Comparison of External and Internal Validity Bias in Experimental and Quasi-Experimental Approaches

Mark White, Ben Hansen, Tim Lycurgus, Brian Rowan

Context:

The ideal experiment randomly samples schools from a defined population and then randomly assigns treatment within this sample. This is rarely achievable given the difficulty of recruiting schools into an experiment (Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017), leading to growing concerns about the external validity of experimental effect estimates when schools are non-randomly sampled. On the other hand, some quasi-experimental designs can be run on all schools implementing an intervention, improving external validity at the cost of internal validity. This potential trade-off may undermine claims that RCTs are the "gold standard" when estimating a causal effect in a specific target population. One study has found external validity bias can be larger than internal validity bias (Bell, Olsen, Orr, & Stuart, 2016), but no studies have directly measured and compared the external validity bias due to non-random sampling to the internal validity bias due to non-random treatment assignment.

Purpose:

Our goal is to compare the estimated external validity bias to the estimated internal validity bias to identify the best method of estimating a treatment effect within the population of current program users, which we argue is the set of schools in which a causal estimate is most useful.

Program:

The Burst©:Reading program is a personalized, beginning reading program that uses a proprietary algorithm to assign students to small group, supplementary instruction on the basis of students' DIBELS test scores and formative assessment results. Groups receive specific targeted lesson plans based on skill deficits. Burst is targeted towards struggling students (Tier 2) as a supplemental instructional program.

Data:

We combined data from three data sets to form a population frame: 1) school data from the Common Core of Data (CCD) for 2009 through 2016, 2) district test score data and community level characteristics from the Stanford Data Education Archives (Reardon, et al., 2017), and 3) the percentage of third graders proficient in math and English from SchoolDigger.com. Proficiency rates were standardized within state/year. The population frame was reduced to include only schools that served students in Kindergarten, 1st grade, 2nd grade, or 3rd grade were in one of the 50 states (or DC); were classified by the CCD as a "regular school" and not closed before 2016; and contained more than 0 students or had missing information for number of students. This resulted in 54,683 schools in the population frame. Due to data limitations, all data is at the school level. The experimental sample comes from an IES goal 3 study and the Burst user population comes from the Burst provider.

Research Design:

We conducted analyses on three separate samples (see figure 1). The first is the experimental sample, including all schools that participated in the experiment. The effect estimate of this sample contains bias due to non-random selection of schools into the experiment. The second sample is the experimental treatment schools with matched controls. The effect estimate of this sample contains both internal and external validity bias. The difference in the effect estimates from samples 1 and 2 isolate the internal validity bias caused by non-randomly assigning treatment in quasi-experiments. The third sample is the set of all schools using Burst with matched controls. The effect estimate from this sample has no external validity bias because schools were not sampled, but has internal validity bias. The difference in the effect estimates from samples 2 and 3 isolate the external validity bias caused by selection into the experiment.

Data Analysis:

Analyses proceeded in three steps. First, propensity score matching was done to select matched control schools for the RCT treatment schools and the Burst population. Propensity scores were estimated with a Bayesian Additive Regression Tree (Kapelner & Bleich, 2013) using demographic and test score variables for the county, district, and schools. Matching was done using full matching with a maximum of 2 control or treatment schools in each match (Hansen & Klopfer, 2006). Second, rebar estimates (see Sales, Hansen, Rowan, 2018) of school achievement were estimated using the schools in the population frame not matched to any schools. School achievement was adjusted for these rebar estimates to remove the dependence on demographic characteristics and time trends.

Third, all samples were analyzed with the interrupted time series model:

$$e_{ts} = \beta_0 + I(T > baseline)_{ts} * \beta_1 + (Burst)_s * \beta_2 + I(T > baseline)_{ts} * (Burst)_s * \beta_3 + \gamma_m + \mu_s + \epsilon_{ts}$$

where t is time, s is school, β_0 is a baseline mean score, β_1 is deviation from the baseline common to all schools after treatment begins, β_2 is the mean difference between Burst and non-Burst schools, β_3 is the treatment effect of interest, estimated as a difference-in-difference, γ_m are match fixed effects, $\mu_s \sim N(0, \tau)$ is a random effect, and $\epsilon_{ts} \sim N(0, \tau)$ is a residual. The internal validity bias is estimated as $\widehat{\beta_3^{2-1}} = \widehat{\beta_3^2} - \widehat{\beta_3^1}$, where the superscript indicates the population from which the estimate is taken. The external validity bias is estimated as $\widehat{\beta_3^{2-3}} = \widehat{\beta_3^2} - \widehat{\beta_3^3}$.

Results:

Table 1 shows the results of the analyses. As can be seen, the effect of Burst on school proficiency rates is not significant in any of the samples. Despite this, we can still estimate the internal and external validity bias. The internal validity bias is estimated to be $\widehat{\beta_3^{2-1}} = 0.15 -$

$0.11 = 0.04$. The external validity bias is estimated to be $\widehat{\beta_3^{2-3}} = 0.15 - 0.04 = 0.11$. Thus, the estimated external validity bias is more than twice the estimated internal validity bias.

Conclusions:

We found that quasi-experimental designs that focus on estimating the treatment effect in the full target population provide a more accurate estimate of the causal effect of a program than a non-randomly sampled randomized control trial. This has important policy implications, especially due to the significantly cheaper cost of running quasi-experimental studies, and the growing push to make research more useful by estimating more targeted causal estimates. Further work is necessary to examine the generalizability of this conclusion, as well as the impact that using school level state test scores, rather than student level scores on a more targeted test. However, our estimates of external validity bias are the same size, though opposite direction, as previous work (Bell, Olsen, Orr, & Stuart, 2016).

