TPM®

EVALUATING METHODS FOR HANDLING MULTILEVEL SELECTION FOR THE PURPOSE OF GENERALIZING CLUSTER RANDOMIZED TRIALS

EVA YUJIA LI
CHRISTOPHER RHOADS
UNIVERSITY OF CONNECTICUT

Over the past decade, the *generalizability* of randomized experiments, defined as the level of consistency between an estimated treatment effect in a nonrandom sample and the true treatment effect in a target population, has received increasing attention from the research community. Existing methods focus on either: (a) prospectively preventing or (b) retrospectively adjusting away, the bias caused by the nonrandom selection of institutions, such as schools, into a study sample. Existing methods overlook the multilevel nature of the selection process that occurs when institutions volunteer for a research study. This study explores methods to adjust away bias caused by this multilevel selection process. A simulation study evaluates the bias reducing properties of different methods for estimating and utilizing inverse probability of participation (IPP) weights to reduce bias. Methods that incorporate both student and school IPP weights reduce more bias than methods that only incorporate the school IPP weights.

Keywords: Generalizability; Multilevel; Randomized trials; Inverse probability weighting; Nonrandom sample.

Correspondence concerning this article should be addressed to Eva Yujia Li, Department of Educational Psychology, University of Connecticut, 239 Glenbrook Road, Storrs, CT 06268, United States of America. Email: eva.li@uconn.edu

Generalization of results from an experiment performed on a nonrandom sample to a population of interest is an important issue for researchers. Policy makers rarely care only about the effect of an intervention on the study sample. Rather, they often reference the average treatment effects estimated by studies to make decisions about the initiation, continuation, or termination of social programs and policies for larger or different populations. Thus, external validity, inference about the extent to which a causal relationship holds over variations in persons, settings, treatments, and outcomes (Shadish, Cook, & Campell, 2002), warrants attention from researchers. In the past, discussions regarding external validity in published randomized experiments were often informal or absent (Blom-Hoffman et al., 2009; Caldwell, Hamilton, Tan, & Craig, 2010; Fernandez-Hermida, Calafat, Becoña, Tsertsvadze, & Foxcroft, 2012). Consequently, it has historically been difficult to gauge the applicability of treatment effects estimated by a randomized experiment to other populations.

Much of the work on generalization of experiments has occurred in the educational research community. In the past decade, this community devoted increasing attention to developing methods for improving the generalizability of randomized controlled trials (RCTs; Hedges, 2013; Olsen, Orr, Bell, & Stuart, 2013). Existing work has focused on the ways in which institutions, for example, schools, that volunteer for experiments differ from those that do not. An implicit assumption is that treatment effects vary only as a function of observable characteristics of schools (school-level moderators). However, for almost all educational RCTs, the study sample is collected in two stages — schools are recruited first, and then

TPM Vol. 27, No. 3, September 2020 – 453-476 – Special Issue – doi:10.4473/TPM27.3.8 – © 2020 Cises

students and/or teachers volunteer for the study. The importance of accounting for nonrandom within school selection processes is evidenced by variations in participation rates across institutions and is an almost inevitable consequence of the need for consent before running research studies.

Bloom-Hoffman et al. (2009) conducted a review of nearly 500 studies of school-based interventions and prevention programs, and found that only 11.5% reported consent procedures and student participation rates. The review does not contain information about school or teacher level participation rates. Of the studies that reported active consent procedures (i.e., required parents to sign consent forms) there was an average student participation rate of 65.5%, with the student participation rate ranging from 11-100% across the studies reviewed. The results of this study bring up two important points for those interested in generalizing RCT results. First, the vast majority of studies did not report participation rates, which makes it difficult to gauge how representative the population studied was of the overall population of eligible students. Second, the wide variation in participation rates suggests that self-selection in to the sample, and therefore, the generalizability of the results, varied considerably across these school-based interventions, and possibly across sites within each intervention.

Additionally, evidence from large scale international assessments shows that student-level nonresponse is related to student characteristics, and in general, less capable students are more likely to be absent from assessments (Rust, 2013). It is plausible, if not likely, that such differential participation related to student characteristics also occurs in RCTs.

There is also reason to believe that teachers who volunteer for research studies are not representative of teachers overall. Kelcey and Phelps (2013) calculated multilevel design parameters of teacher knowledge (e.g., ICC, R^2_{w}, R^2_{b}) using data from several large-scale professional development programs. Despite large sample sizes of these programs, the authors suggested that teacher volunteers in their samples were probably not representative of all teachers nationwide in the United States.

At the moment, the implications for generalization of differential consent rates across different levels of the educational hierarchy is unknown. Interaction effects may be important. For instance, if schools from less privileged districts tend to volunteer for research studies, but more privileged students within those schools volunteer, what does this mean for the generalizability of results? If less experienced teachers from certain districts are more likely to volunteer for research studies, but more experienced teachers from other districts are more likely to volunteer, how will this impact the generalization of the results of RCTs of professional development interventions?

We hope that future empirical research will provide answers to these important questions. The present study considers a multilevel propensity score-based approach to improving generalization from studies with a multilevel selection process. While self-selection into study samples is often a multilevel process with more than two levels (e.g., schools volunteer, teachers within schools consent/volunteer, and parents/students within classrooms consent/volunteer) the current paper only considers two levels of self-selection. This study will extend the existing research on generalization by exploring ways to construct reweighted estimates of treatment effects when there is a two-level selection process. The study proposes to improve the existing inverse probability of participation (IPP) weighting method by adding an additional level of weighting at the within-school level, and then evaluates the proposed method through a simulation study.

EXISTING WORK

Existing work on improving the generalizability of RCTs has suggested two distinct approaches. The first approach is prospective and focuses on coming up with a recruitment plan that will optimize the

generalizability of the estimated sample average treatment effect (SATE) to a population average treatment effect (PATE) of interest (e.g., Tipton, 2013a, 2013b, 2014, Tipton et al., 2014). The second approach is retrospective and focuses on statistical adjustment to understand better how RCT results might apply to different inferential populations of interest (e.g., Chan, 2017, 2018; Cole & Stuart, 2010; Kern, Stuart, Hill, & Green, 2016; O'Muircheartaigh & Hedges, 2014; Stuart, Cole, Bradshaw, & Leaf, 2011). Both approaches assume that treatment effects vary as a function of observable variables that characterize institutions (e.g., school-level moderators). Prospective approaches aim to recruit samples that mirror the population of interest as closely as possible with respect to moderators by minimizing a multivariate distance measure. Retrospective approaches use statistical adjustment to account for the biasing effects of moderating variables in order to obtain an unbiased (or nearly unbiased) estimate of the average treatment effect in an external population. The current paper focuses on retrospective techniques.

Inverse Probability of Participation (IPP) Weighting

IPP weighting is a retrospective adjustment approach which uses estimated participation probabilities to create a synthetic "population" that mimics the target population (Cole & Stuart, 2010; Stuart, Bradshaw, & Leaf, 2015). The first step of the method is to define a *target population* for which inference will be made. For example, the Cognitively Guided Instruction study (Schoen, Lavenia, & Tazaz, 2017) was an RCT conducted with volunteer teachers at 23 participating elementary schools in the state of Florida in the United States. One plausible target population would be all public elementary school math teachers in Florida.

The second step is to collect data on institutions in the study sample and in the target population. Examples of possible variables in the educational context include: school size, number of full-time teachers, percentage of students that qualify for free/reduced priced lunch, and percentage of minority students. Researchers can take advantage of publicly available sources such as the Common Core of Data (CCD), the Stanford Education Data Archive (SEDA), and state-specific data sources (Tipton & Olsen, 2018). These variables must include all covariates that both predict selection into the RCT sample and moderate treatment effects (Tipton, 2013a; Stuart et al., 2011).

The third step is to estimate the selection probability p_{h}, the probability of selecting institution h from a population into a study sample, for each institution in the sample and in the population.

$$p_{\mathrm{h}} = P(S_{\mathrm{h}} = 1|\boldsymbol{V}_{\mathbf{h}}) \tag{1.1}$$

$h = 1, 2, 3, \ldots, H$ is the index of institutions.

$S_{\mathrm{h}} = 1$ if institution h is in the study sample.

$S_{\mathrm{h}} = 0$ if institution h is in the target population but not in the sample.

$\boldsymbol{V}_{\mathbf{h}}$ is a vector of school-level variables which contains *all* variables that explain the selection of institution h into the study sample and the variability of treatment effect in the population.

