External Validity Bias in Cross-National Generalization of Findings from Experimental Impact Evaluations: Causes, Consequences, and Cures

Stephen H. Bell and Robert B. Olsen, Westat
Elizabeth A. Stuart and Larry L. Orr, The Johns Hopkins University

Conference on Rigorous Impact Evaluation in Europe
Torino, Italy
22nd May, 2018

Sponsors/Funders

- Institute for Education Sciences, U.S. Department of Education
- National Science Foundation, United States Government

Problem and Goal

- Random assignment evaluation gives reliable estimates of policy effects for studied sample—but evidence may not generalize to population affected by policy decision
  - e.g., impact varies by location (Greenberg et al, 2003; Nisar, 2010) and sites are not representative
- "External validity bias" = expected value of impact estimate - average impact in the policy population
  - formalized by Olsen et al. (2012)
- Goal = remove external validity bias by translating evidence from a non-representative set of places (sites) into reliable policy guidance for a broader geography
- Higher goal = don't let external validity bias arise

Topics on External Validity Bias Covered in Talk

- Establishing relevance
- Finding ways to measure its degree
  - examples from education research in the U.S.
- Attaining externally valid experiments
  - Design vs. analysis
- Contemplating Europe and the future

Westat

Prevalence of the Problem

266 of 273 early U.S. social experiments were conducted in purposive sets of sites (Greenberg & Shroder, 2004)

More recent appraisal: "Random site selection is very rare in [U.S.] social experiments" (Olsen et al., 2013)

- Why? "Evaluators and evaluation sponsors often justify selecting sites purposively instead of randomly to reduce the costs of the evaluation" (Olsen et al., 2013)

Random selection, when used, usually does not result in representative inclusion

- Why? Local agencies or polities have to agree to random assignment
- Often don't

Scrutinize and Quantify (Bell & Stuart, 2016)

- Check alignment on moderators of impact
- Simulate size of bias
- Directly measure size of bias

Moderator Alignment: Sample vs. Population

Do the background characteristics of the sample sites look like those of the population?

- Surprisingly rare to check and report
- Can be glaring differences (Stuart et al., 2017)
  - large, urban, low-to-mid-performing school districts = 50% vs. 4%
- Guidelines for "close enough" (Stuart et al., 2011; Tipton, 2014)
- First identify factors that moderate impact (Bloom et al., 2003; Peck et al., 2017; many others)

Outcome Alignment: Sample vs. Population

- For new interventions, control group Y should line up with “untreated” population Y
  - “Placebo test” (Hartman et al., 2015)
- For ongoing programs, treatment group Y should line up with “treated” population Y
  - “Drugged test” (Bell & Stuart, 2016)
- Challenge = getting the outcome data for the population
- Payoff = matched outcomes give much greater assurance than matched moderators

External Validity Bias Simulations

- Standard approach (Kern et al., 2016)
  - Get experimental impact estimates for many sites
  - Impose hypothetical non-representative site selection algorithms
  - Calculate impacts in selected sites vs. all sites
- Stronger if
  - Begin with set of sites that represents some population
  - Selection algorithms follow inclusion patterns from actual studies

Impact Alignment: Sample vs. Population

- Best way to measure external validity bias
- Imposes unusual data requirements
  - Highly rigorous impact estimates for all sites in a population
  - Knowledge of which of those sites would be included in a real-world purposive-site experiment

One Has Been Done (Bell et al., 2016)

- Elementary school reading intervention
- Population impact estimate for entirety of 9 states
- Compared to impact estimates for 11 purposive samples of school districts
- Purposive samples on average "missed" by .10 of a standard deviation (= 1.5 months of student progress)
- First entrant to the "population replication" literature
  - Like the "design replication" literature checking for internal validity bias (LaLonde, 1986; many others)
  - Know the right answer
  - Checking if second-best answer is good enough

Attaining Externally Valid Experiments: Overview

- Good design gets you there
  - if a probability sample of sites can be “harvested”
- Analytic adjustments do not

A Good Analysis Won't Get You There

- Critically depends on having all moderator variables for sample *and population*
  - No one does
- Basic or complex analysis techniques
  - Weighting (Stuart et al., 2010)
  - Regression adjustment (Bell et al., forthcoming)
  - Subclassification (Tipton, 2014)
  - Bayesian additive regression trees (Kern et al., 2016)

There're all the same: some bias reduction but far from bias elimination

Recipe for a Good Design—and "Harvest" (Olsen & Orr, 2016)

- Identify the population of interest -- far too rare / can be multiple
- Develop a sampling frame that's population-wide
- Select sites randomly
- Set sample size to account for random site selection
- Minimize non-inclusion – make random assignment procedures more attractive, reimburse local costs
- Look for a non-experimental method to compare impacts between "harvested" sites and all initially selected sites (Kaizar, 2011)

Lessons for Europe: The Big Picture

[Extrapolating, without knowing the lay of the land in Europe]

- It's very hard to generalize reliably from a non-random set of sites
- The amount of external validity bias could be large ➔ external validity should not be neglected in pursuit of internal validity
- Both must undergird "Rigorous Impact Evaluation in Europe", since internal and external bias are equally capable of creating misleading policy guidance

Practical Recommendations for Europe Going Forward: One Scholar's View

- Define the population of interest at the *start* of every impact evaluation
- If possible, randomly select—and then fully "harvest"—sites from the population ➔ ideal of no internal validity bias and no external validity bias
    - it has been done
- Do *some type* of impact analysis for "non-harvested" sites (i.e., without random assignment)
- When short of the ideal, "scrutinize and quantify" the extent of external validity bias, in every way possible
    - and report results at the *end* of every evaluation