SREE Proposal Draft

Evaluating Methods for Handling Multilevel Selection for the Purpose of Generalizing Cluster Randomized Trials

Authors: (Eva) Yujia Li and Christopher Rhoads

Background

In the past decade, the educational research community has devoted increasing attention into developing methods for improving the generalizability of randomized controlled trials (RCTs) (Hedges, 2013; Olsen, Orr, Bell & Stuart, 2013). Existing work focuses on the ways in which institutions, e.g. schools, that volunteer for experiments differ from those that do not. These methods focus on either prospectively recruit representative school samples (e.g. Tipton, 2013a, 2013b, 2014) or retrospective adjust away bias caused by nonrandom school recruitment (e.g., Cole & Stuart, 2010; Kern et al., 2016; O'Muircheartaigh & Hedges, 2014). An implicit assumption is that treatment effects vary only as a function of observable variables that characterize schools (school-level moderators). However, for almost all educational RCTs, the study sample is collected in two stages - schools are recruited first, and then students or teachers volunteer for the study. The importance of accounting for non-random within school selection is evidenced by variations in participation rates across institutions (Blom-Hoffman et al., 2009). Large scale international assessments show that student-level non-response is related to student characteristics, and in general, less capable students are more likely to be absent from assessments (Rust, 2013). It is plausible that such differential participation related to student characteristics also occurs in RCTs. Therefore, unless the consenting teachers and students are representative of all teachers and students in the school, existing methods that adjust estimates based only on hypothesized school-level moderators may fail to remove all of the bias.

One existing method to estimate a population average treatment effect (PATE) is to weigh schools by the inverse of their probabilities of participation in an RCT (e.g. Stuart, Bradshaw & Leaf, 2015). This study extends this method by weighing both schools and students in order to account for non-random selection at both the school and the within school level. Due to the fact that students are nested within schools and their selection processes often vary across schools, there are two viable options for estimating student participation probabilities - both have been adopted in the multilevel propensity score literature (e.g. Rosenbaum, 1986; Kim & Seltzer, 2007) and can be easily adapted to the generalization context by changing the outcome variable from treatment assignment to participation status. The first option is to run separate models for students within each participating school. The second option is to run one multilevel model pooling all student information in all participating schools, with random intercepts and slopes for each school.

Purpose/Objective/Research Questions

This study evaluates through a simulation the effectiveness of PATE estimators that apply student and/or school inverse probability of participation (IPP) weights for reducing bias from nonrandom selection in a cluster randomized trial. The research questions are: (a) under what conditions do methods of accounting for the within school selection process in educational studies reduce bias in estimates of the PATE, compared to only considering the between school selection process and (b) how much is bias reduced under different simulation scenarios?

Data Analysis Procedures

First, a population of schools and students was generated. Two school variables, $V_{1,h} \sim N(0,1), V_{2,h} \sim Bern(0.5)$ and one student variable $X_{h,k} \sim N(V_{1,h}, 1)$ were generated. For each student, two

potential outcomes were generated (EQ 1). Student treatment effect is the difference between two potential outcomes and is a linear combination of school characteristics, student characteristics and their interactions. School selection probabilities were generated as a linear function predicted by $V_{I,h}$ (EQ 2). Student selection probabilities were generated as a linear combination of school and student characteristics (EQ 3). Second, schools and students were selected using the generated selection probabilities. A sample of participating schools was selected and within these schools, students were selected. Participating schools were randomly assigned to treatment or control conditions to mimic a cluster randomized trial.

Third, four estimators were computed using the selected sample. The unadjusted ATE is estimated by an unweighted multilevel model with only one predictor of treatment condition indicator (EQ 4). The IPP-School estimator computed school IPP weights and then applied them to level-2 of EQ 4. The IPP-School+Student separate (IPPSSS) estimator computed student participation weights by separate models within each participation school (EQ 5), and then applied school IPP weights to level-2 and student IPP weights to level-1 of EQ 4. IPP-School+Student multi (IPPSSM) estimator also applied both school and student IPP weights, and the student IPP weights were estimated by a multilevel model pooling information about participants and non-participants in all participating schools (EQ 6).

The study varied three simulation conditions. School population sizes varied between small (H = 50) and large (H = 2000). Within school participation rates varied between low (25%) and medium (50%). Sample selection processes varied among different levels of random selection. Each condition was replicated 200 times. Evaluation criteria for estimator performance were standardized bias and root standardized mean square error (RSMSE).