Typically, p_{h} can be estimated via logistic regression, as in Equation 1.2. Researchers have also applied generalized boosted regression and random forest methods to estimate the probability of participation (Kern et al., 2016). These methods have the advantage of being less sensitive to functional form assumptions compared to logistic regression (Cole & Stuart, 2010; Stuart et al., 2011; Stuart et al., 2015).

$$\ln\left[\frac{p_{\mathrm{h}}}{1-p_{\mathrm{h}}}\right] = \alpha_0 + \alpha_1 V_{1\mathrm{h}} + \alpha_2 V_{2\mathrm{h}} + \cdots + \alpha_{\mathrm{m}} V_{\mathrm{mh}} \tag{1.2}$$

After obtaining the predicted probabilities of participation for each institution, sample observations are weighted by the inverse of their participation probabilities. When the selected study sample is a subset of the target population, the weight for institution h is $w_{\mathrm{h}} = \frac{1}{\widehat{p_{\mathrm{h}}}}$. The researcher should check that the

weighted sample is similar to the target population with regard to observed covariates by computing balance statistics. If covariate balance between the sample and the population is not sufficient, the researcher should try other model specification to estimate w_{h}.

The last step after checking covariate balance is to estimate the population average treatment effect (PATE). Assuming multiple individuals per institution a researcher can fit a weighted multilevel regression model adding IPP weights to the second level (Stuart et al., 2015). A weighted multilevel regression model has the capacity to incorporate unequal selection probabilities for units at each level of sample selection (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). The estimated coefficient for the treatment assignment indicator is the estimated PATE. This estimator is similar to the Hortitz-Thompson (HT) estimator from the survey sampling literature, which is sometimes used to adjust for nonresponse in surveys (Lohr, 2009).

The unconfounded sample selection assumption is the cornerstone of both the prospective and retrospective methods — all covariates that both predict selection into the RCT sample and moderate treatment effects must be included in the construction of distance measures or post-hoc adjustment models (Tipton, 2013a; Stuart et al., 2011). Additionally, assumptions that are required for causal inference in more general settings, that is, the stable unit treatment value assumption (SUTVA) and strongly ignorable treatment assignment in the focal study, must also be satisfied.

The SUTVA idea requires some additional discussion in the context of the current paper. We explore re-weighting methods assuming randomization at the school level. As argued in Rubin, Stuart, and Zanutto (2004) and Hill (2013), when schools are randomized it is not so important that the SUTVA assumption hold within schools in order to produce coherent, internally valid, estimates of treatment effects. However, when attempting to generalize from an experiment on self-selected students to a within-school population of students, we must assume that the potential outcomes of all students in the school do not depend on which students elect to participate in the study. Similarly, when generalizing to a population of schools we must assume that the school average potential outcomes do not depend on which schools elect to participate in the study. These new types of SUTVA assumptions have not previously been discussed in the literature, but are necessary in order to ensure that average treatment effects are well-defined for the entire population of interest. As such, the rest of our paper makes these assumptions.

CURRENT STUDY

The existing IPP weighting method assumes that covariate information that can be used to improve generalizability is only available at a single level of a population that potentially has a multilevel structure. In educational applications to date, only school-level covariate information has been utilized. However, studies should leverage information about nonparticipating individuals within participating institutions, as well as information about nonparticipating institutions, when such information is available. To extend the existing IPP weighting method, a natural choice is to estimate and utilize the individual IPP weights in addition to the institution level IPP weights, which takes into account both within-institution and between-institution selection processes. To fix ideas, the rest of the paper will assume an educational context and use “students” instead of “individuals” and “schools” instead of institutions, however, the conclusions are equally applicable to other multilevel settings.

We define student IPP weight w_{hk} as the inverse of the probability that student k in participating school h will participate. Similar to the steps of estimating the school IPP weight, the first step for estimat-

ing a student IPP weight is to define a *within school target population*. For example, the student sample of the Cognitively Guided Instruction study (Schoen et al., 2017) consisted of volunteer Grades 1 and 2 students in 23 elementary schools. For this study, plausible within school target populations are all Grade 1 and 2 students in each school. The next step is to collect information on students in the study sample and the target within school population. Examples of such information includes demographic variables, pre-intervention achievement measures, and pre-intervention noncognitive measures. For students in the study sample, data are usually collected as a part of the study. For students who are in the participating schools but not in the consenting study sample, their information must be obtained from a different source, for instance, an administrative database maintained by a state department of education.

The third step is to estimate the student sampling propensity score, p_{hk}, defined as the probability of student *k* within school *h* would participate in the study, provided that school *h* participates in the RCT

$$p_{\mathrm{hk}} = P(S_{\mathrm{hk}} = 1 | \boldsymbol{X}_{\mathrm{hk}}, \boldsymbol{V}_{\mathrm{h}}, S_{\mathrm{h}} = 1) \tag{1.3}$$

$h = 1, 2, 3, \ldots, H$ is the index of schools.

$k = 1, 2, 3, \ldots, n_{\mathrm{h}}$ is the index of students in school *h*.

$S_{\mathrm{hk}} = 1$ if the student *k* is in the study sample of school *h*.

$S_{\mathrm{hk}} = 0$ if the student *k* is in the target population within school *h* but not in the sample.

$\boldsymbol{X}_{\mathrm{hk}}, \boldsymbol{V}_{\mathrm{h}}$ contain *all* variables that explain the selection student *k* into the study sample in school *h* and the variability of treatment effect in the within school population.

To estimate p_{hk}, researcher should use a method that takes into consideration the nested structure of students within schools and variation in the selection process across schools. Past research has suggested that participation rates differ across schools (Bloom-Hoffman et al., 2009) and selection of students into studies is related to student and school characteristics (Rust, 2013). In addition, the relationship between student characteristics and selection may vary across schools (Kim & Seltzer, 2007). This study considers two options for estimating p_{hk}, both of which might be termed multilevel propensity scores (Kim & Seltzer, 2007; Li, Zaslavsky & Landrum, 2013; Rosenbaum, 1986). The multilevel propensity score is an extension of the standard propensity score (Rosenbaum & Rubin, 1983; Stuart, 2010) to settings where individuals are clustered in higher level units and selection occurs at the individual level. The multilevel propensity score is defined as the probability of selection given individual and cluster characteristics. Similarly, the student sampling propensity score that we seek to estimate in our study is the probability of participating in the within-school study sample given student and school characteristics.

The first estimation option is to run separate models, such as Equation 1.4, for each participating school *h*. Running one model per school allows for differential within-school selection because each school has its own slope and intercept. The slopes indicate the relationships between student characteristics and selection probability. This approach is only feasible when the within-school sample and population sizes are large enough and when there is sufficient overlap in the distribution of the covariates in the study and population distributions.

$$\ln\left(\frac{p_{\mathrm{hk}}}{1-p_{\mathrm{hk}}} | S_{\mathrm{h}} = 1\right) = \eta_{0\mathrm{h}} + \eta_{1\mathrm{h}} X_{1,\mathrm{hk}} + \eta_{2\mathrm{h}} X_{2,\mathrm{hk}} + \eta_{3\mathrm{h}} X_{3,\mathrm{hk}} + \cdots \tag{1.4}$$

When the above conditions are not met researchers will find it useful to borrow strength from other clusters by using a multilevel random effects model that pools student-level information from all participating schools, such as in Equation 1.5. This model accounts for the variability in the slopes relating student-level covariates to the selection probability across institutions by using school-specific random slopes. The advantage of the random effects model is its robustness in the presence of small schools, because the empirical Bayes estimator in the random effects model allows small clusters to "borrow strength" from large clusters. On the other hand, it has a shortcoming of not guaranteeing balance of the study sample and

target population within each cluster, because the empirical Bayes estimator shrinks the random Level-1 coefficient toward the grand mean (Li et al., 2013, Raudenbush & Bryk, 2002). The study samples within participating schools will be balanced with the target within school populations in participating schools as a whole, but balance is not guaranteed within any school in particular.

$$\ln\left(\frac{p_{\mathrm{hk}}}{1-p_{\mathrm{hk}}}\middle|S_{\mathrm{h}}=1\right)=\eta_{0\mathrm{h}}+\eta_{1\mathrm{h}}X_{\mathrm{hk}} \tag{1.5}$$
$$\eta_{0\mathrm{h}}=\tau_{00}+\tau_{01}V_{\mathrm{h}}$$
$$\eta_{1\mathrm{h}}=\tau_{10}+\tau_{11}V_{\mathrm{h}}$$

After student IPP weights have been estimated, they should be applied to observations in the sample. Researcher should check that the weighted student samples are similar to the within school target populations with respect to observed covariates. Next, student and school IPP weights should be applied to Level 1 and Level 2 of a weighted multilevel regression model, respectively, to obtain an estimate for PATE.

