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

Yujia Li
University of Connecticut - Storrs, yujia.li@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
https://opencommons.uconn.edu/dissertations/2325
Evaluating Methods for Handling Multilevel Selection for the Purpose of Generalizing Cluster Randomized Trials

Yujia Li, PhD
University of Connecticut, 2019

Over the past decade, the generalizability of randomized experiments, defined as the level of consistency between an estimated treatment effect in a sample and the true treatment effect in a target population, has received increasing attention from the educational 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. This study explores methods to adjust away the bias caused by both the between-institution and within-institution selection processes. For instance, in educational studies we desire to account for both nonrandom school-level selection and non-random selection of students and/or teachers. I conduct a simulation study to evaluate the bias reducing properties of different methods for estimating and utilizing inverse probability of participation (IPP) weights in this two-level context. The simulation study found that methods that incorporated both student and school IPP weights reduced more bias than methods that only incorporated the school IPP weights. Using data from a cluster randomized trial of a math professional development intervention, this study finds that within participating schools, the participating sample of students were more “advantaged” than the non-participating students, suggesting the need to adjust for within school nonrandom selection. The study estimated and applied IPP weights to reduce bias in the estimated population average treatment effect for the intervention.
Evaluating Methods for Handling Multilevel Selection for the Purpose of Generalizing
Cluster Randomized Trials

Yujia Li

B.A., Mount Holyoke College, 2011
Ed.M., Harvard University, 2013
M.S. University of Connecticut, 2018

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2019
Copyright by

Yujia Li

2019

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

Presented by
Yujia Li, B.A., M.S., Ed.M.

Major Advisor
Christopher Rhoads

Associate Advisor
H. Jane Rogers

Associate Advisor
Hariharan Swaminathan

Associate Advisor
Tania Huedo-Medina

Associate Advisor
Aarti Bellara

University of Connecticut
2019
ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my academic advisor, Professor Christopher Rhoads, for his guidance, encouragement and continuous support during my Ph.D. study. His incredible knowledge, enthusiasm and patience helped me in my research endeavor and writing of this dissertation. Despite his busy schedule, Chris was always quick to respond to my questions. Through numerous discussions, he helped me solidify ideas, clarify research questions, interpret findings and put research into context of practice. He was painstaking in helping me develop a clear and concise writing style. Without his support and nurturing, the completion of my Ph.D. study would not have been possible.

I would like to extend my sincere thanks to my dissertation proposal and advisory committee, Professors H. Jane Rogers, Hariharan Swaminathan, Tania Huedo-Medina, Aarti Bellara and Robert B. Olsen. I am grateful for their time, advice and insightful comments. Special thanks to Professor Robert Schoen for providing data for the case study, and Amanda Tazaz and Kristy Farina for preparing the data sets. Additionally, I would like to thank my graduate assistantship supervisors, Professors Megan Welsh, Bianca Montrosse-Moorhead, Betsy McCoach and Donald Leu, for the opportunities to work with them on interesting projects.

Finally, I am grateful to my parents and friends. My heartfelt appreciation goes to my parents, who told me to go as far as I dream and come home whenever my heart desires. Special thanks to my friends Shirely Li, Tanesia Beverly, Huihui Yu, Yifan Cao, Nathan Lally, Ziyun Deng, Tiffany Wong and Yian Xu for their friendship and company along the journey.
# Table of Contents

**Introduction**  
1

**Review of Literature**  
3  
Generalizability of RCT Based on School-Level Moderators  
3  
Applicability of Multilevel Propensity Score to Generalizability Research  
11  
Software Considerations  
18

**Research Questions and Hypotheses**  
19

**Simulation Study**  
21  
Simulation Design  
21  
Simulation Results  
34  
Summary of Simulation Results  
48

**Application of IPP Weights to an Example**  
51  
Participants and Target Population  
52  
Data Analysis Procedures  
53  
Results  
56  
Discussion of School and Student-Level Reweighting Applied to CGI Study  
63

**General Discussion**  
65

**Reference**  
69

**Appendix A**  
74  
R Codes for generating population  
74

**Appendix B**  
77  
STATA codes for sample selection and estimating population average treatment effects  
77

**Appendix C**  
81
Introduction

Generalization of results from an experiment done on a limited sample to a population of interest is an important issue for educational 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 was often difficult to gauge the applicability of the treatment effects from one randomized experiment to other populations.

In the past decade, the educational research community has devoted increasing attention to 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. 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 and/or teachers volunteer for the study. The importance of accounting for non-random 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. A review of school-based interventions found that among 93 studies that used active consent procedures, the average consent rate was 65.5%, ranging from 11-
100% (Blom-Hoffman et al., 2009). In a review of 37 randomized trials that compared recruitment strategies, the percentage of the \( n \) recruited for the trial that ultimately participated in the trial ranged from 2-80%, with the mean percentage being 32% (Cadwell et al., 2010). Evidence from large scale international assessments shows 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.

