nothing in cluster
- Recruitment cost significant, but may be lower in future
Concluding Thoughts

- Use of random assignment has grown greatly in function, areas of social policy and numbers
- Some barriers fixed
- Developing practice both creates and overcomes barriers
- Every reason to believe expansion of the boundaries of the possible will continue
- Other sessions to provide further examples

For more information, please contact:

**Howard Rolston**
Principal Associate
Social & Economic Policy Division
T: 301.634.1820
E: Howard_Rolston@abtassoc.com
Session 1 at “Social Experiments in Practice: The Why, When, Where, and How of Experimental Design & Analysis”

When is a Program Ready to be Evaluated?

Presented by:

Jacob Alex Klerman
APPAM-Abt Institutional Member Forum
Washington, DC

Outline

- The Challenge
- Falsifiable Logic Model
- Discussion
“Randomistas Rule”

- We only want to fund programs that work
  - But, most demonstrations are shown not to work

- So, we require a “rigorous impact evaluation” (RIE)
  - Usually random assignment

- Leading to a strategy of the form …
Randomistas OVER Rule?

- But, this must be—and has been—a frustrating strategy
  - RIE is expensive and has long timelines
  - And, most programs will be found not to work

Is It Possible To Do Better?

- Is it possible to know: When is a program ready to be evaluated?
  - i.e., (more) likely to work?

- Apparently, not …
  - This looks like an “impact question”
  - But, the premise of this “Forum” is that determining impact requires a “social experiment”
  - And, nothing else will do!
Yes: Pilot

- Is it possible to know: *When is a program ready to be evaluated?*
  - i.e., (more) likely to work?
- Apparently, not …
  - This looks like an “impact question”
  - But, the premise of this “Forum” is that determining impact requires a “social experiment”
  - And, nothing else will do!

Outline

- The Challenge
- *Falsifiable Logic Model*
- Discussion
Logic Model for Training Welders

**Inputs**
- Staff
- Space
- Materials
- Partnerships

**Outputs**
- Sessions held
- Teaching with fidelity
- Internships

**Outcomes**

<table>
<thead>
<tr>
<th>Short-Term</th>
<th>Long-Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attend sessions</td>
<td>Earnings</td>
</tr>
<tr>
<td>Graduate</td>
<td>Transfer program participation</td>
</tr>
<tr>
<td>Pass external exam</td>
<td>Taxes paid</td>
</tr>
<tr>
<td>Employed in target industry</td>
<td></td>
</tr>
</tbody>
</table>

Estimating “Impact” Is Hard

**Inputs**
- Staff
- Space
- Materials
- Partnerships

**Outputs**
- Sessions held
- Teaching with fidelity
- Internships

**Outcomes**

<table>
<thead>
<tr>
<th>Short-Term</th>
<th>Long-Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attend sessions</td>
<td>Earnings</td>
</tr>
<tr>
<td>Graduate</td>
<td>Transfer program participation</td>
</tr>
<tr>
<td>Pass external exam</td>
<td>Taxes paid</td>
</tr>
<tr>
<td>Employed in target industry</td>
<td></td>
</tr>
</tbody>
</table>

“Impact”: Treatment vs. Control for “Long-Term Outcomes”
- Estimating impact requires: (i) random assignment, (ii) measurement for T and C, (iii) measurement well after program ends

*Making RIE long and expensive*
A Complementary Approach: Exploit Earlier Steps of Logic Model

- Logic model posits that these “inputs, outputs, and short-term outcomes” are necessary for impacts on long-term outcomes

Earlier step are necessary; they may not be sufficient
A process evaluation of a pilot can verify that the “inputs, outputs, and short-term outcomes” posited by the logic model actually occur

- Process evaluation does *not* require: (i) random assignment, (ii) measurement for C, (iii) measurement well after program ends

---

Piloting Is Only Useful If …

- Appropriate “inputs, outputs, and short-term outcomes” exist
  - And failures can be detected quickly and cheaply

- These “inputs, outputs, and short-term outcomes” discriminate; i.e.,
  - Some programs “pass” their own logic model,
  - (many) Other programs “fail” their own logic model
Piloting Is Only Useful If …

- Appropriate “inputs, outputs, and short-term outcomes” exist
  - And failures can be detected quickly and cheaply
- These “inputs, outputs, and short-term outcomes” discriminate; i.e.,
  - Some programs “pass” their own logic model
  - (many) Other programs “fail” their own logic model

Surprisingly often, both of those conditions are satisfied

But, utility of this strategy appears to vary by program type

Examples of Logic Model Failure

1. Inputs: staff, space, partnerships
2. Enough participants
3. Program completion
4. Insufficient fidelity
5. Pre/post progress
6. Pass external exams
7. Employment in targeted industry

Program details matter; examples have caveats
i.e., Such Outcomes Exist

- These failures could reasonably have been specified as an “input, output or short-term outcome” in a (falsifiable) logic model
- Instead, measurement need not be expensive
  - Sometimes conventional process evaluation (e.g., partnerships, fidelity of implementation)
  - Sometimes in program records (e.g., initial enrollment rates, attendance rates, pre/post tests)
  - Sometimes immediate post-program follow-up of participants (e.g., external pass rates, employment in targeted industry)

Outline

- The Challenge
- Falsifiable Logic Model
- Discussion
“Randomistas” Process

Program Idea

Rigorous Impact Evaluation (RIE) “Toll Gate”

Broad Program Implementation

Revised Process

Program Idea

Pilot

Formative Evaluation (to improve satisfaction of LM)

Process Evaluation (to verify satisfaction of LM)

Logic Model “Toll Gate”

Only programs that pass their own logic model proceed to RIE
Leading to Higher Success Rates at RIE through …

1. **Logic Model** → improves program at initial pilot
2. **Formative Evaluation** → improves program at subsequent pilot
3. **Process Evaluation** → winnows out programs that do not satisfy their own logic model

Only programs that pass their own logic model proceed to RIE

Two Caveats

1. **Logic Models**: Requires a falsifiable logic model
2. **Elapsed Time**: Total time through the “RIE Tollgate” increases substantially
For More on These Ideas


For more information, please contact:

**Jacob Alex Klerman**
Principal Associate/Scientist & Abt Senior Fellow
Social & Economic Policy Division
T: 617.520.2613
E: Jacob_Klerman@abtassoc.com