Results

Estimator performance (Table 1 & Figure 1) showed that when schools and students were randomly selected, all estimators performed similarly well. When schools were not randomly selected but students within schools were randomly selected, IPP-School had smaller standardized bias than the unadjusted ATE, but only smaller RSMSE when school population size was large (2000). When both schools and students were non-randomly selected, the IPPSSS and IPPSSM had smaller standardized bias and RSMSE than the IPP-School and unadjusted ATE. 25% within school participation rates conditions had larger standardized bias and RSMSE than 50%. For all weighted estimators, standardized bias reduced more than RSMSE, showing a trade-off between bias and variance.

Discussion

This simulation study showed that when the within school sample was not randomly selected and the unconfounded sample selection assumption held, ignoring the within school selection process led to bias in the estimated population average treatment effect. Two estimators that involved student IPP weights (IPPSSS and IPPSSM), applied in addition to the school IPP weights, significantly reduced bias in the estimated population average treatment effect compared to applying the school IPP weights alone. Small sample size created challenges for estimating PATE through retrospective adjustment, because the variance inflation of the estimate may override the reduction in bias.

References

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American journal of epidemiology*, *172*(1), 107-115.

Hedges, L. V. (2013). Recommendations for Practice: Justifying claims of generalizability. *Educational Psychology Review. Vol 15,* pp 331 – 337.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, *9*(1), 103-127.

Kim, J., & Seltzer, M. (2007). Causal Inference in Multilevel Settings in Which Selection Processes Vary across Schools. CSE Technical Report 708. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. University of California, Los Angeles.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*(1), 107-121.

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*(3), 207-224.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis, London, Chapman & Hall/CRC, 117.

Tipton, E. (2013a) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties and contexts. *Journal of Educational and Behavioral Statistics. Vol 38* (3), pp 239 – 266.

Tipton, E. (2013b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation review*, *37*(2), 109-139.

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics. Vol 39* (6), pp 481 – 501.

Appendix A

Data Generation and Estimation Models

EQ 1. Models for generating student potential outcomes $y_{hk}(1)$, $y_{hk}(0)$ and treatment effect in the population. Schools are indexed by *h* and students are indexed by *k*.

$$y_{hk}(1) = w_0 + w_1 X_{hk} + \phi_0 + \phi_1 X_{hk}$$
$$w_0 = \pi_{00} + \pi_{01} V_{1,h} + \pi_{02} V_{2,h}$$
$$w_1 = \pi_{10} + \pi_{11} V_{1,h} + \pi_{12} V_{2,h}$$
$$\phi_0 = \pi_{30} + \pi_{31} V_{1,h} + \pi_{32} V_{2,h}$$
$$\phi_1 = \pi_{40} + \pi_{41} V_{1,h} + \pi_{42} V_{2,h}$$

$$y_{hk}(0) = w_0 + w_1 X_{hk}$$
$$w_0 = \pi_{00} + \pi_{01} V_{1,h} + \pi_{02} V_{2,h}$$
$$w_1 = \pi_{10} + \pi_{11} V_{1,h} + \pi_{12} V_{2,h}$$

$$Treatment\ effect_{hk} = Y_{hk}(1) - y_{hk}(0) = \phi_0 + \phi_1 X_{hk}$$
$$= \pi_{30} + \pi_{31} V_{1,h} + \pi_{32} V_{2,h} + \pi_{40} X_{hk} + \pi_{41} V_{1,h} X_{hk} + \pi_{42} V_{2,h} X_{hk}$$

EQ 2. Model for generating school selection probability p_h in the population.
School selection probability is fixed at 12% for all conditions by fixing magnitudes of α_0.

$$ln\left(\frac{p_h}{1-p_h}\right) = \alpha_0 + \alpha_1 V_{1,h}$$

EQ 3. Model for generating student selection probability p_{hk} in the population.

$$ln\left(\frac{p_{hk}}{1-p_{hk}}\right) = \eta_{0h} + \eta_{1h} X_{hk}$$
$$\eta_{0h} = \tau_{00} + \tau_{01} V_{1,h} + \tau_{02} V_{2,h}$$
$$\eta_{1h} = \tau_{10} + \tau_{11} V_{1,h} + \tau_{21} V_{2,h}$$

