

Assessing generalizability: Case study of Positive Behavior Interventions and Supports

Elizabeth Stuart
Johns Hopkins Bloomberg School of Public Health
Departments of Mental Health and Biostatistics

estuart@jhsph.edu
www.biostat.jhsph.edu/~estuart

Funding thanks to NIMH 1K25MH83646

January 28, 2009

- 1 Introduction: The need for methods to assess generalizability
- 2 Motivating example: PBIS
- 3 Examining generalizability in MD PBIS study
- 4 Next steps and conclusions

- 1 Introduction: The need for methods to assess generalizability
- 2 Motivating example: PBIS
- 3 Examining generalizability in MD PBIS study
- 4 Next steps and conclusions

Goal of this work: Moving towards dissemination

- Increased emphasis on evidence-based practices
- Often get wide implementation after appearing on various lists
 - e.g., Cochrane and Campbell Collaborations, What Works Clearinghouse, NREPP
- Focus in randomized experiments on internal validity
- But how do we know the intervention will be effective beyond the original study sample (external validity)?
- Main idea: Make individuals in the trials look as similar to the population of interest as possible

Previous work on external validity

- Causal generalizability (Shadish, Cook, & Campbell, 2002, Cook 2007)
 - Factors that limit internal and external validity
 - Primarily conceptual, raising issues to be considered
- Decomposing bias in treatment effect estimates (Imai, King, & Stuart, 2008)
 - Bias due to non-random sample selection
 - Bias due to non-random treatment assignment
 - e.g., experiments have smaller bias due to non-random treatment assignment but may have larger sample selection bias
 - Idea here: Use methods developed to deal with non-random treatment assignment to deal with non-random sample selection
- Simple metrics to assess generalizability (Glasgow et al., 2006)
 - RE-AIM framework
 - Summary indices assessing reach, effectiveness

Current approaches to facilitate generalization

In design:

- Random sampling from population: Great, but rarely done
 - Just a few examples: Upward Bound, Job Corps
- Purposive sampling: Typical instances, heterogeneous instances
 - Ever done?
- Practical clinical trials (Tunis et al., 2003; Glasgow et al., 2006)
 - Generally very expensive, large-scale

In analysis:

- Post-stratification
 - Averages subgroup effects using population proportions
 - Doesn't require individual-level data
 - But very restrictive in terms of the number of covariates that can be used
- Research synthesis approaches (e.g., meta-analysis, cross-design synthesis, response surface modeling) to combine results across studies
 - Requires multiple studies on the same topic—rare in fields such as education, policy

A lot of discussion primarily conceptual, especially with respect to single studies

- 1 Introduction: The need for methods to assess generalizability
- 2 Motivating example: PBIS**
- 3 Examining generalizability in MD PBIS study
- 4 Next steps and conclusions

Positive Behavioral Interventions and Supports (PBIS)

- School-wide behavior improvement program (Lewis & Sugai, 1999)
 - Implemented in over 5,000 schools across 40 states (www.pbis.org)
 - President Obama previously had introduced Senate legislation to provide federal funding for PBIS
- PBIS helps schools create systems (discipline, reinforcement) and procedures (office referral, reinforcement) that promote positive student and teacher behaviors
- Very few randomized trials of PBIS, and even less known about broad effectiveness
- Motivating question: What would the effects of PBIS be if implemented across the state of Maryland?

- Randomized trial in 37 MD elementary schools
- Data from state Department of Education on all elementary schools in MD
- Both datasets have:
 - Demographics, test scores, suspensions, teacher characteristics, school funding
- Covariates measured in 2002 (pre-trial)
 - Demographics, school funding, achievement test scores
- Outcomes measured 2004-2006
 - School average achievement test scores, % meeting NCLB proficiency levels

- 1 Introduction: The need for methods to assess generalizability
- 2 Motivating example: PBIS
- 3 Examining generalizability in MD PBIS study**
- 4 Next steps and conclusions

How similar are the trial schools to those across MD?