The current study compares the existing IPP weighting method that address school-level selection only with new methods that address both within and between school selection process through a simulation study. The research questions that this study investigates are: (1a) 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 (1b) how much is bias reduced under different simulation scenarios? (2) How do different methods for estimating IPP weights perform under different simulation scenarios?

Simulation Design

Data Generation

Population data is generated in the R software (R Core Team, 2018). Schools have the subscript *h* ($h = 1, \ldots, H$). Students within school *h* have the subscript *k* ($k = 1, \ldots, n_{\mathrm{h}}$, where n_{h} is the total number of students in school *h*, or the within school population size). We generated two school-level random covariates, $V_{1,\mathrm{h}} \sim N(0,1)$, and $V_{2,\mathrm{h}} \sim \mathrm{Bernoulli}(0.5)$. We generate one student-level covariate $X_{\mathrm{hk}} \sim N(V_{1,\mathrm{h}}, 1)$, and the mean of X_{hk} within school *h* equals to the school-level variable $V_{1,\mathrm{h}}$, because student characteristics are usually correlated with school characteristics.

We simulated two potential outcomes for each individual. $y_{\mathrm{hk}}(0)$ is the response when student *k* in school *h* is assigned to the control condition. It depends on school- and student-level covariates, and their interactions, as detailed in Equation 2.1. $y_{\mathrm{hk}}(1)$ is the response when student *k* in school *h* is assigned to the treatment condition. It depends on the student's potential outcome under the control condition, $y_{\mathrm{hk}}(0)$, plus a treatment effect, $\phi_0 + \phi_1 X_{\mathrm{hk}}$. ϕ_0, and ϕ_1 reflect the degree of impact by school-, student-level variables and their interactions on student treatment effects. Specifically, π_{30} is the unconditional average treatment effect; π_{31} and π_{32} are the main effects of school-level variables $V_{1,\mathrm{h}}$ and $V_{2,\mathrm{h}}$ on the average treatment effect; π_{40} is the main effect of the student-level variable X_{hk} on the average treatment effect; π_{41} and π_{42} are the impacts of cross-level interactions on the average treatment effect.

$$y_{\mathrm{hk}}(1)=w_0+w_1X_{\mathrm{hk}}+\phi_0+\phi_1X_{\mathrm{hk}} \tag{2.1}$$
$$w_0=\pi_{00}+\pi_{01}V_{1,\mathrm{h}}+\pi_{02}V_{2,\mathrm{h}}$$
$$w_1=\pi_{10}+\pi_{11}V_{1,\mathrm{h}}+\pi_{12}V_{2,\mathrm{h}}$$
$$\phi_0=\pi_{30}+\pi_{31}V_{1,\mathrm{h}}+\pi_{32}V_{2,\mathrm{h}}$$
$$\phi_1=\pi_{40}+\pi_{41}V_{1,\mathrm{h}}+\pi_{42}V_{2,\mathrm{h}}$$

$$y_{\mathrm{hk}}(0) = w_0 + w_1 X_{\mathrm{hk}}$$
$$w_0 = \pi_{00} + \pi_{01} V_{1,\mathrm{h}} + \pi_{02} V_{2,\mathrm{h}}$$
$$w_1 = \pi_{10} + \pi_{11} V_{1,\mathrm{h}} + \pi_{12} V_{2,\mathrm{h}}$$
$$\text{Treatment effect}_{\mathrm{hk}} = Y_{\mathrm{hk}}(1) - y_{\mathrm{hk}}(0) = \phi_0 + \phi_1 X_{\mathrm{hk}} = \pi_{30} + \pi_{31} V_{1,\mathrm{h}} + \pi_{32} V_{2,\mathrm{h}} + \pi_{40} X_{\mathrm{hk}} + \pi_{41} V_{1,\mathrm{h}} X_{\mathrm{hk}} + \pi_{42} V_{2,\mathrm{h}} X_{\mathrm{hk}} \quad (2.1.1)$$

Next, we used selection models at the school and student levels to determine participating schools and students in the study for each simulated experiment. Schools were selected for the study based on the result of a randomly generated Bernoulli variable S_{h}, $S_{\mathrm{h}} \sim \mathrm{Bernoulli}(p_{\mathrm{h}})$, $S_{\mathrm{h}} = 1$ or 0, indicating whether school *h* is in the sample. The other school level covariate only impacts within school selection probabilities. The average percentage of the schools that select into a study can be determined by the values of coefficients in Equation 2.2. In any one replication, the percentage of schools that self-select into the study is random. Within a replication, for each school, *h*, selected for the study, we generated a treatment indicator $Z_{\mathrm{h}} \sim \mathrm{Bernoulli}(0.5)$ which on average assigns 50% of schools into the treatment group (the simulation study assumes a school-randomized experiment).

$$\ln\left(\frac{p_{\mathrm{h}}}{1-p_{\mathrm{h}}}\right) = \alpha_0 + \alpha_1 V_{1,\mathrm{h}} \quad (2.2)$$

Selection of students within schools into the study sample is determined based on the outcome of the random variable $S_{\mathrm{h}} \sim \mathrm{Bernoulli}(p_{\mathrm{hk}})$, where $p_{\mathrm{hk}} = \mathrm{P}(S_{\mathrm{hk}} = 1 | S_{\mathrm{h}} = 1)$. $S_{\mathrm{hk}} = 1$ or 0, indicating whether student *k* in school *h* is in the sample. Student selection probability is determined by a multilevel logistic regression based on the student-level covariate, school-level covariates, and their interactions. The average percentage of the students that select into the study can be determined by the values of coefficients in Equation 2.3. In any one replication, the percentage of students who self-select into a study is random. Both $V_{1,\mathrm{h}}$ and $V_{2,\mathrm{h}}$ are predictors of the selection of students into the within school study sample in Equation 2.3, but only $V_{1,\mathrm{h}}$ predicts the school selection in Equation 2.2. The distinction is intentional because it is unlikely that the school level variables that predict school and student selection are exactly the same.

$$\ln\left(\frac{p_{\mathrm{hk}}}{1-p_{\mathrm{hk}}}\right) = \eta_{0\mathrm{h}} + \eta_{1\mathrm{h}} X_{\mathrm{hk}} \quad (2.3)$$
$$\eta_{0\mathrm{h}} = \tau_{00} + \tau_{01} V_{1,\mathrm{h}} + \tau_{02} V_{2,\mathrm{h}}$$
$$\eta_{1\mathrm{h}} = \tau_{10} + \tau_{11} V_{1,\mathrm{h}} + \tau_{21} V_{2,\mathrm{h}}$$

True SATE and Estimators of PATE

All estimation was run in STATA 15 SE. First, the true sample average treatment effect (SATE) was calculated. Even though it is not an estimator, we compare it with the estimators to show how much bias is caused by the within and between and school selection processes after eliminating error caused by the randomization process. The true SATE is the average of $y_{hk}(1)$ minus $y_{hk}(0)$ for the selected sample for each replication.

The first estimator that is considered is labelled the *unadjusted ATE*. It is the unadjusted, internally valid estimator of the ATE in the study sample. It is the estimate of γ_{01} in the unweighted hierarchical linear model specified by Equation 2.4. This estimator is internally valid because the treatments are randomly assigned within each study sample for a given replication. It is not externally valid unless samples of schools and students are randomly selected from the population.

$$y_{\mathrm{hk}} = \beta_{0\mathrm{h}} + \varepsilon_{\mathrm{hk}}, \varepsilon_{\mathrm{hk}} \sim N(0, \sigma^2) \quad (2.4)$$
$$\beta_{0\mathrm{h}} = \gamma_{00} + \gamma_{01} Z_{\mathrm{h}} + \mathrm{u}_{0\mathrm{h}},\ \mathrm{u}_{0\mathrm{h}} \sim N(0, \tau)$$

The second estimator considered is labelled *IPP-School*, and is estimated by applying the school-level IPP weights to the second level of the outcome model in Equation 2.4. The school selection probabil-

ity is estimated using a single level logistic regression model as described in Equation 2.2. School weights are computed as the inverse of the selection probabilities, $\widehat{w_{\mathrm{h}}} = \frac{1}{\widehat{p_{\mathrm{h}}}}$.