This study will extend the existing research on generalization from randomized experiments by illustrating possibilities and difficulties that may arise when considering generalization from non-random within school samples. First, I discuss different methods for estimating inverse probability of participation (IPP) weights when information on both non-participating schools and non-participating students within schools is available to the researchers. Next, I conduct a simulation study to evaluate the effectiveness of different methods of estimating student and/or school-level IPP weights for reducing bias from nonrandom selection of the study sample. Finally, I apply these methods to an IES-funded cluster randomized trial, *Replicating the Cognitively Guided Instruction Experiment in Diverse Environments* and discuss the difference between the unadjusted treatment effect estimate and the adjusted estimates (under various adjustment methods).
Review of Literature

I first provide an overview of existing methods for generalizing from experiments that focus exclusively on the school-level selection process. Then I review the literature on multilevel propensity score models and argue for its applicability to the problem of constructing IPP weights for the purpose of generalizing experimental results when both within school and between school covariate information is available for non-participating units.

Generalizability of RCT Based on School-Level Moderators

Existing work on improving the generalizability of RCTs has suggested two distinct approaches to deal with the problem. The first approach is prospective and focuses on coming up with a recruitment plan that will optimize the generalizability of the resulting RCT to a population average treatment effect (PATE) of interest (e.g., Tipton, 2013a, 2013b, 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; Nguyen, Ebnesajjad, Cole & Stuart, 2017; O’Muircheartaigh & Hedges, 2014; Stuart, Cole, Bradshaw & Leaf, 2011). Both approaches assume that treatment effects vary as a function of observable variables that characterize schools (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. Applications of these methods have taken advantage of publicly available school-level information from sources such as the Common Core of Data (CCD), the Stanford Education Data Archive (SEDA) and state-specific data sources (Tipton & Olsen, 2018).
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, 2013; Stuart et al., 2011). Additionally, assumptions that are required for causal inference in more general settings, i.e. the Stable Unit Treatment Value Assumption (SUTVA) and strongly ignorable treatment assignment in the focal study, must also be satisfied.

**Prospective recruiting approach.** When researchers recruit from the population of interest using a probability sampling scheme, e.g. simple random sampling, cluster sampling, or stratified sampling (Lohr, 2009), and apply appropriate statistical methods, they can obtain unbiased estimates of PATE (Hedges, 2013). In practice, it is difficult to conduct a randomized trial using units selected by a probability sampling scheme from a clearly defined population of interest. Most studies rely on volunteers, with the exception of federally mandated programs such as the National Assessment for Education Progress (which is not a randomized trial). Even after a principal agrees to allow researchers to conduct a randomized trial using his or her school, parents and/or teachers may not give consent. To alleviate the bias caused by nonrandom selection of schools, researchers have developed tools to help recruit the most representative sample possible.

**Stratified sampling.** Stratified random sampling is a widely used technique that protects researchers from obtaining a very unusual simple random sample, potentially reducing costs and often yielding more precise population estimates than simple random sampling (Lohr, 2009). This method divides the inference population into strata based on one or more classifiers. Within each stratum, units are randomly sampled. Proportional allocation or optimal allocation are often used to determine the number of units to select within each stratum. Proportional allocation is allocating the number of sampled units in each stratum
proportional to the size of the stratum. Optimal allocation is allocating more units to strata with larger variances.

Stratified sampling has been applied to the recruitment of RCTs with several adaptations (Tipton, 2013b; Tipton et al., 2014). There are many plausible school-level moderators to stratify on, such as school size, title I status, proportion of minority students and urbanicity of school location. As the number of variables increases, defining strata by discretizing continuous variables and cross-classifying becomes inefficient at best and impossible at worst. Tipton et al. (2014) proposed stratifying on propensity scores. The idea is that in an educational experiment, there is usually a target population of schools \( P \) and a population of eligible schools \( E \), where \( E \subset P \). In practice, eligibility criteria for schools to participate in a study are usually determined by power analysis, financial and practical concerns. In order to obtain a generalizable sample, researchers should sample from \( E \) in a way so that the sample of schools \( S \) is as closely representative of \( P \) as possible. To achieve the goal of representative sampling, a multivariate distance measure is necessary to quantify the degree of similarity between each school in the eligible set \( E \) and the target population \( P \). Tipton et al. (2014) calculate the eligibility propensity score: \( g(X) = \Pr (Q = 1|X) \), where \( 0 < g(X) < 1 \) by assumption, \( Q = 1 \) indicates that the school is eligible, and \( Q = 0 \) otherwise. \( X \) is a vector of all variables that explain variability of treatment effect in the population - the unconfounded sample selection assumption. The inference population is then divided into \( k \) strata based on the eligibility propensity score. Within each stratum, eligible schools are ranked on their distance from the stratum average of the inference population, with the ones with the smallest distance ranked on top. To recruit schools, research would start from the top of the rank, and go down the list. If a school declines to participate, researcher would go to the next school on the list. The number of schools to sample from each stratum is determined by proportional allocation. In additional to the unconfounded
sample selection assumption, this approach relies on additional assumptions that the eligible population is a subset of the inference population, and that X does not contain any covariate that define the set of eligibles E.