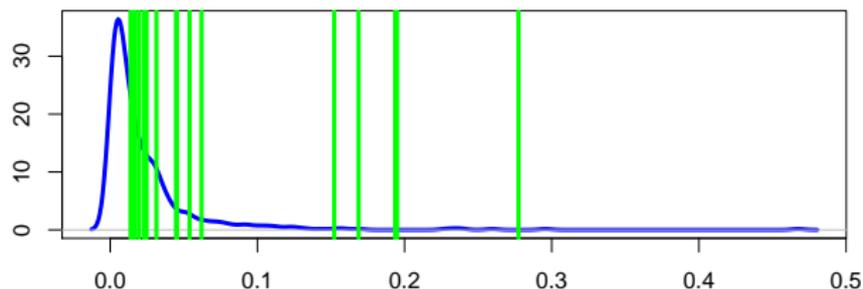
	Trial	Statewide	p-value
Enrollment	485	494	0.73
Attendance (%)	95.3	95.3	0.75
Limited English (%)	2.1	4.2	0.00
Free meals	39.7	35.6	0.25
County wealth per student	\$250,000	\$270,000	0.04
County expend. per student	\$7,500	\$8,000	0.00
Achievement Test Scores (% Advanced or Proficient)			
Grade 3 Math	27.4	32.3	0.07
Grade 3 Reading	32.9	34.9	0.50
Grade 5 Math	44.6	51.8	0.04
Grade 5 Reading	54.2	53.9	0.92

Propensity scores as summary of differences

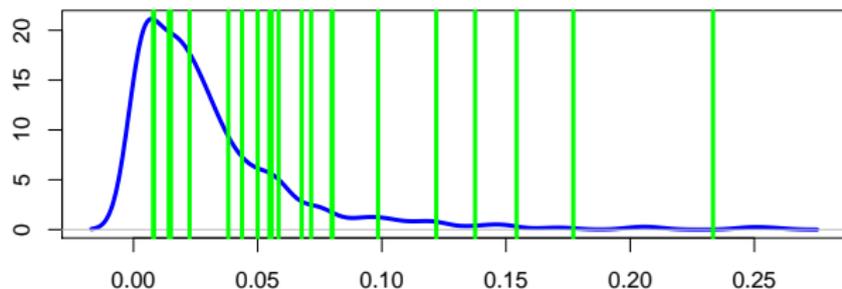
- Quite a few significant differences on individual variables
- Hard to combine these into one measure
- Use ideas behind propensity scores to do so
 - Generally used in non-experimental studies to identify/reduce extrapolation
 - Used to ensure that groups being compared are similar
 - In our case the “treatment” = being in the randomized trial
 - Fit logistic regression predicting membership in trial given characteristics
 - Propensity scores = predicted probabilities from that model (p_i)
 - Done separately for treated and control groups
- Rosenbaum & Rubin (1983), Stuart & Rubin (2007)

Propensity scores across state and in trial

Propensity scores in control schools and across the state



Propensity scores in treated schools and across the state



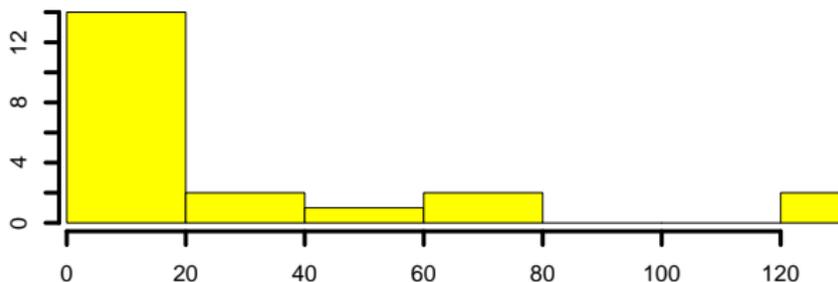
- Overall difference in each group: 0.7 standard deviations
 - Quite large! (Rubin, 2004)
- Not many differences on individual characteristics, but they do combine to create differences in the propensity score

But can we make the trial schools look like the state?

