Abstract Title Page
Not included in page count.

Title: Estimates of external validity bias when impact evaluations select sites purposively

Author(s):

Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health
Robert B. Olsen, Abt Associates
Stephen H. Bell, Abt Associates
Larry L. Orr, Johns Hopkins Institute for Policy Studies

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

Many impact evaluations in education research are carried out in a set of sites, for example schools or school districts. Conclusions from those evaluations are often then used to guide policy decisions for a larger set of schools or districts. However, the sites in the evaluation are generally not selected randomly from that larger population of interest and may not be representative of that population. Instead, evaluation sites are usually selected in a purposive manner. It is unknown how such purposive site selection impacts the external validity of impact estimates in these evaluations. While much attention has been paid to methods to increase internal validity—the ability of an evaluation to generate unbiased impact estimate for the sites in the study—less attention has been paid to how purposive site selection may impact external validity. This is of particular concern in contexts where impacts are heterogeneous, particularly across sites. If there is no treatment effect heterogeneity there will be no problem in generalizing results from one sample to another and thus no external validity bias. However, as described below, if there is treatment effect heterogeneity then the bias from selecting a purposive sample of sites (rather than a random sample) can be large.

In previous work, Olsen et al. (2011) developed a conceptual model of the external validity bias that can result from purposive site selection when the goal of the evaluation is to produce impact estimates for a broader population of sites. That model of purposive site selection is based on the premise that purposive samples can be treated as probability samples where the probabilities of inclusion in the study are unknown. The bias formula derived shows that the external validity bias from purposive site selection can be expressed as a function of the variation in impacts across sites in the population, the variation in the unobserved site selection and self-selection (i.e., inclusion) probabilities across these sites, and the correlation between site-level impacts and the site inclusion probabilities. This model thus helps researchers consider the factors that contribute to external validity bias.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

While there has been some increasing interest in external validity, most work to this point has been in assessing the similarity of a randomized trial sample and a population of interest (e.g., Stuart et al., 2010; Tipton, 2011). The goal of the current research is to calculate empirical estimates of the external validity bias in educational intervention evaluations that can result from selecting sites purposively. This will help researchers and funding agencies understand the implications of purposive site selection.

Setting:

Description of the research location.

(May not be applicable for Methods submissions)

We use longitudinal data from 10 states across the country. These 10 states are a subset of the 15 states used in a recent evaluation of Reading First (Gamse et al., 2011). The 10 states were selected because they offer student achievement data for at least two years before the implementation of Reading First. These pre-intervention data allow us to account for selection

into treatment (receipt of Reading First funds) and estimate comparative interrupted time series models to estimate the effects of the Reading First program on 3rd grade test scores.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)
All schools eligible for Reading First in the 10 states included in this study.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
(May not be applicable for Methods submissions)
Reading First is a federal program that provided funds to states (and then from states to districts and schools) to improve reading instruction (Gamse et al., 2011). Schools that received Reading First funds had to use evidence-based reading curricula, offer professional development to teachers to help them implement evidence-based instruction in reading, and screen and monitor students to identify early reading difficulties. Six-year grants were awarded to state education agencies following a formal application process. States then distributed funds to school districts using a competitive process, with priority given to districts and schools with the lowest reading proficiency and highest poverty statuses. Nationally there are a total of nearly 6,000 Reading First schools in over 1,800 school districts (Gamse et al., 2011). To estimate the effects of Reading First we will use the year of receipt of Reading First funds to define an "interruption" in a comparative interrupted time series design, as detailed below.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.
The study provides the first empirical estimates of the external validity bias that can result from selecting sites purposively. The analysis focuses on a single intervention, Reading First, using data that cover all school districts in 10 selected states, which, for purposes of this methodological exercise, we treat as the population of interest. These data, then, allow us to estimate the effect of Reading First in the target population. We then use the samples of districts selected for 10 rigorous impact evaluations in education to estimate the effect of Reading First that would be obtained if we used the purposively selected districts. Our estimate of the external validity bias for a particular purposive sample is the difference between the impact estimate calculated using the purposively selected set of schools and the true population effect, calculated using all schools in the selected states. While all of these studies were conducted for the Institute of Education Sciences, they span different interventions, grade levels, and research organizations. This allows us to produce empirical evidence on the consequences of purposive site selection that span 10 purposive samples in educational evaluations.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.
This study obtains empirical estimates of the bias from purposive site selection. This empirical exercise requires data from an evaluation with a representative sample of sites or the universe of sites in the defined population of interest, to provide a benchmark for comparison with purposively selected samples of sites. For this exercise, we use data from Abt Associates' most recent evaluation of Reading First, which included all districts that received Reading First

funding in 15 states (Gamse et al., 2011). For our study, we focus on the subset of 10 states for which there are at least two years of pre-Reading First data.