In the condition where the entire inference population is eligible for the study, \( E \cap P = P \), calculating the eligibility propensity score becomes impossible, because \( Q = 1 \) for the entire population. Tipton (2013b) proposed a different stratified sampling procedure based on cluster analysis. The k-means clustering method is used to classify the population of schools into \( k \) strata, and schools are sampled within each stratum. Researcher needs to specify a set of school-level covariates, \( X \), the number of clusters \( k \) and a distance metric. If \( X \) consists solely of continuous covariates, Euclidean distance is the most commonly used distance metric. If \( X \) includes dichotomous, categorical and continuous variables, or if any covariate contains missing values, other more general distance metrics are available, such as \( S_{ij} \) proposed by Gower (1971).

\[
S_{ij} = \sum_{t=1}^{v} S_{ijt} / \sum_{t=1}^{v} \delta_{ijt} \tag{1.1}
\]

\( S_{ij} \) measures the distance between two schools \( i \) and \( j \) on a set of covariates \( X_t \), \( t = 1, \ldots, v \). For categorical \( X_t \)'s, \( S_{ijt} = 1 \) if the two schools agree on \( X_t \), and \( S_{ijt} = 0 \) if they differ. For quantitative \( X_t \)'s, \( S_{ijt} = 1 - \frac{|x_i - x_j|}{R_t} \), where \( R_t \) is the range of \( X_t \). \( \delta_{ijt} = 1 \) when \( X_t \) can be compared, and 0 if the covariate value is missing for either school \( i \) or \( j \).

The elbow graph, which is similar to the scree plot commonly used in factor analysis, is a criterion for selecting the number of strata \( k \). Once the schools are classified into \( k \) strata, researchers can sample schools within each stratum randomly. They can also rank order the schools based on their distance from the center of the stratum, and select the most “stratum-typical” schools. Proportional allocation is suggested to decide how many schools to sample within each stratum.
**Retrospective adjustment approach.** These methods apply statistical adjustment with respect to school-level moderators to internally valid average treatment effects (ATE) in order to create an unbiased (or nearly unbiased) estimate of the average treatment effect in an external population. The retrospective adjustment approach can provide estimates of the PATE for multiple populations of interest, which is particularly useful when the population of interest is unknown in advance (Hedges, 2013). Project STAR, for example, was conducted by the Tennessee State Department of Education to study the effect of reducing class size on student achievement. The project randomly assigned students within 79 schools to small and large class sizes (Achilles et al., 2008). This study has been used to draw inferences about class size effects in states all over the U.S., despite the fact that only schools in Tennessee were in the study sample (Hedges, 2013). Retrospective adjustments can be useful for making inferences about the effect of class size interventions to schools in other states.

**Post-stratification based on the propensity score.** O’Muircheartaigh and Hedges (2014) estimated a PATE from an unrepresentative sample of schools by subclassifying schools and weighting the schools in each subclass by the proportion of schools with similar sampling propensity scores in the population. First, they compute the sampling propensity score $S(X)$ for each school in the sample and the population.

$$S(X) = P(Z = 1|X)$$  \hspace{1cm} (1.2)

$Z = 1$ if the school is in the study sample.

$Z = 0$ if the school is in the population but not in the sample.

$X$ contains all school-level variables that explain variability of treatment effect in the population and predict selection into the RCT sample.

The sampling propensity score is estimated using a logistic regression model.

$$\ln\left[\frac{S(X)}{1-S(X)}\right] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_m X_m$$  \hspace{1cm} (1.3)
Next, they categorized schools in the sample and in the population into $N$ strata based on their propensity scores. O’Muircheartaigh and Hedges (2014) recommended five strata, because Cochran (1968) found that dividing the distribution of $X$ into five strata removed 90% of bias. Third, they estimate the sample average treatment effects within each stratum. Lastly, they estimate the PATE as follows: 

$$\text{PATE} = \sum_{v=1}^{5} p(v)T(v),$$

where $p(v)$ is the proportion of the units in stratum $v$ in the population and $T(v)$ is the sample average treatment effect of stratum $v$. They argued that although the assumption of unconfounded sample selection is difficult to verify, subclassification on an incomplete set of variables still makes the inference less biased than no adjustment at all.

**Inverse probability of participation (IPP) weighting.** This method applies the inverse of the estimated probability of participation to create a synthetic “population” that mimics the target population (Stuart, Bradshaw & Leaf, 2015). It is similar to the subclassification based on propensity scores method, but instead of assigning the same weight to all observations within the same subclass, every school is assigned its own weight.

The first step is to estimate the sampling propensity score, or the probability of participation for each school. Typically, this is estimated via logistic regression, as in equation (1.3) (Cole & Stuart, 2011; Stuart, Cole, Bradshaw & Leaf, 2011; Stuart, Bradshaw & Leaf, 2015). However, researchers have also applied generalized boosted regression models and random forest 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 the linear model. After obtaining the predicted probabilities of participation for each school, sample observations are weighted by the inverse of their participation probabilities. All students within the same school receive the same weight. To estimate PATE, one can apply the IPP weights to the observations and obtain the mean difference between the weighted treatment and control groups (Cole & Stuart, 2011). A doubly-robust
estimator that applies IPP weights has also been utilized to estimate the PATE (Kern et al., 2016). Doubly robust estimators combine weighting to account for different participation probabilities with regression adjustment to account for the covariance between potential confounders and the outcome variable. In a different example of doubly robust estimation, Stuart, Bradshaw & Leaf (2015) ran a weighted multilevel regression model, adding the school IPP weights to the second level. The estimated coefficient value for the treatment assignment indicator was the PATE. The weakness of the inverse propensity score method is that when the overlap between the sample and the population is small, some observations receive large weights so that they represent a large portion of the population, which makes these observations highly influential. Correspondingly, poor overlap between the sample and the population means that the variance of the IPP weights will be large, which increases the standard error of the treatment effect estimate.

