Design of Social Security Administration Demonstration Evaluations

Burt S. Barnow
George Washington University
David H. Greenberg
University of Maryland, Baltimore County

STATE OF THE SCIENCE MEETING

Evaluation Design in SSA Demonstrations

Non-Experimental

- 2 proof-of-concept →
- 2 impact analyses →

Experimental

- 11 classical experiments →
- 1 natural experiment →

Examples

- B.E.S.T., HSPD
- HOPE, Work Incentives Demo

- TETD, NetWork, AB, BOND, BPOD, MHTS, YTD, PROMISE, POD, DMIE, Timely Wage Reporting Nudge
- Ticket to Work

Population Representativeness

- Findings from evaluations in selected states or sites may not generalize to entire nation
- Findings from evaluations where volunteers participated may not generalize to mandatory programs
- Findings from evaluations focusing on selected conditions (e.g., diabetes, mental illness) may not generalize to other conditions
- Conclusion: *population representativeness* (of locations, kinds of people) cannot always be achieved, but should always be considered

Statistical Power and Effect Sizes

- In planning: some evaluated conducted a power analysis to determine sample size needed to detect impacts
- In practice, evaluations noted:
  - Too small a sample size means not being able to detect impacts as statistically significant (unless impacts are large)
  - Even with large enough overall sample, there might not be for subgroups
- Conclusion: power analysis is important to determining overall and subgroup sample sizes needed for an evaluation

Outcome Measures

- Employment and earnings (most commonly annual, but also relative to SGA level)
- Benefit receipt and amounts
- Time for processing applications (but not accuracy)
- Health outcomes (from self-assessments or existing scales)
- Varied other outcomes (e.g., hours, wages, income, quality of life, motivation)
- Conclusion: evaluations focused on most important outcomes; SSA may want to standardize some measures

Impact Estimation Issues

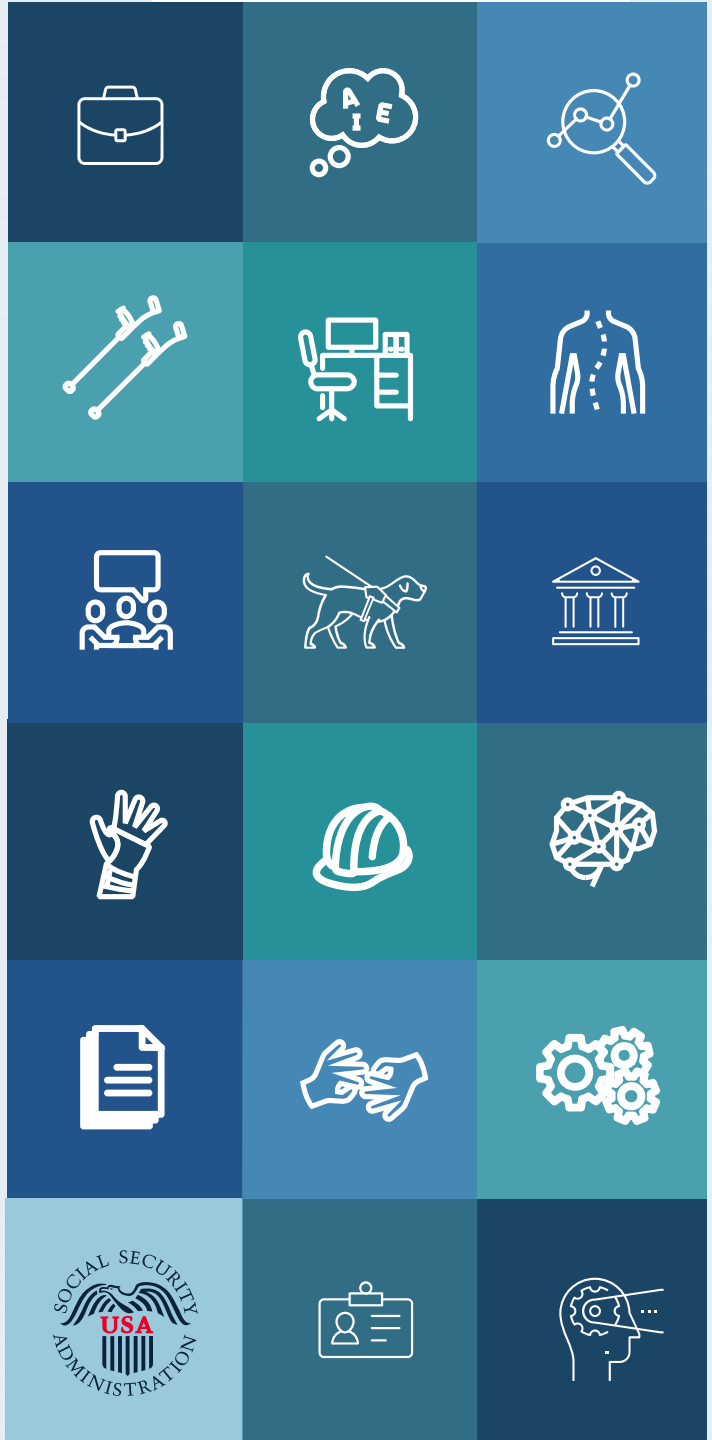
- Most evaluations used…
  - Intent to treat (ITT) approach
  - Ordinary least squares for continuous outcomes, logit for dichotomous outcomes, often weighted analysis
  - Data from survey and administrative sources
- Considerations:
  - Treatment on the treated (TOT), which requires more assumptions but sometimes answers important policy questions
  - Weighting strategies and sensitivity analysis
  - Findings from surveys versus administrative data

Impact Estimation Issues (cont.)