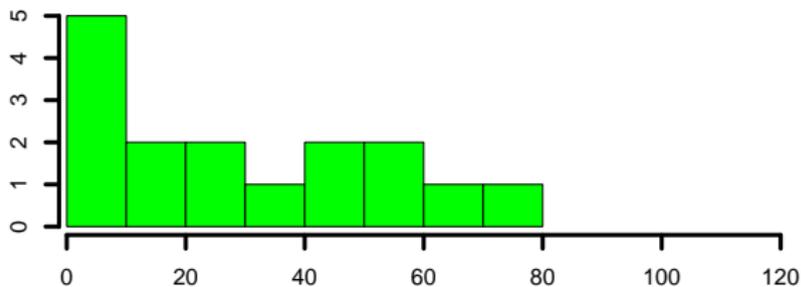
- Main idea: weight schools in trial to look like schools across the state
 - Inverse probability of treatment weighting (Cole & Hernan, 2008)
 - Each school in trial receives weight $\frac{1}{p_i}$
 - Treated like a sampling weight, but estimated rather than known
 - Then take weighted average of outcomes among trial schools
 - Weights both groups (treated and control) up to full population (the state as a whole)
- Diagnostics:
 - How extreme are the weights?
 - How similar are the weighted control group outcomes to those of the state as a whole?

What do the weights look like?