The third estimator is *IPP-School+Student-separate (IPPSS)*. It is estimated by adding school IPP weights to the second level, and also student IPP weights to the first level of the outcome model in Equation 2.4. School IPP weights are again estimated by Equation 2.2. Student probability of selection into the study is estimated using a separate single-level logistic regression model for each participating school ($S_{\mathrm{h}} = 1$) (Equation 2.5). This predicted selection probability is a conditional probability, $p_{\mathrm{hk|h}} = \Pr(S_{\mathrm{hk}} = 1|S_{\mathrm{h}} = 1)$. These logistic models are estimated using the STATA command *logit* and individual probabilities are predicted by the *pred* command.

$$\ln\left(\frac{p_{\mathrm{hk}}}{1-p_{\mathrm{hk}}}|S_{\mathrm{h}} = 1\right) = \eta_{0\mathrm{h}} + \eta_{1\mathrm{h}}X_{\mathrm{hk}} \tag{2.5}$$

Even though this student selection model does not include any school level covariates that are in the true student selection model (Equation 2.3), it allows each school to have its own intercept $\eta_{0\mathrm{h}}$ and slope $\eta_{1\mathrm{h}}$. We expect that school-specific intercepts η_{0h} will account for the variation in selection probabilities due to school characteristics ($\tau_{00} + \tau_{01}V_{1,\mathrm{h}} + \tau_{02}V_{2,\mathrm{h}}$), and school-specific slopes $\eta_{1\mathrm{h}}X_{\mathrm{hk}}$ will account for the variation due to student characteristics and their interaction with school characteristics ($\tau_{10}X_{\mathrm{hk}} + \tau_{11}V_{1,\mathrm{h}}X_{\mathrm{hk}} + \tau_{12}V_{2,\mathrm{h}}X_{\mathrm{hk}}$).

The fourth estimator is labelled *IPP-School+Student-multi (IPPSSM)*. It is estimated by adding the school IPP weights to the second level, and student IPP weights to the first level of the outcome model in Equation 2.4. School IPP weights are again estimated by Equation 2.2. The student weights are estimated using a multilevel logistic regression model with student-level and school-level covariates (using observations in selected schools only), and their interactions (Equation 2.6). These multilevel logistic models are estimated using STATA command *gllamm* (Rabe-Hesketh, Skrondal, & Pickles, 2004; StataCorp, 2017). Individual probabilities are predicted by the *pred* and *gllapred* commands, respectively.

$$\begin{aligned}\ln(\frac{p_{\mathrm{hk}}}{1-p_{\mathrm{hk}}}|S_{\mathrm{h}} = 1) &= \eta_{0\mathrm{h}} + \eta_{1\mathrm{h}}X_{\mathrm{hk}} \\ \eta_{0\mathrm{h}} &= \tau_{00} + \tau_{01}V_{1,\mathrm{h}} + \tau_{02}V_{2,\mathrm{h}} + \mathrm{u}_{0\mathrm{j}} \\ \eta_{1\mathrm{h}} &= \tau_{10} + \tau_{11}V_{1,\mathrm{h}} + \tau_{21}V_{2,\mathrm{h}} + \mathrm{u}_{1\mathrm{j}}\end{aligned} \tag{2.6}$$

The fifth estimator is called *IPP-School+Student-multi miss (IPPSSM (miss))*. It is estimated by adding the school IPP weights to the second level, and student IPP weights to the first level of the outcome model in Equation 2.4. School IPP weights are again estimated by Equation 2.2. The student weights are estimated using Equation 2.7. The model is missing a school-level covariate $V_{2,\mathrm{h}}$ from Equation 2.6. We hypothesize that a multilevel random effects model protects against missing level-2 covariates because the missing information goes into the cluster specific random slopes and intercepts. In other words, the missing terms $\tau_{02}V_{2,\mathrm{h}}$ and $\tau_{02}V_{2,\mathrm{h}}X_{\mathrm{hk}}$ from Equation 2.6 will add to the variation of the $\mathrm{u}_{0\mathrm{j}}$ and $\mathrm{u}_{1\mathrm{j}}$ terms in Equation 2.7.

$$\begin{aligned}\ln(\frac{p_{\mathrm{hk}}}{1-p_{\mathrm{hk}}}|S_{\mathrm{h}} = 1) &= \eta_{0\mathrm{h}} + \eta_{1\mathrm{h}}X_{\mathrm{hk}} \\ \eta_{0\mathrm{h}} &= \tau_{00} + \tau_{01}V_{1,\mathrm{h}} + \mathrm{u}_{0\mathrm{j}} \\ \eta_{1\mathrm{h}} &= \tau_{10} + \tau_{11}V_{1,\mathrm{h}} + \mathrm{u}_{1\mathrm{j}}\end{aligned} \tag{2.7}$$

The *mixed* procedure in STATA 15 was used to run all the multilevel (weighted) outcome models, due to its capability of correctly handling survey weights (West & Galecki, 2011). All weights are specified as sampling weights.

Simulation Conditions

The simulation varies the population size, the within school participation rate, and the random/nonrandom selection process. Table 1 shows a summary of the simulation conditions. Table 2 shows the parameter values that correspond to each simulation condition.

TABLE 1
Simulation conditions

	Conditions	Explanation
Population size	H = 2,000, $n_h \sim N(200, 80)$, $\min = 10, \max = 700$	Two thousand schools in the population; simulates schools in a State as the population.
	H = 50, $n_h \sim N(250, 60)$, $\min = 25, \max = 330$	Fifty schools in the population; simulates a midsized school district as the population.
Population treatment effect distribution	TE main effect	In the population, treatment effect is impacted by the characteristics of schools and students. The strength of these impacts are constant in all schools.
	TE interaction	In the population, treatment effect is impacted by the characteristics of schools and students. The strength of the impacts of the student characteristics depend on the characteristics of the school the student is in.
Sample selection process	Random school and student	Schools are randomly selected into the study. Within those schools, students are randomly selected into the study.
	Nonrandom school, random student	Schools volunteer into the study with unequal probabilities. Within those schools, students are randomly selected into the study.
	Nonrandom school and student	Schools volunteer into the study with unequal probabilities. Within those schools, students volunteer into the study with unequal probabilities.
	Nonrandom school, nonrandom student, and interaction	Schools volunteer into the study with unequal probabilities. Within those schools, students volunteer into the study with unequal probabilities. The impact of student characteristics on the probability of volunteering into a study depends on the characteristics of the school the student is in.
Participation rate	School 12%, within school 50%	On average over replications, 12% schools participate in the study; 50% of students in each participating school participate in the study.
	School 12%, within school 25%	On average over replications, 12% schools participate in the study; 25% student in each participating school participate in the study.

Population size. We generated two different populations, corresponding to two different, potentially policy relevant populations of interest. The first is a large population of H = 2,000 schools. The school size follows a truncated normal distribution with a mean of 200, standard deviation of 80, a minimum of 10, and maximum of 700 students. The distribution mimics the population of Grade 1 and Grade 2 students in public elementary schools in the state of Florida in the United States. The second population is a small group of H = 50 schools, which mimics the population of Grade 1 and Grade 2 students in a single large school district in the state of Florida.

(1) $H = 2000,\ n_h \sim N(200, 80),\ \min = 10,\ \max = 700$
(2) $H = 50, n_h \sim N(250, 60),\ \min = 25,\ \max = 330$

TPM Vol. 27, No. 3, September 2020
453-476 – Special Issue
© 2020 Cises

TABLE 2

Simulation results for unbalanced sample parameter values for Equations 2.1.1, 2.2, and 2.3 in correspondence with simulation conditions

A. Population size $H = 50$; Population treatment effect model = TE main effect.

$H = 50$	Parameters	Sample Selection Process							
		Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction	Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction
Population treatment effect model (Equation 2.1.1)	π_{30}	1	1	1	1	1	1	1	1
	π_{31}	1	1	1	1	1	1	1	1
	π_{32}	1	1	1	1	1	1	1	1
	π_{40}	1	1	1	1	1	1	1	1
	π_{41}	0	0	0	0	0	0	0	0
	π_{42}	0	0	0	0	0	0	0	0
School selection parameters (Equation 2.2)	α_0	–2	–2.5	–2.5	–2.5	–2	–2.5	–2.5	–2.5
	α_1	0	1	1	1	0	1	1	1
Student selection parameters (Equation 2.3)	τ_{00}	0	0	-4	-5	-1	-1	-6	-9
	τ_{01}	0	0	1	1	0	0	1	1
	τ_{10}	0	0	2	2	0	0	2	2
	τ_{11}	0	0	0	1	0	0	0	1
	τ_{02}	0	0	1	1	0	0	1	1
	τ_{12}	0	0	0	1	0	0	0	1
Participation rates		School 12%, student 50%				School 12%, student 25%			