**Other retrospective adjustment methods.** Machine learning methods have been applied to the generalization of ATEs. Compared to parametric, linear models, these models have the advantage of avoiding linearity and additive assumptions. Kern et al. (2016) explored two approaches for using machine learning models. The first is to apply retrospective adjustment with IPP weights, and instead of estimating the IPP weights with the simple logistic regression propensity score model, they used a machine learning approach. They explored both random forest (RF) and generalized boosted regression (GBM) methods for estimating the weights. The second approach was to avoid propensity score based approaches entirely and instead to use Bayesian Additive Regression Trees (BART) to predict counterfactual outcomes for units in the target population, but not in the study. BART models were fit to the combined experimental and target population data, in order to generate predicted potential outcomes for the treatment and control conditions for each unit in the target population. The average of the difference between these potential outcomes is the
estimated average treatment effect. Kern et al. (2016) found that when the unconfounded sample selection assumption was satisfied, RF, GBM and BART outperformed parametric methods. When the assumption was violated, all estimators performed poorly.

Highly variable weights inflate the standard errors of the treatment effect estimate (see Little & Rubin, 2002). Zubizarreta (2015) developed a method to produce stable weights by simultaneously optimizing covariate balance between groups and minimizing the variance of the weights. It solves the convex optimization problem of minimizing the variance of weights under constraints of (1) the difference between weighted covariates of the sample and the covariates of the population must be smaller than a vector of constants $\delta_p$; (2) the sum of all weights equals 1, (3) all weights are positive. The first constraint can be used to balance statistics other than means, such as both the mean and the variance of covariates. This method explicitly solves for stable weights and thus address the problem caused by extreme weights. While this technique has not yet been applied to the generalization of RCTs, it seems to be a promising approach for reducing generalizability bias.

**Generalizability Index.** Tipton (2014) formulated a Generalizability Index to assess the similarity between a sample and a target population of schools. Researchers can compute this index prospectively or retrospectively to quantify how representative the sample (or a prospective sample) is of a particular population of interest.

The index is based on the Bhattacharyya coefficient – a measure of similarity between two groups. To calculate the generalizability index, the researcher should calculate the sampling propensity scores for schools in the sample and schools in the target population. The sampling propensity score, $S(x)$, is the predicted probability of being selected into a sample from the target population given covariates $X$ and it is estimated using formula (1.2) and (1.3). This procedure yields two distributions - $f_s(s)$ is the distribution of propensity scores for units in the sample, and $f_p(s)$ is the distribution of propensity scores of units not in
the sample but in the population. If the two distributions are continuous and known, the
Generalizability Index is defined by:

$$\beta = \int \sqrt{f_s(s)f_p(s)} ds$$  \hspace{1cm} (1.4)

If the distributions of the propensity scores do not follow any particular known
distributions, or if the density for some values of \(s\) is zero, researcher should convert \(f_s(s)\)
and \(f_p(s)\) into discrete distributions by dividing propensity scores into \(k\) bins. The discrete
version of the Bhattacharyya coefficient is:

$$\hat{\beta} = \sum_{i=1}^{k} \sqrt{w_{pi}(s)w_{si}(s)}$$  \hspace{1cm} (1.5)

where \(w_{pi}(s)\) is the proportion of the population that fall into a bin of propensity scores, and
\(w_{si}(s)\) is the proportion of the sample that fall into the same bin.

It can be easily shown that \(0 \leq \beta \leq 1\). \(\beta = 0\) when the two distributions have no
overlap and \(\beta = 1\) when the two distributions are identical. Through simulation studies,
Tipton (2014) developed rules of thumb for judging the magnitude of the index:

- Very high: \(0.9 \leq \hat{\beta} \leq 1\)
- High: \(0.8 \leq \hat{\beta} < 0.9\)
- Medium: \(0.5 \leq \hat{\beta} < 0.8\)
- Low: \(\hat{\beta} < 0.5\)

**Applicability of Multilevel Propensity Score to Generalizability Research**

The methods discussed above all assume that covariate information that can be used
to improve generalizability is only available at a single level of a population that potentially
has a multi-level structure. In educational applications to date only school-level covariate
information has been utilized. However, the application presented later in the dissertation
leverages information about non-participating students within participating schools, as well
as information about non-participating schools. In other words, covariate information is
available at two levels of a hierarchically structured population. For simplicity, I use “students” to refer to participating individuals within schools, which can also be teachers. This dissertation explores some methods for utilizing this additional within school information, and compares the reduction in bias from these methods to the bias reduction when only school-level information is available.

One method for utilizing within-school covariate information is to estimate propensity scores and form IPP weights, as described above. However, when covariate information is available at two levels, there are new choices to be made about how to utilize information to form IPP weights. This section summarizes the existing literature on multilevel propensity score methods. I use the term “multilevel propensity score methods” to refer to all approaches that use propensity score methods in a multilevel context, not necessarily propensity score models that explicitly utilize multilevel models.

