Obstacles to and Limitations of Social Experiments: 15 False Alarms

19 October, 2012

Stephen H. Bell & Laura R. Peck, Abt Associates

Competing Qualities of Field Experiments for Social Policy Evaluation

- Random assignment removes unknown confounders (selection bias)
- Concerns about
  - ethical appropriateness
  - scientific reliability
  - feasibility
  - cost
- With . . .

  creative design

  scrutiny of assumptions

  adequate commitment of funders

. . . many of these go away; they are “false alarms”

Overview of Talk

- Review concerns in the U.S. about limitations of field experiments as a method for measuring social policy/ program effectiveness

- Convey our reasons for believing these objections do not hold up in the American context

- Invite discussion about their importance in Europe

Possible Flaws Shared with Other Study Designs

- Missing outcome data
- Limited sample size (esp. for subgroup analyses)
- Inability to sort out causes of cross-site variation
- Findings not nationally representative

These challenges and limitations can face ***any*** impact study design, experimental or non-experimental

1. Ethical Concerns

- Unfair or unethical to deny services to control group
- If program is oversubscribed, have to limit inflow somehow
  - – lottery may be the fairest way (Orr, 1999)
- Doing an evaluation because ***don't know*** if the intervention helps ➔ can't presume Cs are hurt
- Examples of where Cs are ***helped*** compared to Ts: male youths in the National JTPA Evaluation (Orr et al., 1996)

Scientific Concerns Regarding Experiments

Does the experiment give scientifically unreliable estimates of impact? Several concerns here:

2. Impact on those assigned to, not on those who get the intervention
3. Control group is not a "no services" counterfactual
4. Control group members have easier access to alternative services than if there were no program
5. Treatment group's experience is distorted by change in program scale or population served
6. Eliminate selection bias only for the policy exposure controlled by randomization / care about other Qs

2. Impact on the Assigned, Not the Treated

- Matters if < 100% participation in the T group
- Make “no-show adjustment” – attributes entire T-C difference to the participants (Bloom, 1984)
- No assumptions about similarity of participants and non-participants, nor about similarity of either one to Cs
- Needed assumption = zero impact on non-participants
  – widely viewed as innocuous

3. Not “No Services” Counterfactual

- Multiple agencies supply similar services

- A given agency should be looking at “our services” versus “everything else that’s out there”, to justify its portfolio

- Were that agency not offering its particular intervention – and people could do just as well with services from other sources – the agency’s services are truly having no impact

  ➔ That’s what you want the findings to show

4. Cs Have Easier Access to Other Services

- With no program, all field experiment participants (Ts & Cs) would compete for assistance from available sources
- Cs face less competition as a result of Ts getting nto the focal program ➔ Cs get too much help ➔ impact estimate is biased downward (if effective)
- Not if policy decision is about program expansion/ contraction at the margin . . . or if other programs would expand to fill (at least part of) the void

  ➔ C group with full access to alternatives = better approximation of the ideal than one with no access

5. Change in Scale or Population Served

- Removing Cs necessarily results in
  - operating below capacity, or
  - broadening the population served

  Both could change impacts
- One option = broaden the population but have local staff identify applicants they would ordinarily have served (incentive is higher probability of T-group assignment)
  - do impacts for “ordinary” group
  - compare to impacts on “extras” (Olsen et al., 2007)
- Other option: spread Cs very thinly across many sites (e.g., Job Corps; Schochet et al., 2001)

6. Eliminates Selection Bias Only Once

- Experiments don’t strengthen inference about the consequences of program facets not randomized

- For example, can’t estimate impacts of different service sequences that emerge after random assignment without reverting to non-experimental methods and facing selection bias

- True . . . but other types of impact evaluations have all of these same problems plus one more (selection bias when estimating the ***main*** impact that randomization ***addresses***)

Feasibility Concerns Fixed with Adequate Funding

7. Saturation interventions affect entire communities
   - It's a big continent ➔ randomize communities
8. Programs that struggle to meet enrollment targets
   - Spread Cs thinly + fund added outreach
9. Need lifelong "treatment" to get full behavioral response
   - Fund the lifelong treatment
10. Low T group participation ➔ small average effects
    - Large samples reveal small effects
11. Programs/policies that pose questions of effectiveness in multiple areas
    - Multi-stage random assignment

(continued…)

Feasibility Concerns Fixed with Adequate Funding (continued)

12. Interventions with general equilibrium consequences beyond the experimental sample
    - General equilibrium analysis methods needed whether measure direct effects with experiments or with non-experimental methods

13. National / EU policy needs to be guided by representative findings
    - Field experiments with statistically representative samples of sites are feasible (e.g., Head Start; Puma et al., 2010)

14. Evidence of policy effectiveness needed quickly
    - Do regular, temporally overlapping experimental studies of existing programs ➔ new impact findings are always emerging

15. Experiments Are Too Expensive

- Need to consider costs compared to alternative studies tackling the same policy questions
- Obtaining broadly representative data on thousands of people with and without a policy is not cheap – whatever the study's design
- Exception = evaluations using existing large surveys; comparison group designs based on such data were the first ones discredited (LaLonde, 1986, Barnow, 1987)
- Low-cost experiments are possible (Baron, 2012)
- The costs of *failing* to do experiments is extremely high, if ineffective programs continue

Summary and Discussion

- Are field experiments sufficiently feasible to serve as foundation of social policy decision-making?

- View from the States
  - – "yes," scientifically, ethically, and operationally
  - – in an era of heightened fiscal accountability and results-focused policy making, it pays to ***fund*** the best possible impact evidence

  but some skeptics remain (e.g., Brickman & Reich, 2009)

- View from Europe: *? ? ?* *[discussion]*

Investigator Contact Information

Stephen H. Bell, Ph.D.

Senior Fellow
Abt Associates
Bethesda, Maryland, U.S.A.
(301) 634-1721
stephen_bell@abtassoc.com

Laura R. Peck, Ph.D.

Principal Scientist
Abt Associates
Bethesda, Maryland, U.S.A.
(301) 347-5537
laura_peck@abtassoc.com

Abt
ASSOCIATES

BOLD
THINKERS
DRIVING
REAL-WORLD
IMPACT