B. Population size H = 50; Population treatment effect model = TE interaction.

H = 50	Parameters	Sample Selection Process							
		Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction	Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction
Population treatment effect model (Equation 2.1.1)	π_{30}	0	0	0	0	0	0	0	0
	π_{31}	1	1	1	1	1	1	1	1
	π_{32}	1	1	1	1	1	1	1	1
	π_{40}	1	1	1	1	1	1	1	1
	π_{41}	1	1	1	1	1	1	1	1
	π_{42}	1	1	1	1	1	1	1	1
School selection parameters (Equation 2.2)	α_0	–2	–2.5	–2.5	–2.5	–2	–2.5	–2.5	–2.5
	α_1	0	1	1	1	0	1	1	1
Student selection parameters (Equation 2.3)	τ_{00}	0	0	–4	–5	–1	–1	–6	–10
	τ_{01}	0	0	1	1	0	0	1	1
	τ_{10}	0	0	2	2	0	0	2	2
	τ_{11}	0	0	0	1	0	0	0	1
	τ_{02}	0	0	1	1	0	0	1	1
	τ_{12}	0	0	0	1	0	0	0	1
Participation rates		School 12%, student 50%				School 12%, student 25%			

C. Population size = 2,000; Population treatment effect model = TE main effect

H = 2,000	Parameters	Sample Selection Process							
		Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction	Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction
Population treatment effect model (Equation 2.1.1)	π_{30}	1	1	1	1	1	1	1	1
	π_{31}	1	1	1	1	1	1	1	1
	π_{32}	1	1	1	1	1	1	1	1
	π_{40}	1	1	1	1	1	1	1	1
	π_{41}	0	0	0	0	0	0	0	0
	π_{42}	0	0	0	0	0	0	0	0
School selection parameters (Equation 2.2)	α_0	–2	–2.5	–2.5	–2.5	–2	–2.5	–2.5	–2.5
	α_1	0	1	1	1	0	1	1	1
Student selection parameters (Equation 2.3)	τ_{00}	0	0	–3	–4	–1	–1	–5.5	–8
	τ_{01}	0	0	1	1	0	0	1	1
	τ_{10}	0	0	2	2	0	0	2	2
	τ_{11}	0	0	0	1	0	0	0	1
	τ_{02}	0	0	1	1	0	0	1	1
	τ_{12}	0	0	0	1	0	0	0	1
Participation rates		School 12%, student 50%				School 12%, student 25%			

D. Population size = 2,000; Population treatment effect model = TE interaction

H = 2,000	Parameters	Sample Selection Process							
		Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction	Random school and student	Nonrandom school, random student	Nonrandom school and student	Nonrandom school, non-random student, and interaction
Population treatment effect model (Equation 2.1.1)	π_{30}	0	0	0	0	0	0	0	0
	π_{31}	1	1	1	1	1	1	1	1
	π_{32}	1	1	1	1	1	1	1	1
	π_{40}	1	1	1	1	1	1	1	1
	π_{41}	1	1	1	1	1	1	1	1
	π_{42}	1	1	1	1	1	1	1	1
School selection parameters (Equation 2.2)	α_0	–2	–2.5	–2.5	–2.5	–2	–2.5	–2.5	–2.5
	α_1	0	1	1	1	0	1	1	1
Student selection parameters (Equation 2.3)	τ_{00}	0	0	–3	–4	–1	–1	–5.5	–8
	τ_{01}	0	0	1	1	0	0	1	1
	τ_{10}	0	0	2	2	0	0	2	2
	τ_{11}	0	0	0	1	0	0	0	1
	τ_{02}	0	0	1	1	0	0	1	1
	τ_{12}	0	0	0	1	0	0	0	1
Participation rates		School 12%, student 50%				School 12%, student 25%			

Population treatment effect model. In the population, the treatment effects are moderated by the school level covariates, student level covariates, and their interactions, as shown in Equation 2.1. We specified two conditions for the treatment effect model, the *TE main effect* condition and the *TE interaction* condition. In the TE main effect condition, the treatment effects are predicted by school and student level covariates, but there is no interaction effect. It means that the strength of the impact of the school and student characteristics on individual level treatment effects are equal in all schools. We set the value of $\pi_{30} = \pi_{31} = \pi_{32} = \pi_{40} = 1$ and $\pi_{41} = \pi_{42} = 0$ in Equation 2.1.1, resulting in the following equation for the treatment effect:

$$\text{Treatment effect}_{\text{hk}} = 1 + V_{1,\text{h}} + V_{2,\text{h}} + X_{\text{hk}} \quad (2.8)$$

In the TE interaction condition, the treatment effects are predicted by school and student level covariates, and their interactions. It means that the strength of the impact of the student characteristics on individual level treatment effects depend on the school's characteristics. In this scenario, the inclusion of student IPP weights may be more important because the student characteristics lead to greater difference in school-specific average treatment effects between schools with low and high values of $V_{1,\text{h}}$ and $V_{2,\text{h}}$. We set the value of $\pi_{30} = 0$ and $\pi_{31} = \pi_{32} = \pi_{40} = \pi_{41} = \pi_{42} = 1$ in Equation 2.1.1, resulting in the following equation for the treatment effect:

$$\text{Treatment effect}_{\text{hk}} = V_{1,\text{h}} + V_{2,\text{h}} + X_{\text{hk}} + V_{1,\text{h}}X_{\text{hk}} + V_{2,\text{h}}X_{\text{hk}} \quad (2.9)$$

Holding other factors constant, the true PATE for the TE main effect condition and the TE interaction condition are approximately the same (differing only as a result of simulation error in the population level simulations). We expected that for the TE interaction condition, inclusion of student IPP weights will reduce more bias than for the TE main effects condition.

Sample selection processes. We varied the sample selection processes at the school and student levels by varying the coefficient value of the school and student covariates in Equations 2.2 and 2.3. There are four sample selection conditions. First, the *random school and student selection* condition selects school and student randomly at both stages. The parameters in the selection models (Equations 2.2 and 2.3) are set to be zero except for the intercepts α_0 and τ_{00}. Second, the *nonrandom school, random student* condition sets $\alpha_1 = 1$. All other parameters in Equation 2.3 are set to be zero except the intercept τ_{00}. Third, the *nonrandom school, nonrandom student* condition again sets $\alpha_1 = 1$ in Equation 2.2. It also sets $\tau_{01} = \tau_{02} = 1, \tau_{10} = 2$ and $\tau_{11} = \tau_{12} = 0$ in Equation 2.3. By setting $\tau_{11} = \tau_{12} = 0$, student selection into the within school sample are affected by the student's characteristics and the characteristics of the school, and the effects are the same across all schools. Fourth, the *nonrandom school, nonrandom student, interaction* condition again sets $\alpha_1 = 1$ in Equation 2.2. It sets $\tau_{01} = \tau_{02} = \tau_{11} = \tau_{12} = 1, \tau_{10} = 2$ in Equation 2.3. Student selection into the within school sample is affected by the student's characteristics and the characteristics of the school, and the strength of the impact of student characteristics depend on the characteristics of the school the student is in.

Within school participation rates. In this study, the selection of a school into a study is a Bernoulli random variable. The selection of a student into the study in his or her school is also a Bernoulli random variable. Therefore, for any particular replication, the number of schools that select into the sample and the number of students that select into the within school samples are random. The school and within school selection rates were set by varying the values of intercepts (α_0 and τ_{00}) in the selection models, Equations 2.2 and 2.3. The specific values of these parameters were varied for the different simulations, as shown in Table 2. Each condition requires different intercept values to achieve the desired participation rates, because the values of other parameters are different across conditions and the intercept values have to be adjusted accordingly. The value of α_0 was selected so that in each simulation condition, approximately 12% of schools select into the sample across replications. The value of τ_{00} was set so that in each simulation

condition, the average percentage of students who select into the within school sample across all schools that select into the sample in that particular replication, is approximately 25% or 50%.