The multilevel propensity score is an extension of the standard propensity score (Rosenbaum & Rubin, 1983; Stuart, 2010) to settings where individuals are clustered into higher level units and selection occurs at the individual level. An example from the causal inference literature is patients in many different hospitals self-selecting into a new type of treatment. The unadjusted difference in clinical outcomes between patients who receive the new treatment and patients who undergo the old treatment ignores both differences in the types of individuals who select treatment (e.g. gender, age, disease stage) and the types of clusters that select treatment (e.g. geographic location, median survival rate, proportion of enrollment). Failure to properly account for the different cluster and individual-level selection processes may result in incorrect conclusions. Correct specification of propensity scores can balance the treatment and control groups in both the individual and cluster-level covariates.
The strong ignorability assumption must hold for the multilevel propensity score to work (Kim & Seltzer, 2007). In settings where the objective is to remove bias due to self-selection into treatment, this means that treatment assignment and responses must be conditionally independent given a vector of observed covariates X, which must include all covariates that are both related to treatment assignment and potential outcomes. The analog in the generalization context is that X must contain all covariates related to selection into the study and that moderate the treatment effect (i.e. unconfounded sample selection).

Estimating multilevel propensity scores can take either one of two approaches, one propensity score model per cluster (Rosenbaum, 1986), or one multilevel generalized linear model (also named generalized linear mixed model) using all observations in all clusters (Kim & Seltzer, 2007). The first approach balances the treatment and control units within each cluster, thus balancing the entire sample of treatment and control units on both individual and cluster characteristics. This approach is feasible when cluster sizes are large enough and there are enough treatment and control individuals within each cluster to run propensity score models.

The second approach accounts for variability in the average probability of receiving treatment across institutions by using school-specific random intercepts. It accounts for the variability in the slopes relating student-level covariates to the probability of receiving treatment across institutions by using school-specific random slopes. The Empirical Bayes Estimator in the random effects model allows the clusters to “borrow strength” from each other, making the estimates more stable in the presence of small clusters. In addition, school-level characteristics can be included as predictors of the random intercepts and slopes, thus addressing the impact of cross-level interaction between student and school-level covariates on the probability of being selected into treatment. This approach does not require that each cluster has both treatment and control units. The clusters where all individuals are assigned to
the same condition can be included in the analysis. One disadvantage is that the random effects model does not guarantee balance within each cluster due to the shrinkage of random effects toward zero (Li, Zaslavsky & Landrum, 2013).

In multilevel observational studies, once propensity scores have been estimated, individuals can be matched or weights can be created in order to balance treatment and control units with respect to confounders in order to estimate an internally valid ATE. Matching can be done within the same cluster (Kim & Seltzer, 2007; Rosenbaum, 1986), across different clusters (Arpino & Mealli, 2011) or within homogeneous groups of clusters (Kim & Steiner, 2015). Alternatively, inverse propensity score weights (IPSW) can be used.

Li, Zaslavsky & Landrum (2013) showed that if the multilevel model for propensity score estimation is correctly specified, the internally valid average treatment effect (ATE) can be estimated with consistency by comparing appropriately weighted averages of the response variable. The weights $w_i$ are $w_i = \frac{1}{e(X)_i}$ for units in the treatment group, and $w_i = \frac{1}{1-e(X)_i}$ for units in the control group, where $e(X)_i$ is the propensity score for unit $i$. Li, et al. (2013) conducted a simulation study to evaluate several estimators: (1) “marginal” estimators which ignore clustering in both the propensity score model and the outcome model, (2) inverse propensity score weighting estimators that model the multilevel structure of the data in the propensity score stage but not in the outcome stage, (3) estimators that model the multilevel data structure in the outcome model but not in the propensity score stage, and (4) doubly-robust estimators that model the multilevel data structure at both stages. The results showed that estimators in (2), (3) and (4) which took into account of the clustered structure of the data in either the propensity score model or the outcome model greatly reduced bias in the ATE compared to methods that ignored the clustered data structure in both stages in (1). In addition, they found that using a random effects outcome model provided protection against misspecification due to a missing cluster-level confounder.
The multilevel propensity score is a natural choice for adjusting internally valid ATEs when there is both school- and student-level selection in to the study. In the within-school population, the “treatment condition” for the students is participation in the study. Instead of estimating a students’ probability to be assigned into the treatment condition, the multilevel propensity score models can estimate the students’ probability of participating in a study. If individual-level data is available on the entire target population, one can directly model an individual’s probability of selection into the study using the multilevel random effects model. In the case where individual-level data is only available in the schools that volunteered for the study, one can model the probability of school participation using existing methods (e.g. Cole & Stuart, 2013), and then the probability of student participation within volunteer schools using multilevel propensity score models (Kim & Seltzer, 2007; Li, Zaslavsky & Landrum, 2013; Rosenbaum, 1986).

This study assumes that the student-level information is only available in the schools that participate in the study, because it is more realistic than assuming information on the entire population of students is available. Therefore, selection into the study takes two stages - the selection of schools into the study and the selection of students into the study. To model the within school selection of students into the sample, one can run either single-level propensity score models for each school or to run one multilevel propensity score models pooling data from different schools. Running single-level propensity score models for each school is conceptually simple and when specified correctly, guarantees balance within clusters. It becomes infeasible, however, for small clusters and in clusters where there are not enough students who volunteer for the study.