- More data issues
  - Follow-up periods (<1 year versus 10+ years)
  - Missing in data (unit and item nonresponse)
- The "multiple comparisons" problem
  - Identified confirmatory hypotheses to reduce problem
  - Used statistical corrections to adjust significance levels
- Pooling
  - Most evaluations with similar treatments across sites pooled data
  - Evaluations with diverse treatments did not pool across sites

Beyond Impact Evaluation…

- Implementation evaluation (or "process analyses")
  - document what intervention was implemented
  - document counterfactual conditions
  - analyze fidelity of treatment, understanding of treatment, and cross-site variation
- Participation analysis
  - examines understanding of intervention, how participation varies by characteristics
- Cost-benefit analysis
  - assesses if intervention is worthwhile to participants, government, and society as a whole



Thank you

STATE OF THE SCIENCE MEETING

Discussion: Design of Social Security Administration Demonstration Evaluations

Discussant: Jesse Rothstein, University of California, Berkeley

OVERVIEW

Excellent paper – really useful catalog.

My comments:

- Which aspects of design deserve more attention?
- Interplay between evaluation design and policy usage of results

Exhibit 1. Reviewed SSA Evaluations

Non-experimental

Proof-of-concept Studies

Benefits Entitlement Services Team Demonstration Project (B.E.S.T.)
Homeless with Schizophrenia Presumptive Disability (HSPD)

Impact analyses

Homeless Outreach Projects and Evaluation ((HOPE)
SSI Work Incentives Demonstration Project

Experimental

Classical experiments

Transitional Employment Training Demonstration (TETD)
Project NetWork
Accelerated Benefits (AB) Demonstration
Benefit Offset Pilot Demonstration ((BOPD)
Benefit Offset National Demonstration (BOND)
Mental Health Treatment Study (MHTS)
Youth Transition Demonstration (YTD)
Promoting Readiness of Minors in SSI (PROMISE)
Promoting Opportunity Demonstration (POD)
Demonstration to Maintain Independence and Employment (DMIE)[1]
Timely Wage Reporting Nudge Experiment

Natural experiment

Ticket-to-Work Evaluation

IMPORTANT DESIGN ASPECTS

- **Statistical power**
- **Population representativeness & self selection**
- Missing data
- Site effects

Statistical Power

- Evaluators emphasize power calculations. But the importance is in the decisions.
- Essential to figure out ex ante how big/small effects need to be to matter.
- Ex post, compare results to that. What did we learn?

Population representativeness & self selection

- Selection on gains is a huge issue.
- Build study of this into design. E.g., extra incentives for some participants – are effects smaller for them?

IMPORTANT DESIGN ASPECTS

- Statistical power
- Population representativeness & self selection
- **Missing data**
- **Site effects**

Missing data

- Explore multiple imputation for item missingness.

Site effects

- These demonstrations are in few sites.
- But in other studies, site effects are often very large.
- Gets at efficacy vs. efficiency distinction

Interplay between evaluation design and policy usage

What questions are we trying to answer?

Paper

Emphasizes efficacy vs. effectiveness

Efficacy: "test the optimum implementation of an intervention, at a small scale"

Effectiveness: "consider the program in a 'real world' setting, often increasing the scale of operations"

What do we want to know?

- Does a particular mechanism operate?
- Can we design a program to activate that mechanism?
- What is the best program among those that do?
- Is the program we've implemented successful?

Program gave SSI and SSDI participants vouchers for training and employment services. Did this increase employment?

Mechanism: Are trainable skills / job matching barriers to employment?

Can we design programs that overcome those barriers? What program features are successful at enrolling participants and getting them over the hurdles?

Scalability. Once we've designed the program, does it continue to work at scale?

Important points

- The evaluation that tests the mechanism may use a program that we would never scale.
- We often jump straight to testing effectiveness, without ever verifying mechanisms.
  - We focus on a very specific program, without preliminary evidence that it works.
  - We don't really test at scale, but we act as if we have.
  - If the program fails, we interpret it as falsifying the mechanism.
  - If it succeeds, we rarely get around to testing it again at scale.

Mechanism → Program design → Efficacy → Effectiveness






STATE OF THE SCIENCE MEETING

Thank you!

Jesse Rothstein | rothstein@berkeley.edu | @rothstein_jesse

Discussion: Design of Demonstrations

Discussant: Jack Smalligan, The Urban Institute

STATE OF THE SCIENCE MEETING

Lessons Learned from SSA Demonstrations: A State of the Science Meeting

We are on a break.
Content will resume shortly.