Due to the fact that the selection of a school is a Bernoulli random variable, for a particular replication, there is a nonzero probability that zero schools are selected into the sample. To solve this problem, in the large school population (H = 2,000) conditions, the selection of schools was repeated until at least 10 schools are selected. In the small school population size (H = 50) conditions, the simulation procedure repeated the Bernoulli trials until exactly six schools were selected into the sample. In the treatment assignment step, we randomly assigned three schools to the treatment and three schools to the control condition.

Evaluation criteria. For each condition, 200 replications were simulated. The estimators were evaluated by computing the mean standardized bias and root standardized mean square error (RSMSE). The bias and root mean squared error were divided by the standard deviation of the real treatment effects in the population to make the magnitude of the bias and RMSE comparable across simulation conditions.

Simulation Results

Sample Selection Conditions

The standardized bias and RSMSE of the estimators are shown in Table 3.

Random school and student. When the selection process is random at both the school and the student level, all estimators perform similarly well. The standardized biases of all estimators are less than 0.1 standard deviations away from the true PATE when school population size is H = 50 and less than 0.01 standard deviations away from the true PATE when school population size is H = 2,000. The RSMSE is less than one standard deviation in all cases. All estimators have bigger RSMSEs than the true SATE because the true SATE is only affected by error from the within and between selection process, while other estimators are also affected by the randomization error and the estimation error. The standardized bias and RSMSE of the unadjusted ATE and the IPP weighted estimators are similar. The unadjusted ATE has slightly smaller standardized bias than the IPP weighted estimators, but the difference is within 0.01 standard deviations of the treatment effects in their respective populations. This result suggests that even when the school and student samples are both randomly selected, applying the IPP weights does not damage the performance of the estimators.

Nonrandom school, random student. When the selection process is nonrandom at the school level and random at the student level, the IPP-school estimator has smaller standardized bias than the unadjusted ATE in both population sizes. The three IPP-school+student estimators show similar performances to the IPP-School estimator. The RSMSE of the IPP-school compared to the unadjusted ATE show different patterns in the two population sizes. When the population size is large (H = 2,000), the RSMSE of the IPP-school is substantially smaller than the unadjusted ATE, with its magnitude being a quarter of the RSMSE of the unadjusted ATE. When population size is small (H = 50), the RSMSE of the IPP-school, surprisingly, is larger than that of the unadjusted ATE, indicating that the unadjusted ATE is the more accurate estimator. This result can be attributed to the trade-off between bias and variance. The IPP-school is less biased than the unadjusted ATE in both populations because the IPP school weights adjusts away bias caused by the nonrandom selection of the school sample. On the other hand, adding weights to the sample increases variance (Lohr, 2009). When the school population size is 50 and school participation rate is 12%,

TABLE 3
Standardized bias and RSMSE for true SATE and PATE estimators

Population parameters	Within school participation rate	Estimators	Sample Selection Process							
			Random school and student		Nonrandom school random student		Nonrandom school and student		Nonrandom school nonrandom student interaction	
			Std. Bias	RSMSE	Std. Bias	RSMSE	Std. Bias	RSMSE	Std. Bias	RSMSE
$H = 50$ TE main effects	50%	True SATE	0.001	0.361	0.765	0.820	1.451	1.472	1.527	1.533
		Unadjusted ATE	0.017	0.806	0.826	1.183	1.212	1.835	1.049	1.683
		IPP-School	–0.018	0.779	0.549	1.294	0.913	1.645	0.765	1.708
		IPPSSS	–0.019	0.778	0.548	1.292	0.614	1.462	0.476	1.601
		IPPSSM	–0.022	0.785	0.538	1.295	0.608	1.468	0.483	1.572
		IPPSSM (miss)	–0.018	0.779	0.549	1.294	0.625	1.470	0.489	1.571
	25%	True SATE	0.003	0.367	0.765	0.821	1.697	1.716	1.852	1.856
		Unadjusted ATE	0.014	0.805	0.832	1.189	1.332	1.913	1.488	2.014
		IPP-School	–0.020	0.780	0.556	1.300	1.081	1.740	1.384	2.130
		IPPSSS	–0.019	0.778	0.547	1.291	0.793	1.546	1.108	1.901
		IPPSSM	–0.024	0.787	0.541	1.297	0.781	1.512	1.097	1.898
		IPPSSM (miss)	–0.020	0.782	0.553	1.298	0.789	1.514	1.114	1.905
$H = 50$ TE interaction	50%	True SATE	–0.024	0.334	0.751	0.830	1.733	1.795	1.760	1.775
		Unadjusted ATE	–0.018	0.750	0.780	1.120	1.215	1.673	1.101	1.511
		IPP-School	–0.093	0.719	0.434	1.156	0.737	1.294	0.685	1.352
		IPPSSS	–0.094	0.718	0.433	1.152	0.429	1.117	0.396	1.230
		IPPSSM	–0.092	0.723	0.426	1.160	0.428	1.122	0.395	1.201
		IPPSSM (miss)	–0.092	0.720	0.434	1.155	0.440	1.122	0.399	1.201
	25%	True SATE	–0.021	0.338	0.750	0.831	2.193	2.254	2.470	2.480
		Unadjusted ATE	–0.021	0.748	0.785	1.126	1.413	1.824	2.032	2.265
		IPP-School	–0.095	0.721	0.441	1.162	0.971	1.455	1.863	2.220
		IPPSSS	–0.095	0.719	0.432	1.151	0.623	1.233	1.476	1.861
		IPPSSM	–0.095	0.724	0.429	1.162	0.601	1.189	1.422	1.835
		IPPSSM (miss)	–0.095	0.722	0.437	1.159	0.606	1.188	1.452	1.842

(Table 3 continues)

Table 3 (continued)

Population parameters	Within school participation rate	Estimators	Random school and student		Nonrandom school random student		Nonrandom school and student		Nonrandom school nonrandom student interaction	
			Std. Bias	RSMSE	Std. Bias	RSMSE	Std. Bias	RSMSE	Std. Bias	RSMSE
H = 2,000 TE main effects	50%	True SATE	–0.003	0.055	0.735	0.737	1.441	1.442	1.420	1.421
		Unadjusted ATE	–0.003	0.134	0.708	0.753	1.072	1.100	1.094	1.110
		IPP-School	0.004	0.133	0.010	0.208	0.611	0.649	0.684	0.730
		IPPSSS	0.004	0.133	0.010	0.207	0.304	0.358	0.396	0.455
		IPPSSM	0.004	0.133	0.010	0.208	0.319	0.375	0.405	0.465
		IPPSSM (miss)	0.004	0.133	0.010	0.208	0.320	0.376	0.405	0.465
	25%	True SATE	–0.002	0.055	0.736	0.740	1.821	1.821	1.824	1.824
		Unadjusted ATE	0.001	0.132	0.695	0.733	1.313	1.338	1.487	1.499
		IPP-School	0.008	0.131	0.010	0.181	0.944	0.969	1.306	1.323
		IPPSSS	0.008	0.131	0.008	0.182	0.637	0.662	1.035	1.052
		IPPSSM	0.007	0.131	0.011	0.181	0.643	0.673	1.052	1.071
		IPPSSM (miss)	0.007	0.131	0.011	0.181	0.644	0.674	1.053	1.072
H = 2,000 TE interaction	50%	True SATE	–0.002	0.053	0.773	0.778	1.770	1.772	1.713	1.714
		Unadjusted ATE	0.005	0.117	0.725	0.759	1.145	1.166	1.169	1.181
		IPP-School	0.010	0.117	0.013	0.142	0.427	0.461	0.603	0.652
		IPPSSS	0.010	0.117	0.013	0.142	0.118	0.186	0.265	0.344
		IPPSSM	0.009	0.117	0.013	0.143	0.136	0.202	0.278	0.351
		IPPSSM (miss)	0.009	0.117	0.013	0.143	0.136	0.201	0.278	0.352
	25%	True SATE	–0.002	0.053	0.771	0.775	2.527	2.528	2.526	2.527
		Unadjusted ATE	0.005	0.116	0.726	0.759	1.512	1.530	1.805	1.814
		IPP-School	0.010	0.117	0.012	0.143	0.854	0.874	1.440	1.462
		IPPSSS	0.010	0.116	0.011	0.141	0.449	0.473	1.010	1.034
		IPPSSM	0.010	0.117	0.012	0.143	0.440	0.469	1.018	1.042
		IPPSSM (miss)	0.009	0.117	0.012	0.143	0.440	0.470	1.020	1.044

Note. The table shows standardized bias and RSMSE of the true SATEs and five 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 SATE refers to the true sample average treatment effects in 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. IPPSSM (miss) refers to the IPP-school+student-multi miss estimator. It applies the school-level weight and student-level weight estimated by a multilevel propensity score model that omits one school-level covariate.

the sample only consists of six schools. The small number of clusters combined with the addition of sampling weights increased the variance of the estimators enough to overcome the decrease in bias in the RSMSE. When the school population size is 2,000 and the school participation rate is 12%, the average school sample consists of 240 schools. The large cluster sample size offset the increase in the variance of the estimators due to weights, and as a result most of the mean squared error is due to bias.

