Session 3: On The Frontier of the “How” of Social Experiments: Getting Inside the Black Box

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

Opportunities for Implementation Research

Lawrence M. Mead
Department of Politics
New York University
Forms of implementation research:

- **High-tech:**
  - As part of an experimental evaluation.
  - Uses analysis of evaluation data.

- **Low-tech:**
  - Outside of evaluations but supportive of them.
  - Uses field interviewing and analysis of program data.

High-tech implementation research:

- Seeks to enrich experimental findings through additional analyses.
- Multiple treatment groups, as in the NEWWS evaluation.
- Analyzing data across sites, as in Bloom et al., “Linking Program Implementation and Effectiveness.”
- Questions about feasibility.
Low-tech implementation research: supports evaluation through:

- Establishing evaluability.
- Performance analysis.
- Generating new ideas.

Establishing evaluability:

- Implementation research is needed to show that a program is ready to evaluate.
- Implementation problems weakened Great Society programs.
- Stronger implementation helped welfare reform succeed.
- Uses interviews and program data to test what a program means “on the ground.”
Performance analysis:

- Going beyond evaluability to connect program policies to performance.
- Staff say in interviews what practices they think maximize performance.
- One tests those claims by modeling performance measures with program data on client services plus controls.
- Not an impact finding but suggests policies to test with evaluations.

Generating new ideas:

- Evaluators need plausible programs to test.
- Implementation research often generates new ideas.
- Inquiry into one program dimension may suggest the importance of others.
- Ideas are then tested against program data.
- How research on welfare work programs uncovered the importance of participation requirements.
New ideas I had in the field:

- Effective welfare work programs must:
  - Require participation.
  - Be demanding yet positive toward clients.
  - Supervise client activities closely.
- Work programs for poor men can be financed out of child support collections.
- Some—not all—of these ideas have checked out against program data.

Policy learning depends on both evaluation and implementation:

- Evaluations are rigorous but rigid, can test only established ideas.
- Same for high-tech implementation research.
- Low-tech implementation research is less rigorous but more open to serendipity.
- The two styles are complementary.
- Data analysis is not enough—we must also lay hands on the institutions.
On the Frontier of the “How” of Social Experiments: Getting Inside the Black Box

Innovations in Experimental Impact Analysis

Presented by: Laura R. Peck

APPAM-Abt Institutional Member Forum
Washington, DC
The Randomized Experiment

- Random assignment to T & C groups creates condition for inferring causality.
- Difference in T & C outcomes is impact of being in treatment group.
- Within the treatment group:
  - No-shows
  - Partial participation
  - Full participation

These are not randomized to, so standard “threats to validity” apply: selection bias, maturation, history, etc.

*Post-hoc* analyses can build on experimental design to say something about impacts on these “subgroups.”

Treatment Heterogeneity

- **Endogenous subgroups**
  - Participation (take-up; adjusting for no-shows)
  - Potential effects on no-shows
  - Treatment dosage or quality
  - Multi-faceted treatment components
  - Multi-faceted treatment group characteristics
  - Control group “what-ifs”

- **Research questions**
  - Class 1 (simple) = take-up, participation
  - Class 2 (complex) = dosage, quality, program mediators, individual mediators, counterfactual conditions
Analytic Approaches…

…that use experimental data

- Instrumental Variables (IV)
- Analysis of Symmetrically-Predicted Endogenous Subgroups (ASPES)
- Cluster Analysis

Instrumental Variables

- Use random assignment as an “instrument” for Program take-up
- Interact random assignment with site to create an instrument for Program characteristics
- Assumption of “exclusion restriction” requires that a lone pathway influences outcomes
Bonuses of Randomization

Propensity Score-Based Subgroups

- Propensity score-based subgroups, identified within experimental evaluation data, compare treatment subgroup members to their predicted counterparts in the control group.

- Avoid bias introduced by differential T-C subgroup identification by using out-of-sample prediction.
Analysis of Symmetrically-Predicted Endogenous Subgroups (ASPES)

- Step 1: Use baseline data to predict subgroup.
- Step 2: Analyze treatment effect.
### Analysis of Symmetrically-Predicted Endogenous Subgroups (ASPES)

- **Step 1:** Use baseline data to predict subgroup.
- **Step 2:** Analyze treatment effect.
- **Step 3:** Convert from predicted to actual, by assumption.

### Cluster Analysis

- ASPES considers cleanly identifiable groups: none/some (dosage), low/high (quality).
- Extension to cluster analysis allows baseline data’s patterns to reveal complex endogenous subgroups.
Cluster Analysis

- ASPES considers cleanly identifiable groups: none/some (dosage), low/high (quality).
- Extension to cluster analysis allows baseline data’s patterns to reveal complex endogenous subgroups.
Cluster Analysis Example

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Participation (%)</th>
<th>Earnings Impact</th>
<th>Employment Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>49.2%</td>
<td>$4,393</td>
<td>10.3%</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>39.7%</td>
<td>$1,771</td>
<td>6.0%</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>24.9%</td>
<td>$3,041</td>
<td>22.1%</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>17.0%</td>
<td>$2,408</td>
<td>17.8%</td>
</tr>
</tbody>
</table>

Cluster Analysis

- **Internal validity = strong**
  - Differences in T and C subgroups’ mean outcomes are unbiased impact estimate for that subgroup
- **External validity = limited**
  - Generalizing to complex aggregate profiles that may or may not represent specific, real individuals
Conclusion

- Build from randomized experimental design to expose what’s in the black box.