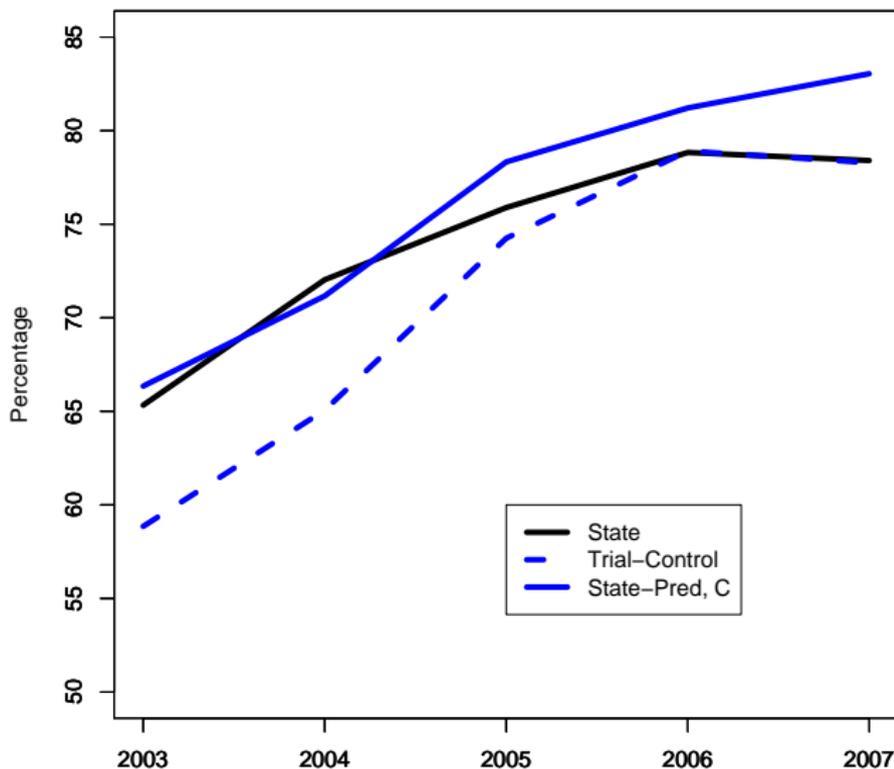
Weights for treated schools

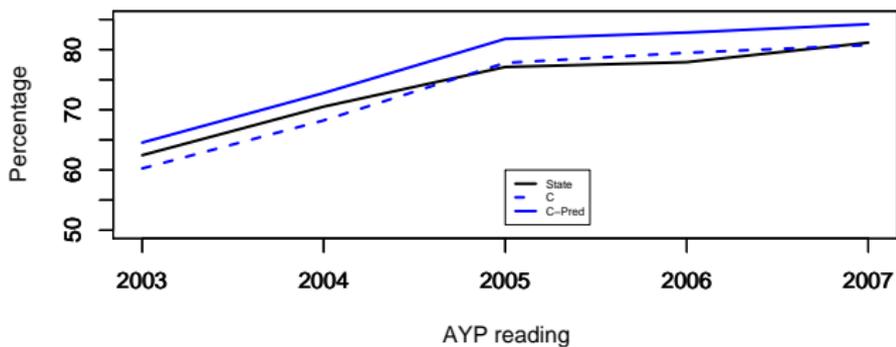
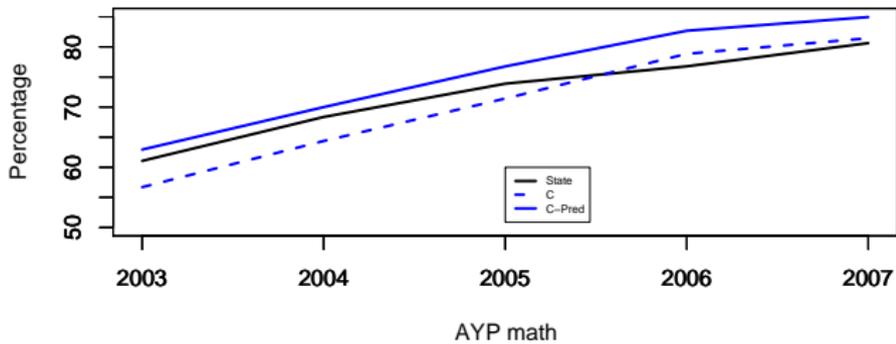


Weights for control schools



Do the weighted control schools look like the state?





- 1 Introduction: The need for methods to assess generalizability
- 2 Motivating example: PBIS
- 3 Examining generalizability in MD PBIS study
- 4 Next steps and conclusions

Next steps: Developing these methods

- Use idea to estimate effect for schools across state
 - Compare post-stratification, propensity score post-stratification, weighting
 - Develop diagnostics, including measure of extrapolation
- Simulations ongoing to investigate performance of methods
- Extensions to utilize multiple trials at once
- Determine guidance for when this generalization is possible
- How should studies be designed to facilitate generalizability?

Next steps: Comparison with other approaches

- In reality, some schools in state implementing PBIS on their own
 - Calculate effect using observational methods with only state data
 - Calculate effect by bridging randomized and observational studies (e.g., research synthesis, confidence profile method)

- In some cases only have one experiment, but also have information on the population
- Goal of this work: Develop methods to use that data to determine whether the experimental results can generalize
 - Give researchers quantitative way to investigate generalizability
 - Take advantage of the good features of both datasets: internal validity of experiment and representativeness of population data
 - Useful for determining whether broad implementation makes sense

References

- Cook, T. D. (2007). Evidence-based practice: Where do we stand? The Gwen Ilding Brogden Distinguished Lecture, The 20th Annual Research Conference A System of Care for Childrens Mental Health: Expanding the Research Base. Available at <http://rtckids.fmhi.usf.edu/rtccconference/20thconference/iding.cfm>.
- Glasgow, R.E., Klesges, L.M., Dzewaltowski, D.A., Estabrooks, P.A., & Vogt, T.M. (2006). Evaluating the impact of health promotion programs: using the RE-AIM framework to form summary measures for decision making involving complex issues. *Health Education Research* 21(5): 688-694.
- Imai, K., King, G., & Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171: 481-502.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.
- Stuart, E. A. & Rubin, D. B. (2007). Matching methods for causal inference: Designing observational studies. In J. Osborne, ed., *Best practices in quantitative social science*. Thousand Oaks, CA: Sage Publications.
- Tunis, S.R., Stryer, D.B., & Clancy, C.M. (2003). Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. *Journal of the American Medical Association* 290(12): 1624-1632.