Nonrandom school and student. When the selection processes are nonrandom at both the school and student level, and the selection probabilities was predicted by the main effects only, the three IPP-school+student estimators outperform the IPP-school and the unadjusted ATE estimators with respect to both standardized bias and RSMSE. The reduction in standardized bias and RSMSE is on the order of 20 to 40% when H = 50, and 40 to 70% when H = 2,000. Amongst the three IPP-school+student estimators, performance is similar. The IPPSSM (miss) always underperform compared to the IPPSSM since the former is missing a school-level covariate in the model for estimating the student IPP weight. However, the difference in performance is small. When school population size is 50, the difference in standardized bias between the two estimators is less than 0.02 standard deviations. When school population size is 2,000, the performances of these two estimators are almost exactly the same. This result suggests that the misspecification of the model for estimating the student IPP weights is offset by the inclusion of random intercepts and slopes. The performance of IPPSSS compared to the IPPSSM depends on school population size. When school population size is 50, the IPPSSM has smaller standardized bias than the IPPSSS. The IPPSSM also has smaller RSMSE than the IPPSSS when the within school participation rate is 25%, but not necessarily when the within school participation rate is 50%. When school population size is 2,000, the IPPSSS has smaller standardized bias and RSMSE than the IPPSSM across most conditions. The order is reversed, however, under the TE interaction and within school participation rate is 25%. The difference in performance between these two estimators is small. The difference in standardized bias and RSMSE between these two estimators is less than 0.04 standard deviations in all population size, population treatment effect and participation rate conditions. The observed patterns were expected, as the IPPSSM method is more advantaged when there are fewer schools and smaller within school sample sizes.

*Nonrandom school, nonrandom student, interaction***.** When the selection processes are nonrandom at both school and student levels, and student selection is predicted by school and student characteristics and their interaction, the three IPP-school+student estimators again perform better than the IPP-school and the unadjusted ATE. The performance of the IPP-school compared to the unadjusted ATE depends on the school population size and population treatment effect distribution. When school population size is 50, the IPP-school always has smaller standardized bias than the unadjusted ATE. The IPP-school estimators have larger RSMSE than the unadjusted ATEs under TE main effects. When the school population size is 2,000, the IPP-school always perform better than the unadjusted ATE, having smaller standardized bias and RSMSE. Amongst the three estimators that adjust for within school selection, the performance was similar. The IPPSSM (miss) always underperforms compared to the IPPSSM, due to the obvious reason that the former is misspecified in the model for estimating the IPP student weight. The relative performance of the IPPSSS and IPPSSM depends on the school population size. When the school population size is 50, the IPPSSM has lower standardized bias than the IPPSSS when the within school participation rate is 25%, and similar or higher standardized bias than the IPPSSS when the within school participation rate is 50%. The IPPSSM has lower RSMSE than IPPSSS when the population size is 50. The difference in the performance of these three estimators is small — within 0.1 standard deviations of the treatment effects in their respective populations. When school population size is 2,000, the IPPSSS performs slightly better than the IPPSSM across population treatment effect models and within school participation rates, but the difference

is less than 0.02 standard deviations of treatment effects in their respective populations. Again, the observed patterns were expected, as the IPPSSM method is more advantaged when there are fewer schools and smaller within school sample sizes.

Within School Participation Rate

When the sample selection process is random at the student level, each estimator has similar standardized bias and RSMSE under the 50% and 25% within school participation rate conditions. As long as within school selection is random, smaller within school participation rates have little impact on the bias of the estimators. When the sample selection process is nonrandom at student level, each estimator performs better when the within school participation rate is 50% than when it is 25%. This is because the 25% within school participation rate selects not only fewer students per schools, but also more biased student samples.

TE Main Effect versus TE Interaction

Averaging across the other conditions, the TE interaction conditions have higher standardized bias and RSMSE in the SATE than the TE main effects conditions. Comparing performance of the same estimator between TE main effect and TE interaction while holding all other conditions constant, the unadjusted ATE generally performs better in the TE main effects condition and the IPP weighted estimators generally perform equally well or better under the TE interaction condition. The performance of the unadjusted ATE can be easily explained by the fact that the TE interaction conditions have higher standardized bias and RSMSE in the SATE to begin with than the TE main effects conditions. Consequently, the superior performance of the IPP weighted estimators under the TE interaction conditions compared to the TE main effect conditions means that the IPP weights are able to reduce more bias under the TE interaction conditions than under the TE main effect conditions. The stronger reduction in bias by the IPP weights can be explained by the fact that under TE interaction conditions, the individual treatment effects in the population are more dependent upon the school and student-level covariates, and adjusting for the nonrandom selection of schools and students is more impactful in reducing bias and improving the accuracy of the estimates of the PATE. The only exception is when the school population size is 50, sample selection process is nonrandom school, nonrandom student, and interaction, and the within school participation rate is 25%. Under these conditions, each estimator has smaller standardized bias in the TE main effect than in the TE interaction condition. This condition may be anomalous because the selection probabilities in this condition are highly variable and the average participation rate is low, thus yielding schools with very small within school sample sizes. The combination of small between and within school sample sizes may make it too difficult to estimate weights accurately. Since standardized bias in the SATE in the TE interaction condition is higher to begin with, it remained higher after adjustment when the weights are imprecisely estimated.

School Population Size

The performance ranking of estimators within each condition are for the most part the same in the H = 2,000 and H = 50 conditions, with a few aforementioned exceptions. All estimators perform remarkably better in the H = 2,000 than in the H = 50, with smaller standardized bias and RSMSE. This is the result of larger sample sizes in the H = 2,000 condition.

Summary of Simulation Results

This simulation study found that applying IPP weights that account for both levels of a multilevel selection process generally improves the performance of the estimators of PATE. When schools are non-randomly selected, applying the school IPP weight reduces the standardized bias of the estimator compared to the unadjusted ATE. However, when the number of schools in the study is small the increased variance due to reweighting means that no improvement in RSMSE is observed. The RSMSEs of the rewighted estimators are smaller than the unadjusted ATE only when the number of schools in the experimental sample is large. When schools and students are both nonrandomly selected, applying both the school IPP weight and the student IPP weight improves the performance of the estimator relative to the IPP-school and the unadjusted ATE estimators. Therefore, applying IPP weights always reduces the standardized bias in the estimator for PATE, when the sample is nonrandomly selected. This is true regardless of how the IPP weights are constructed. However, reweighting increases the RSMSE of the estimator through the inflation of variance if the sample size is small.

The model for estimating the student IPP weights has little impact on the performance of these weights. IPPPSSM performs better than IPPSSS when the school population size is small and the reverse is true when the school population size is large. IPPSSM (miss) shows slightly worse performance compared to IPPSSM, but the difference is small. This indicates that the random effects model for estimating student probability of participation provides protection against missing school-level covariates due to school-specific random intercepts and slopes. In addition, the IPPSSS estimator also protects against missing school-level covariates because it does not need school level covariates in the model (since different models are used for each school).

Given the same school-level participation rate of 12%, all estimators perform better when the school population size is large, which can be explained by the larger sample sizes of schools. The smaller (25%) within school participation rates in this study leads to smaller and more biased samples. Consequently, the estimators perform less well in the 25% within school participation rate conditions than in the 50% conditions. However, the within school participation rate does not affect the order of the performance rankings among the estimators.

Under the TE interaction conditions, the IPP weighted estimators generally perform better than under TE main effect conditions, and have larger reduction in standardized bias and RSMSE compared to the unadjusted ATE. This is because under the TE interaction conditions, the individual treatment effects in the population are more impacted by school and student characteristics than in the TE main effect conditions. Consequently, adjusting away bias caused by nonrandom selection is more effective in improving the accuracy of the estimator for PATE. The distribution of treatment effects in the population does not affect the order of the performance rankings among the estimators.