In the case of running one multilevel propensity score model that includes all available student-level information, one needs to choose the functional form of the model. For multilevel generalized linear models, one needs to choose between fixed and random
effects for the clusters. The fixed effects model has the shortcoming of producing unstable propensity scores in the scenario of a large number of small clusters. The random effects model is more robust in the presence of small clusters as it allows borrowing of information across clusters. On the other hand, it does not guarantee balance within clusters because the Empirical Bayes Estimator shrinks the random level-1 coefficient toward the grand mean (Radenbush & Bryk, 2002). Kim & Seltzer (2007) considered fitting propensity score models with a random intercept, a random slope, or both a random intercept and a random slope. They recommended the random intercept and slope model because it accommodates variation in the unconditional selection probabilities (random intercepts) and variation in the impact of student characteristics on selection into treatment (random slopes) across clusters. This paper applies the random intercept and slope model for estimating the within school selection of students into the sample.

Among the methods for estimating an ATE, i.e. matching, weighting, subclassification and covariate adjustment (Rosenbaum and Rubin 1983; Schafer and Kang 2008; Steiner and Cook 2013), weighting or subclassification are most appropriate as applied to the generalization of RCTs. Matching is not appropriate because outcome data is typically only available for the participating sample. Subclassification is a coarsened weighting method where units in the same subclass are assigned the same weights (Stuart et al., 2011). Subclassification has the advantage of protection against extreme weights and thus ensures the estimator will not have a very large variance. On the other hand, it adds questions about how many subclasses one needs to assign at the school level and at the student level, and whether students in the same school will necessarily be in the same subclass. Using inverse probability of participation (IPP) weights (Cole & Stuart, 2010) results in an estimator that 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). These weights are defined
as the inverses of the estimated propensity scores $\hat{p}(x_i)$ of units in the sample. The school and student IPP weights can be applied to a weighted multilevel regression, which incorporates unequal selection probability for units at each level of sample selection (Pfeffermann et al., 1998). Stuart, Bradshaw & Leaf (2015) adopted this method to estimate PATE by adding school IPP weights to the second level of the multilevel model. A natural extension is to include both school and student IPP weights at each level in a multi-level context to obtain an estimate for PATE.

There exist two distinctions between applying the multilevel propensity score to multisite observational studies and estimating a generalizable average treatment effect in an RCT. In multisite observation studies, it is almost always the case that all clusters have both treated and control units, as the goal is to derive the causal effect of an intervention, and researchers are usually not interested in clusters where no one received the treatment. In large-scale educational experiments, on the other hand, it is almost always the case that a small number of schools participate in the study and the study sample is then used to make inferences about a much larger population of schools. Therefore, the majority of clusters in the population have zero students that participate in the study.

The second distinction is the assumptions. The *strong ignorability* assumption and the *overlap* assumption must hold when using multilevel propensity score methods for observational studies (Kim & Seltzer, 2007). When generalizing from RCTs to a population of interest, one needs to assume that the entire set of covariates that predict selection and moderate treatment effects are included in the analysis. This is generally a smaller set of covariates than what would be necessary to make the strongly ignorable assumption plausible in an observational study context (Tipton, 2014; Stuart et al., 2011). In the context of nonrandom selection at both the between school and the within school level, these set of
covariates can include school-level covariates, student-level covariates and/or cross-level interactions.

**Software Considerations**

In order to estimate propensity scores using a multilevel model, one must select an appropriate software. This section discusses software for the estimation of multilevel propensity scores and the application of these scores to adjust internally valid estimates in order to estimate a more externally valid PATE. In the first step of the analysis, researchers can either estimate the propensity of selection into the study within each school and balance the within school sample and the within school population, or estimate the propensity of selection into the study for all students in the participating schools, and balance the entire sample of consenting students with the overall student population in the participating schools. In the first approach, standard single level propensity score models can be applied and numerous software packages have been developed for this purpose (See Stuart, n.d. for a list of available software and packages).

For the second approach of estimating multilevel propensity scores, the software needs to both estimate generalized linear mixed models for binary outcomes and output predicted probabilities for each student. A variety of software satisfy these needs, e.g, R package *lme4* (1.1-21), STATA (15) commands *gllamm* and *melogit*, and SAS (9.2) command *PROC MIXED*. HLM (7) and Mplus (8) can estimate multilevel models with binary outcomes, however, they do not yet provide the option to output individual predicted probabilities for such models, and thus are inappropriate for this purpose. In the second step of the analysis, the researcher needs to incorporate the IPP weights into the outcome model to estimate PATE. The software needs to correctly handle sampling weights in each level of the multilevel model. West & Galecki (2010) recommended HLM (7), MLwinN (2.22), Mplus (7), STATA (12) commands *gllamm* and *xtmixed*.
In terms of software performance, a simulation study found that the STATA (14) command *mixed* and the R package *lme4* had lower convergence rates for multilevel random effects models when the at least one random slope has near-zero variance, compared to HLM (7), Mplus (7) and SAS (9.4) (McCoach et al., 2018). However, the study did not make performance comparisons among this software for multilevel models with binary or categorical outcomes. Considering the two possible STATA commands, the *melogit* command uses Gauss-Hermite quadrature and the *gllamm* command uses adaptive quadrature. The former tends to work well when the cluster sizes are small to moderate and intraclass correlations are low. The latter works well when the cluster sizes are small or large and regardless of high intraclass correlations (Rabe-Hesketh, Skrondal and Pickles, 2002).