- New Frontiers
  - Guidelines for when/where to apply these tools
  - Project design considerations (research questions, data, sample)
  - Substitute study *design* options for these analytic innovations
  - Build on new study *design* options

For more information, please contact:

*Laura R. Peck*
Principal Scientist
Social & Economic Policy Division
T: 301.347.5537
E: Laura_Peck@abtassoc.com
On the Frontier of the “How” of Social Experiments: Getting Inside the Black Box

Innovations in Experimental Impact Design

Presented by:
Stephen H. Bell
APPAM-Abt Institutional Member Forum
Washington, DC

Research Question and Outline

What random assignment designs best equip analysts to measure the contribution of individual intervention components to impact magnitude?

- Working within the textbook design
- Tweaking individual-level randomization to learn more
- Examining other, even “fancier” approaches
- Randomizing sites
- Caveats and future possibilities
Randomize Individuals to Treatment and Control Status within Sites

- Not designed to reveal experimentally the contributions of intervention features to impact magnitude
- Advantage: Unbiased measures of the impact of different “packages” of program features ➔ trying to attribute the right thing to the causal features of the intervention
- Disadvantage: Findings biased by confounding factors that vary across sites
  - Non-intervention factors that . . .
  - Affect impact magnitude and . . .
  - Are not controlled in the analysis
- Advice: Get dozens of sites, to try to model confounders

Randomize Individuals to Intervention Variants (+ Control) within Site

- Three-arm random assignment
  - T = program without focal component
  - T+ = program with focal component
  - C = control group (no program)
- T+ vs. T isolates the impact of the focal component (“+”), while C arm reveals impacts vs. status quo (as in “textbook” design)
- Advantage: Answers multiple questions experimentally, including contribution of the focal component
- Disadvantage: Smaller samples for each Q . . . or higher costs!
- Advice: Why not do it, if Ns and $$ permit!
What Else Can Be Done with 3-Arm Designs . . . Or with 2-Arm Ones?

- Can natural variation be combined with experimental variation to learn about non-randomized components?
  - Yes, by using the experimental-variation sample to validate and improve methods used to analyze natural variation

- Is a control group always needed?
  - Not if the policy decision concerns the choice between Program A and Program B and “no program” is off the table
  - Strategy is fine if findings show A > B and have confidence that B does not do harm (on other grounds) \( \Rightarrow \) choose A
  - What if findings show neither A > B nor B > A? Maybe neither one works . . . i.e., improves on “no program”

Do Fancier Approaches Help? (p. 1)
Factorial or “Matrix” Designs

- “Cross-tabs” of multiple intervention components to form a grid of cells, with units randomized to components in all combinations

<table>
<thead>
<tr>
<th></th>
<th>Basic Service</th>
<th>Basic + One</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Limited</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Extended</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Do Fancier Approaches Help? (p. 2)  
Factorial or “Matrix” Designs

- Complete or partial “fractional” (empty cells)
- Equivalent to many-armed random assignment except for opportunity to look at “marginals” → answers Qs about the effects of the intervention components used to form cells
  - With larger Ns, but . . . .
  - Only in a mixture of contexts regarding other components
    → meaning of findings is murkier?

Do Fancier Approaches Help? (p. 3)  
Multi-Stage Random Assignment

- Eligibles
  - Phase A Complete
  - Phase B

<table>
<thead>
<tr>
<th>Effect of Phase A</th>
<th>T_{1A} + T_{1B} versus C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of Phase B</td>
<td>T_2 versus T_{1B}</td>
</tr>
<tr>
<td>Total Effect</td>
<td>T_{1A} + T_2 versus C</td>
</tr>
</tbody>
</table>
Do Fancier Approaches Help? (p. 4)  
Multi-Stage Random Assignment

- Preserves Ns while answering impact Qs about multiple intervention components . . . if components naturally occur sequentially
  - Randomize individuals to T/C (mostly Ts)
  - Let first component in sequence happen for Ts who chose it (Phase A)
  - Randomize those Ts into or out of second component
  - Rinse and repeat (Phase B, C, . . . )
- Example = Workforce Investment Act
- Shows incremental and cumulative impacts of the “cascade” of intervention components

Better Still: Randomize Sites, Not Individuals, to Program Variants

- At random, include a program feature of interest in some sites and not others—with everything else about the intervention and all other factors the same
- Advantage: Isolates contribution of the randomly varied factor (no confounding with other factors that vary by site)
- Disadvantage: Need many dozens of sites just to experimentally isolate the effect of one program feature with adequate power
- Advice: Consider only when have big budgets and big universes
Caveats and Coming Attractions

- More ambitious approaches should be tried... remembering that the above strategies are not mutually exclusive

- Crucial consideration = ability to deliver services in keeping with design ➔ achieving experimental compliance when implementing more than one treatment is a big challenge ➔ don’t over-reach

- Think through data analysis plans when making design decisions (innovate/extend on those, too!)

- Build policy knowledge from one study to the next through clever design, to home in on “what works”

For more information, please contact:

Stephen H. Bell
Vice President & Senior Fellow
Social & Economic Policy Division
T: 301.634.1721
E: Stephen_Bell@abtassoc.com