To obtain empirical evidence on the consequences of purposive site selection, we obtained lists of the school districts that participated in 10 educational impact evaluations that selected districts and schools purposively. These evaluations included between 3 and 47 school districts; they included evaluations of classroom interventions, after school programs, and federal education programs sponsored by the U.S. Department of Education. All of these evaluations were conducted for the Institute of Education Sciences and were based on random assignment or quasi-experimental designs. To estimate the bias from purposive site selection, we matched the districts from these 10 studies to the Reading First study data in the 10 states in our analysis. This allows us to *estimate the impacts of Reading First using sets of districts that were selected purposively for evaluations of other programs*. We obtain 11 estimates of the external validity bias—one from each of the 10 purposive samples and one that pools the 10 purposive samples—by taking the difference between each of these impact estimates and the impact estimate calculated using all districts within the 10 analysis states that received Reading First funding.

The primary model used to estimate the effects of Reading First is a comparative interrupted time series model. This model is estimated first in all districts in the Reading First evaluation population (all schools eligible for Reading First in our 10-state data set), and then in the subsets of districts included in the 10 Institute of Education Sciences impact evaluations. For each purposive sample, the external validity bias is calculated as the difference between the true impact in the population of Reading First-eligible schools in the 10 states and the impact estimated using the purposively selected subset of sites.

We use a number of methods to provide benchmarks for interpreting the size of the external validity bias that results. First we express the impact estimates and the bias in effect size units. We also calculate a naïve estimate of the treatment effect estimate within the population, which does not adjust for covariates or time trends, and that provides an estimate of the internal validity bias that would result from a naïve estimation of the impact of Reading First. We are then able to compare the size of the external validity bias to the internal validity bias that might result using the same sample, to relate our findings on external validity bias to the literature on internal validity bias (the so-called "design replication" studies; see, for example, Glazerman, Levy, & Myers, 2003 and Cook, Shadish, & Wong, 2008.)

Usefulness / Applicability of Method:

Demonstration of the usefulness of the proposed methods using hypothetical or real data.

Our approach to estimating external validity bias is illustrated using real data from 1) an impact evaluation of Reading First, and 2) lists of purposively-selected districts used in 10 IES-funded impact evaluations. This is particularly valuable in showing "how far off" estimates of educational effectiveness based on purposively-selected sites can be from the desired population figure.

Research Design:

Description of the research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

(May not be applicable for Methods submissions)

As described above we use a quasi-experimental comparative interrupted time series design to estimate the impacts of Reading First. The impacts calculated from each of 10 sets of purposively selected sites are then compared to the impacts in the full population of interest (all schools eligible for Reading First). This quasi-experimental design may produce estimates of impact that contain some internal validity bias. However, as long as the magnitude of that bias does not correlate with which districts are included in evaluations that select sites purposively, the internal bias will not distort the comparisons of interest between the population estimate and the estimates based on purposive samples.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

We use data from Abt Associates' evaluation of Reading First, collected at the student level and aggregated to the school level. These data include third grade test scores as well as demographics. We also use data that list the sites selected for 10 IES-funded impact evaluations. Analysis methods are described under Research Design above.

Findings / Results:

Description of the main findings with specific details.
(May not be applicable for Methods submissions)

We find that the effect size estimates obtained using each of the 10 purposive samples, as well as from the pooled set of all 10 samples, are all lower than the true population impact of Reading First across all 10 states (0.027; see Figure 1). The effect size estimates obtained from each of the 10 purposive samples range from -0.09 to 0.019; the estimated effect size calculated using the set of sites in any of the 10 purposive samples is -0.015. This implies an average external validity bias due to purposive site selection of -0.042. The impact estimate for the population and the estimated external validity bias are both significant at the 0.05 level. To put this into context, this estimate of the external validity bias is smaller than the internal validity bias produced by an extremely naïve impact estimation model, but is 2.5 times larger than the internal validity bias produced by a better but still suspect impact estimation model.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

While Olsen et al. (2011) shows that purposive site selection can potentially lead to biased impact estimates for the population of interest, little is known about the magnitude of the external validity bias in real educational studies. This paper provides the first empirical estimates of the bias that can result from selecting sites purposively. We find that the estimated effect of Reading First is systematically different for schools included in purposive-site evaluations than for the population of interest. We find consistent negative bias in the 10 purposive-site evaluations examined, but this will likely not generalize to other contexts. In any particular study, the magnitude of the bias will depend on the variation in impacts across sites in the population of interest (and other factors described in Olsen et al. 2011). The results from this paper should be treated as illustrative of the magnitude of the bias that can result when impacts do vary across sites, and sites are selected purposively.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Cook, T.D., Shadish, W.R., & Wong, V.C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management* 27(4): 724-750.

Gamse, B., Boulay, B., Rulf Fountain, A., Unlu, F., Maree, K., McCall, T., McCormick, R. (2011). *Reading First Implementation Study 2008-09 Final Report.* Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.