EQ 4. Model for estimating unadjusted ATE using selected sample.
Z_h is the indicator for school treatment assignments. The unadjusted ATE = $\widehat{\gamma_{01}}$.

$$y_{hk} = \beta_{0h} + \varepsilon_{hk}\,, \varepsilon_{hk} \sim N(0, \sigma^2)\ \textit{(2.4)}$$
$$\beta_{0h} = \gamma_{00} + \gamma_{01} Z_h + u_{0h},\ u_{0h} \sim N(0, \tau)$$

EQ 5. Model for estimating student IPP weight for IPPSSS using selected sample. Student IPP $\widehat{w_{hk}} = \frac{1}{\widehat{p_{hk}}}$.

$$ln\left(\frac{p_{hk}}{1-p_{hk}} | S_h = 1\right) = \eta_{0h} + \eta_{1h} X_{hk}$$

EQ 6. Model for estimating student IPP weight for IPPSSM using selected sample.
Student IPP $\widehat{w_{hk}} = \frac{1}{\widehat{p_{hk}}}$.

$$ln\left(\frac{p_{hk}}{1-p_{hk}} | S_h = 1\right) = \eta_{0h} + \eta_{1h} X_{hk}$$
$$\eta_{0h} = \tau_{00} + \tau_{01} V_{1,h} + \tau_{02} V_{2,h} + u_{0j}$$
$$\eta_{1h} = \tau_{10} + \tau_{11} V_{1,h} + \tau_{21} V_{2,h} + u_{1j}$$

Appendix B

Table 1.

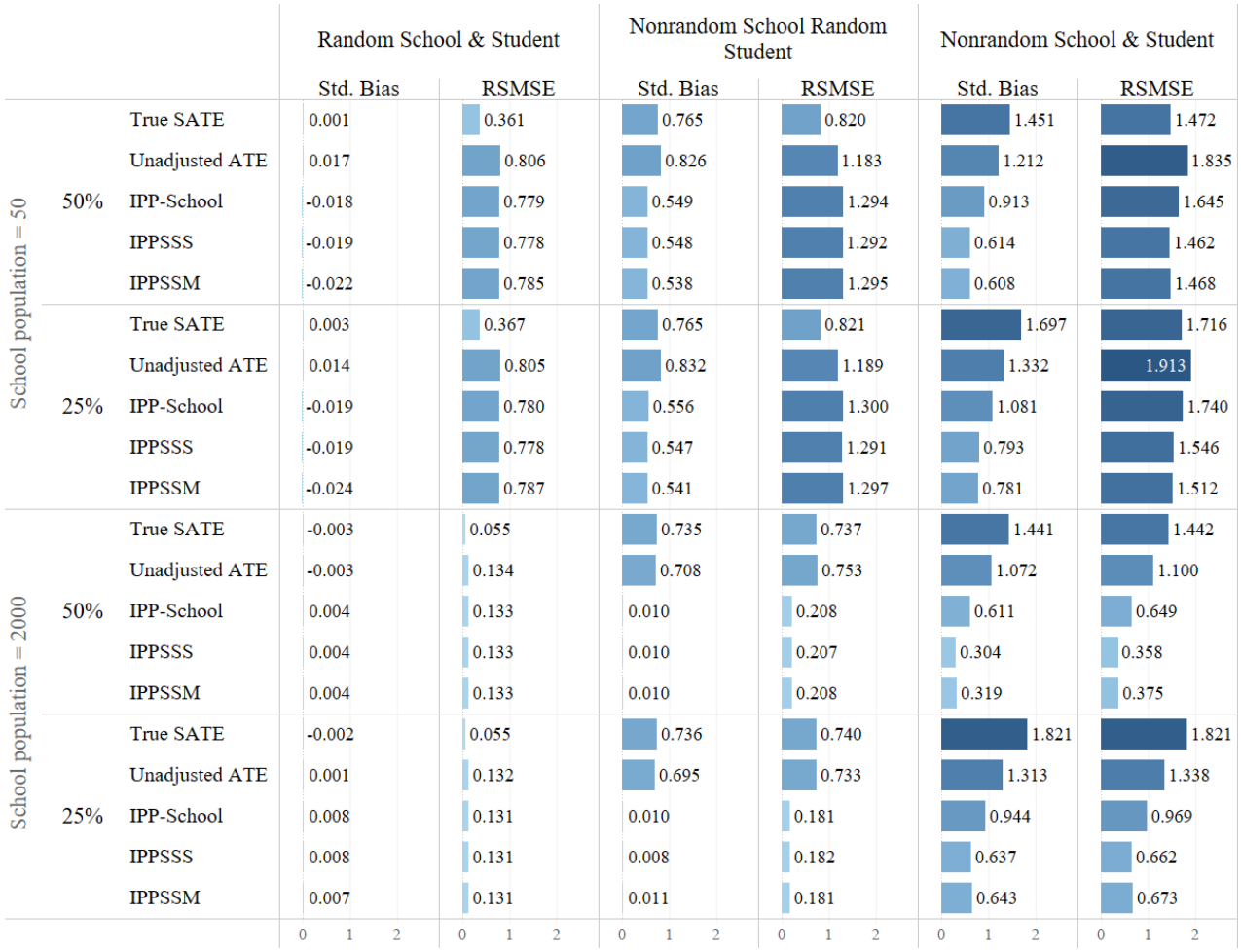
Standardized bias and RSMSE for true SATE and PATE estimators.