In this study, I utilize STATA 15 SE (StataCorp, 2017b) because it has the capacity to handle all necessary computation in the estimation of PATE. It estimates the multilevel propensity scores with the *gllamm* and *melogit* commands, produces individual predicted probabilities with the *pred* command, and estimates PATE using weighted multilevel models with the *mixed* command.

**Research Questions and Hypotheses**

The questions that this study investigates are: (1) (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? (2) How do different methods for estimating IPP weights perform under different simulation scenarios? (3) How does adjusting for within and between school selection processes change estimated average treatment effects for the *Replicating the Cognitively Guided Instruction Experiment in Diverse Environments* Study?
The main study hypotheses are as follows. First, that PATE estimators that adjust for both between and within school selection will perform better than the estimator that only adjusts for between school selection when students are not randomly selected and when the within school participation rate is low. Second that, in the presence of an unmeasured school-level covariate that influences selection in to the study, the misspecified random effects model will perform almost as well as the correctly specified random effects model, because the random intercepts and slopes provide protection against misspecification (Li, et al., 2013).

To answer these research questions, I conducted a simulation study to compare existing methods that address school-level selection with new methods that address both within and between school selection process under a variety of conditions. Then I apply both the existing methods and new methods to the *Replicating the Cognitively Guided Instruction Experiment in Diverse Environments* Study and compare the resulting estimates for the PATE.
Simulation Study

Simulation Design

**Data generation.** Population data is generated in the R software (R Core team 2018). Codes are shown in Appendix A. Schools have the subscript \( h (h = 1, \ldots, H) \). Students within school \( h \) have the subscript \( k (k = 1, \ldots, n_h, \text{where } n_h \text{ is the total number of students in school } h, \text{or the within school population}) \). There are two school-level random covariates, \( V_{1,h} \sim N(0, 1) \) is continuous and \( V_{2,h} \sim Bernoulli(0.5) \) is binary. There is one continuous student-level covariate \( X_{hk} \sim N(V_{1,h}, 1) \), and the mean of \( X_{hk} \) within school \( h \) equals the school-level variable \( V_{1,h} \).

First, I generated two potential outcomes for each individual in the entire population. \( y_{hk}(0) \) is the response when student \( k \) in school \( h \) is assigned to control. It is generated from multilevel models predicted by school- and student-level covariates, and their interactions, as detailed in equation (2.1). \( y_{hk}(1) \) is the response when student \( k \) in school \( h \) is assigned to treatment. It is generated as the student’s potential outcome under the control condition, \( y_{hk}(0) \), plus a treatment effect, \( \phi_0 + \phi_1 X_{hk} \). \( \phi_0 \) and \( \phi_1 \) reflect the degree of impact by school-, student-level variables and their interactions on student treatment effects.

Specifically, \( \pi_{30} \) is the unconditional treatment effect for all students; \( \pi_{31} \) and \( \pi_{32} \) are the impacts of school-level variables \( V_{1,h} \) and \( V_{2,h} \) on student treatment effect; \( \pi_{40} \) is the impact of student-level variable \( X_{hk} \) on student treatment effect; \( \pi_{41} \) and \( \pi_{42} \) are the impacts of cross-level interactions on student treatment effect.
\[ y_{hk}(1) = w_0 + w_1 X_{hk} + \phi_0 + \phi_1 X_{hk} \quad (2.1) \]
\[ w_0 = \pi_{00} + \pi_{01} V_{1,h} + \pi_{02} V_{2,h} \]
\[ w_1 = \pi_{10} + \pi_{11} V_{1,h} + \pi_{12} V_{2,h} \]
\[ \phi_0 = \pi_{30} + \pi_{31} V_{1,h} + \pi_{32} V_{2,h} \]
\[ \phi_1 = \pi_{40} + \pi_{41} V_{1,h} + \pi_{42} V_{2,h} \]

Next, I used selection models at the school and student levels to determine who would “participate” in the study for a particular simulation run. Schools are selected for the study based on the result of a randomly generated Bernoulli variable, \( S_h, S_h \sim Bernoulli(p_h) \), \( S_h = 1 \) or 0, indicating whether school \( h \) is in the sample. School selection probability is determined by a logistic regression based on school-level covariates. This procedure assumes that schools self-select into studies. 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, I generated a treatment indicator \( Z_h \sim Bernoulli(0.5) \), which on average assigns 50% of schools into the treatment group (the simulation study assumes a school-randomized experiment).

\[ \ln \left( \frac{p_h}{1-p_h} \right) = \alpha_0 + \alpha_1 V_{1,h} \quad (2.2) \]

Selection of students within schools into the study sample is determined based on the outcome of the random variable \( S_h \sim Bernoulli(p_{hk}) \), where \( p_{hk} = P(S_{hk} = 1|S_h = 1) \). \( S_{hk} = 1 \) or 0, indicating whether student \( k \) in school \( h \) is in the sample. Student selection probability \( p_{hk} \) 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 a study can be determined by the values of coefficients in equation (2.3). In any one replication, the percentage of students that self-select into a study is random.