Glazerman, S., Levy, D. M., and Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science* 589, 63-93.

Olsen, R., Bell, S., Orr, L., and Stuart, E.A. (2011). The loss of external validity in policy evaluations that choose sites purposively: Bias in the estimates and recommendations for bias reduction. Under review, *Journal for Policy Analysis and Management.*

Stuart, E.A., Cole, S.R., Bradshaw, C.P., and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *The Journal of the Royal Statistical Society, Series A* 174(2): 369-386.

Tipton, E. (2011). Using Propensity Score Matching Methods to Improve Generalization from Randomized Experiments. Presentation at Society for Research on Educational Effectiveness Spring Meeting. Washington, DC: March 2011. Available at http://www.sree.org/conferences/2011/program/downloads/abstracts/43.pdf.

Appendix B. Tables and Figures
Not included in page count.

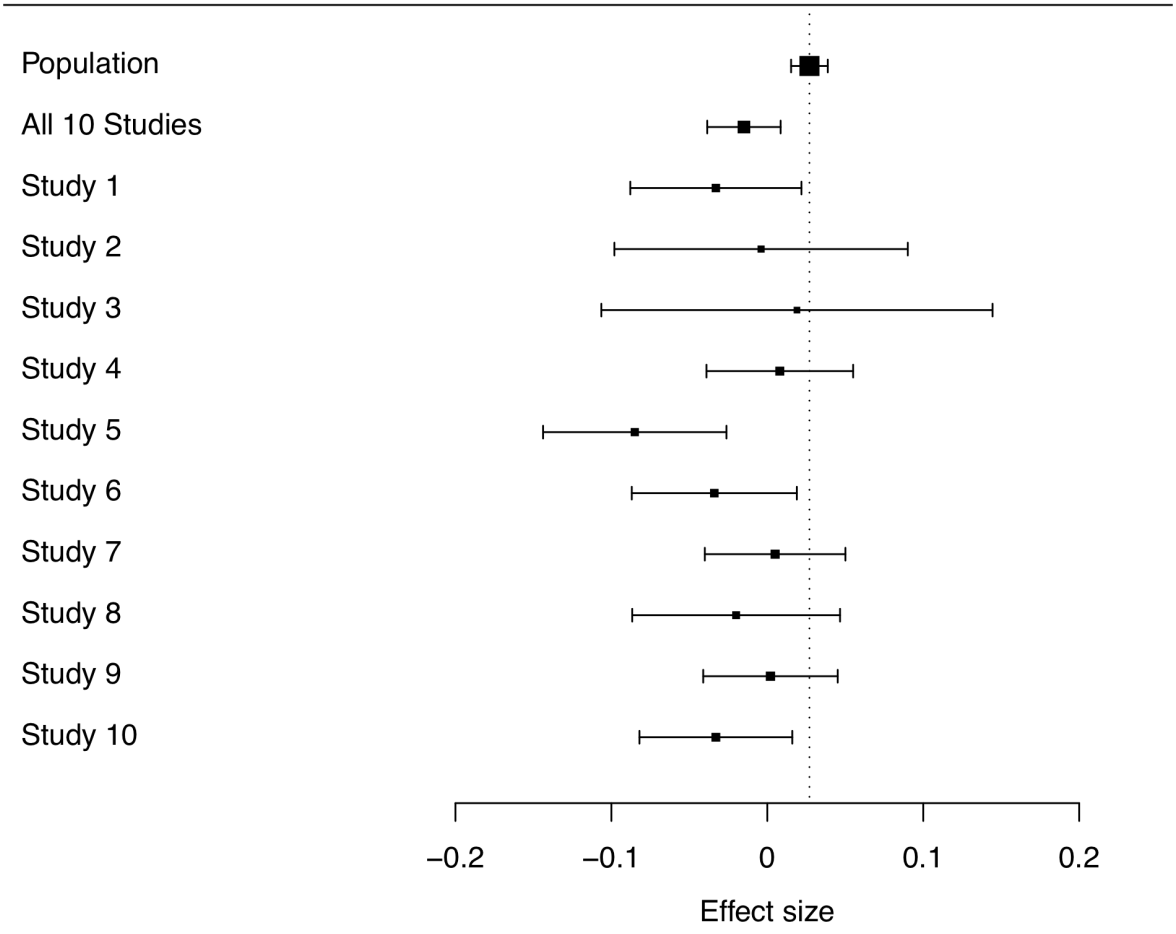

Figure 1: Estimates of the effect of Reading First (in effect size units, relative to the distribution of student test scores; with 95% confidence intervals) in the 10 state population of interest ("Population") as well as in 10 purposively selected samples of sites (Study 1-10) and the pooled sample of purposively selected sites ("All 10 Studies").