School Population	Within school participation rate	Estimators	Sample Selection Process: Random school, random student		Nonrandom school, random student		Nonrandom school, nonrandom student	
			Std. Bias	RSMSE	Std. Bias	RSMSE	Std. Bias	RSMSE
H = 50	50%	**True SATE**	**0.001**	**0.361**	**0.765**	**0.820**	**1.451**	**1.472**
		Unadjusted ATE	0.017	0.806	0.826	1.183	1.212	1.835
		IPP-School	-0.018	0.779	0.549	1.294	0.913	1.645
		IPPSSS	-0.019	0.778	0.548	1.292	0.614	1.462
		IPPSSM	-0.022	0.785	0.538	1.295	0.608	1.468
	25%	**True SATE**	**0.003**	**0.367**	**0.765**	**0.821**	**1.697**	**1.716**
		Unadjusted ATE	0.014	0.805	0.832	1.189	1.332	1.913
		IPP-School	-0.020	0.780	0.556	1.300	1.081	1.740
		IPPSSS	-0.019	0.778	0.547	1.291	0.793	1.546
		IPPSSM	-0.024	0.787	0.541	1.297	0.781	1.512
H = 2000	50%	**True SATE**	**-0.003**	**0.055**	**0.735**	**0.737**	**1.441**	**1.442**
		Unadjusted ATE	-0.003	0.134	0.708	0.753	1.072	1.100
		IPP-School	0.004	0.133	0.010	0.208	0.611	0.649
		IPPSSS	0.004	0.133	0.010	0.207	0.304	0.358
		IPPSSM	0.004	0.133	0.010	0.208	0.319	0.375
	25%	**True SATE**	**-0.002**	**0.055**	**0.736**	**0.740**	**1.821**	**1.821**
		Unadjusted ATE	0.001	0.132	0.695	0.733	1.313	1.338
		IPP-School	0.008	0.131	0.010	0.181	0.944	0.969
		IPPSSS	0.008	0.131	0.008	0.182	0.637	0.662
		IPPSSM	0.007	0.131	0.011	0.181	0.643	0.673

Note. The table shows standardized bias and RSMSE of the true SATEs and four PATE estimators averaged over 200 simulated datasets for each condition discussed in the text. The standardized bias is the bias of the SATE divided by the standard deviation of the treatment effects in the population. RSMSE is the root mean square error of the SATE divided by the standard deviation of the treatment effects in the population. The true SATE refers to the true sample average treatment effects in the sample, computed as the mean of the student treatment effects of the sample. It is not an "estimator" but is computed to show the magnitude of real bias of the sample. Unadjusted ATE refers to the internally valid ATE estimated by a "naive" model that does not take into account sampling bias. IPP-School applies the school-level weight. The IPPSSS refers to the IPP-School+Student separate estimator. It applies the school-level weight and student-level weight estimated by single level propensity score models in each school. IPPSSM refers to the IPP-School+Student multi estimator. It applies the school-level weight and student-level weight estimated by a multilevel propensity score model for all sample schools.

Figure 1.

Standardized bias and RSMSE for true SATE and PATE estimators.



Note. This figure visualizes the data of Table 1. See note under Table 1 for explanation of estimators.