References

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318–335.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006X137047

Kapelner, A., & Bleich, J. (2013). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *ArXiv:1312.2171 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1312.2171

Sales, A. C., Hansen, B. B., & Rowan, B. (2018). Rebar: Reinforcing a Matching Estimator With Predictions From High-Dimensional Covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 3–31. https://doi.org/10.3102/1076998617731518

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of School Districts That Participate in Rigorous National Educational Evaluations. *Journal of Research on Educational Effectiveness*, *10*(1), 168–206. https://doi.org/10.1080/19345747.2016.1205160

Figure 1: Description of Populations and Sources of Bias

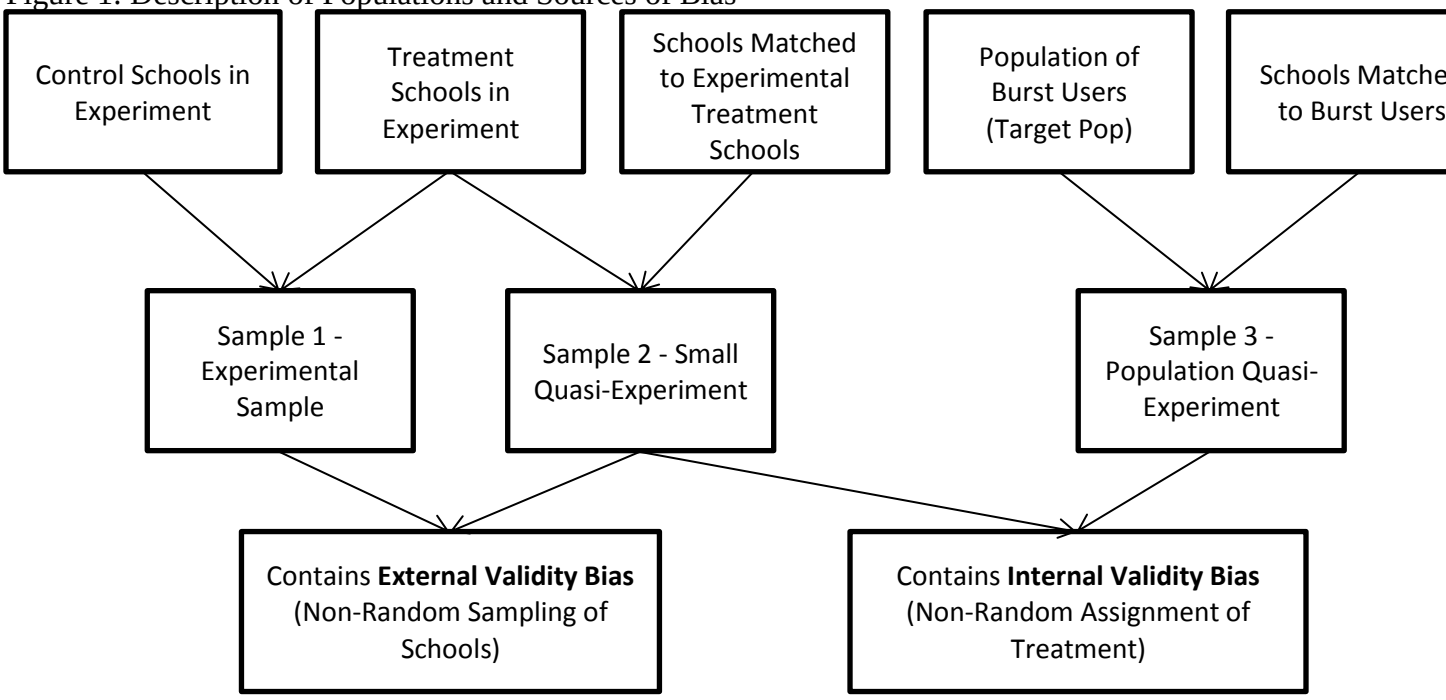

Table 1: Effect Estimates Across Populations

	Estimate	Std. Error	t-value	CI-2.5%	CI-97.5%
Sample 1 – RCT					
Intercept	-0.16	0.23	-0.69	-0.61	0.29
Treatment Time Period (β_1)	-0.01	0.07	-0.20	-0.15	0.12
Using Burst (β_2)	0.02	0.09	0.21	-0.15	0.18
Using Burst Treatment Time Period (β_3)	0.11	0.09	1.15	-0.08	0.29
Sample 2 – Small Quasi-Experiment					
Intercept	0.15	0.05	-3.46	-0.24	0.07
Treatment Time Period (β_1)	0.02	0.01	2.18	-0.03	0.00
Using Burst (β_2)	0.01	0.09	0.11	-0.16	0.18
Using Burst Treatment Time Period (β_3)	0.15	0.07	2.25	0.02	0.28
Sample 3 – Population Quasi-Experiment					
Intercept	-0.10	0.06	-1.74	-0.22	0.01
Treatment Time Period (β_1)	-0.04	0.00	-14.2	-0.04	-0.03
Using Burst (β_2)	-0.02	0.03	-0.86	-0.08	0.03
Using Burst Treatment Time Period (β_3)	0.04	0.02	1.93	-0.00	0.07