Discussion

This study explored methodological approaches for handling multilevel selection of samples into randomized controlled trials for the purpose of generalizing treatment effect estimates to a target population. The simulation study shows that when the within school sample is not randomly selected ignoring the within school selection process leads to bias in the estimated population average treatment effect unless statistical adjustment are used. Furthermore, the two estimators that involve student IPP weights (IPPSSS

and IPPSSM), applied in addition to the school IPP weights, significantly reduce bias compared to applying the school IPP weights alone. In addition, both estimators are robust to missing school-level covariates in the student selection model. The IPPSSS does not directly use school-level covariates in models, since a separate student selection model is estimated in each participating school. The IPPSSM protects against missing school-level covariates in the student selection model because the multilevel model has school-specific intercepts and slopes.

The simulation study also shows that small sample sizes create challenges for estimating PATE through retrospective adjustment, because the variance inflation of the estimate may override the reduction in bias. While the large school population condition (H = 2,000, school sample size ≈ 120) clearly show large reduction in both standardized bias and RSMSE, the small school population condition (H = 50, school sample size = 6) has some reduction in standardized bias, but much smaller reduction in RSMSE. In one particular condition, the RSMSE of the IPP-school is larger than the unadjusted RSMSE, suggesting that not applying any adjustment at all would actually be the best choice for this condition. The precise school and students level sample sizes needed for the IPP weights to effectively reduce both bias and RMSE is a topic for future research.

This study has several implications for future research. The results of the simulation study showed that, while IPP weights always substantially decreases standardized bias, they often have less of an effect on the RSMSE. This shows the possible effectiveness of utilizing the stable weights developed by Zubizarreta (2015), which should, in theory, optimally balance the bias-variance trade-off. In the case of limited covariate overlap between a sample and a target population, stable weights may reduce extreme weights and the resultant inflation of standard errors. In addition, recent research on propensity score methods showed that machine learning methods outperform logistic regression models in terms of bias reduction and mean squared error under conditions of nonlinearity and nonadditivity (Lee, Lessler, & Stuart, 2010). These methods, however, have to be adapted for the multilevel setting.

This study has several limitations. First, this study looked at a two-level selection process, but nonrandom within school selection can occur due to both teachers and students (i.e., a three-level selection process). Exploring two-level selection processes is the first step in understanding multilevel selection processes for generalizing from experiments. Future work should expand this study to think about a three-level selection process with students within teachers and teachers within schools.

Second, the methods explored in this study rely on the strongly ignorable sample selection assumption, which cannot be verified empirically. Nguyen, Ebnesajjad, Cole, & Stuart (2017) did a sensitivity analysis for an unobserved school-level moderator, and future research should further investigate the impact of unobserved student-level moderators. For the conditions explored in our study, adding a student-level variable that is correlated with both selection and outcomes led to substantially more bias reduction than using only a school-level model for adjustment. However, it is possible that there are situations not explored in our simulations where adjustments meant to reduce bias due to student-level observables could inadvertently increase bias due to unobserved confounding variable. By analogy to the literature on observational studies (see, e.g., Pearl, 2011), one can surmise that bias amplification is likely to occur when: (i) there is an observed covariate that strongly predicts selection in to the study but is unrelated to variation in treatment effects and (ii) there is an unobserved covariate that is unrelated to selection in to the study or to the observed covariate but is strongly related to variation in treatment effects. Future studies should explore situations where bias amplification might occur.

Third, the selection and outcome models specified in the simulation study are linear and have only three covariates. In reality, there may be many more covariates at each level, which may include linear or

nonlinear predictors and multiple interactions. Correct model specification involves selection of variables, interactions and polynomial effects at each level. In addition, correct estimation of more complex models may be computationally intensive and there may be convergence issues involved when estimating the necessary multilevel logistic models and weighted linear multilevel outcome models. In the case of many potential confounders, interaction terms and polynomial effects, methods with automated variable selection can be applied (e.g., generalized boosted models).

Fourth, all existing methods and the methods proposed condition on the estimated IPP weights. However, the weights themselves are random variables estimated from the data and so are estimated with uncertainty. Lastly, the simulation study is limited by the particular design factors that were chosen, such as the particular school and student population sizes and the particular distribution of treatment effects in the population. In particular, different results may emerge if smaller within school population sizes were used, and if the interaction terms in the selection/outcome models differ in sign from the main effect terms.

REFERENCES

Blom-Hoffman, J., Leff, S. S., Franko, D. L., Weinstein, E., Beakley, K., & Power, T. J. (2009). Consent procedures and participation rates in school-based intervention and prevention research: Using a multicomponent, partnership-based approach to recruit participants. *School Mental Health*, *1*(1), 3-15. doi:10.1007/s12310-008-9000-7

Caldwell, P. H. Y., Hamilton, S., Tan, A., & Craig, J. C. (2010) Strategies for increasing recruitment to randomized controlled trials: Systematic review. *PLoS Med*, *7*(11), e1000368. doi:10.1371/journal.pmed.1000368

Chan, W. (2017). Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, *10*(3), 646-669. doi:10.1080/19345747.2016.1273412

Chan, W. (2018). Applications of small area estimation to generalization with subclassification by propensity scores. *Journal of Educational and Behavioral Statistics*, *43*(2), 182-224. doi:10.3102/1076998617733828

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. doi:10.1093/aje/kwq084

Fernandez-Hermida, J. R., Calafat, A., Becoña, E., Tsertsvadze, A., & Foxcroft, D. R. (2012). Assessment of generalizability, applicability and predictability (GAP) for evaluating external validity in studies of universal family-based prevention of alcohol misuse in young people: Systematic methodological review of randomized controlled trials. *Addiction*, *107*(9), 1570-1579. doi:10.1111/j.1360-0443.2012.03867.x

Hedges, L. V. (2013). Recommendations for practice: Justifying claims of generalizability. *Educational Psychology Review*, *25*(3), 331-337. doi:10.1007/s10648-013-9239-x

Hill, J. (2013). Multilevel models and causal inference. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The SAGE handbook of multilevel modeling* (pp. 201-220). SAGE Publications Ltd. doi:10.4135/9781446247600

Kelcey, B., & Phelps, G. (2013). Considerations for designing school randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis*, *35*(3), 370-390. doi:10.3102/0162373713482766

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. doi:10.1080/19345747.2015.1060282

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.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*(3), 337-346. doi:10.1002/sim.3782

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine*, *32*(19), 3373-3387. doi:10.1002/sim.5786

Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.

Nguyen, T., Ebnesajjad, C., Cole, S. R., & Stuart, E. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *Annals of Applied Statistics*, *11*(1), 225-247. doi:10.1214/16-AOAS1001

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, *63*(2), 195-210. doi:10.1111/rssc.12037

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. doi:10.1002/pam.21660

Pearl, J. (2011). Invited commentary: understanding bias amplification. *American Journal of Epidemiology*, *174*(11), 1223-1229. doi:10.1093/aje/kwr352

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *60*(1), 23-40. www.jstor.org/stable/2985969

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167-190. doi:10.1007/BF02295939

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol 1, 2nd ed). Sage.

Rosenbaum, P. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*(3), 207-224. doi:10.2307/1165073

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. doi:10.1093/biomet/70.1.41

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, *29*(1), 103-116. doi:10.3102/10769986029001103

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis* (pp. 118-12). Chapman and Hall/CRC.

Schoen, R. C., Lavenia, M., & Tazaz, A. (2017, March). *Effects of a Two-year Cognitively Guided Instruction Professional Development Program on First- and Second-Grade Student Achievement in Mathematics.* Paper presented at the spring Conference of the Society for Research in Educational Effectiveness, Washington, DC.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

StataCorp (2017). *Stata 15 Base Reference Manual*. Stata Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1-21. doi:10.1214/09-STS313

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *174*(2), 369-386. doi:10.1111/j.1467-985X.2010.00673.x

Stuart E. A., Bradshaw, C. P., Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*(3), 475-485. doi:10.1007/s11121-014-0513-z

Tipton, E. (2013a). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239-266. doi:10.3102/1076998612441947

Tipton, E. (2013b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, *37*(2), 109-139. doi:10.1177/0193841X13516324

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 481-501. doi:10.3102/1076998614558486

Tipton, E., Hedges, L., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness*, *7*(1), 114-135. doi:10.1080/19345747.2013.831154

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524. doi:10.3102/0013189X18781522

West, B. T., & Galecki, A. T. (2011). An overview of current software procedures for fitting linear mixed models. *The American Statistician*, *65*(4), 274-282. doi:10.1198/tas.2011.11077

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, *110*(511), 910-922. doi:10.1080/01621459.2015.1023805