\[
\ln\left(\frac{\theta_{hk}}{1-\theta_{hk}}\right) = \eta_{0h} + \eta_{1h}X_{hk} \quad (2.3)
\]

\[
\eta_{0h} = \tau_{00} + \tau_{01}V_{1,h} + \tau_{02}V_{2,h}
\]

\[
\eta_{1h} = \tau_{10} + \tau_{11}V_{1,h} + \tau_{21}V_{2,h}
\]

Both \(V_{1,h}\) and \(V_{2,h}\) are predictors of the selection of students into the within school study sample in (2.3), but only \(V_{1,h}\) predicts the school selection in (2.2). The distinction is intentional because it is unlikely that the school-level variables that predict school and student selections are exactly the same.

**True SATE and Estimators of PATE.** All estimation was run in STATA 15 SE. Codes are shown in Appendix B. First, the true sample average treatment effect is calculated. Even though it is not an estimator, I compare it with the other estimators to show how much variation is due to within-study estimation error as compared with bias and variance caused by the consent and school selection processes. The true SATE is calculated by the average of \(y_{hk}(1) - y_{hk}(0)\) for the selected sample in 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 \(\gamma_{01}\) in the unweighted hierarchical linear model equation (2.4), which appears below. 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.

\[
y_{hk} = \beta_{0h} + \epsilon_{hk}, \epsilon_{hk} \sim N(0, \sigma^2) \quad (2.4)
\]

\[
\beta_{0h} = \gamma_{00} + \gamma_{01}Z_{h} + u_{0h}, u_{0h} \sim N(0, \tau)
\]

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 (2.4). The school
selection probability is modeled using a correctly specified single level propensity score model. School weights are computed as the inverse of the selection probability, \( \hat{\omega}_h = \frac{1}{\hat{p}_h} \), with \( \hat{p}_h \) estimated using the logistic regression model described in (2.2). This estimator was proposed in existing literature to retrospectively adjust away bias caused by nonrandom sampling of schools (Stuart, Bradshaw & Leaf, 2015). I estimate the IPP-School to compare it with the next three proposed estimators that incorporate both school- and student-level IPP weights.

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 (2.4). School IPP weights are again estimated by (2.2). Student probability of selection into the study is estimated using a separate single-level logistic regression model for each participating school \((S_h = 1) (2.5)\). This predicted selection probability is a conditional probability, \( p_{hk|l} = Pr(S_{hk} = 1|S_h = 1) \). These logistic models are estimated using the STATA command \textit{logit} and individual probabilities are predicted by the \textit{pred} command.

\[
\ln \left( \frac{p_{hk}}{1-p_{hk}} | S_h = 1 \right) = \eta_{0h} + \eta_{1h}X_{hk} \tag{2.5}
\]

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

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 (2.4). School IPP weights are again estimated by (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 (2.6). These multilevel logistic models are estimated using STATA commands *melogit* and *gllamm* (Rabe-Hesketh, Skrondal & Pickles, 2004; StataCorp, 2017a). Individual probabilities are predicted by the *pred* and *gllapred* commands, respectively.

\[
\ln\left( \frac{p_{hk}}{1-p_{hk}} \big| S_h = 1 \right) = \eta_{0h} + \eta_{1h}X_{hk} \quad (2.6)
\]

\[
\eta_{0h} = \tau_{00} + \tau_{01}V_{1,h} + \tau_{02}V_{2,h} + u_{0j}
\]

\[
\eta_{1h} = \tau_{10} + \tau_{11}V_{1,h} + \tau_{12}V_{2,h} + u_{1j}
\]

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 (2.4). School IPP weights are again estimated by (2.2). The student weights are estimated using equation (2.7). The model is missing a school-level covariate \( V_{2,h} \) from (2.6). I 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 \( \tau_{02}V_{2,h} \) and \( \tau_{02}V_{2,h}X_{hk} \) from (2.6) will add to the variation of the \( u_{0j} \) and \( u_{1j} \) terms in (2.7).

\[
\ln\left( \frac{p_{hk}}{1-p_{hk}} \big| S_h = 1 \right) = \eta_{0h} + \eta_{1h}X_{hk} \quad (2.7)
\]

\[
\eta_{0h} = \tau_{00} + \tau_{01}V_{1,h} + u_{0j}
\]

\[
\eta_{1h} = \tau_{10} + \tau_{11}V_{1,h} + u_{1j}
\]

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). For estimators involving weights, school weights are added to the second level and student weights are added to the first level. All weights are specified as sampling weights. The *mixed* procedure assumes that the level-2 weights \( w_j \) are the weight of the cluster and
level-1 weights are conditional weights \( w_{ij} \), or the inverse probability of individual \( i \) being selected into the sample given that group \( j \) is already selected into the sample. In the sample selection scenario of this study, the schools are the primary sampling units (PSUs) and students are the secondary sampling units (SSUs). Student selection is dependent upon the school being selected into the study. Therefore, this assumption is plausible. In the scenarios outside of the scope of this study, where students are PSUs and they are nested within schools, rescaling of the level-1 weights may be necessary.

**Simulation Conditions.** This simulation varies the population size, the within school participation rate, and the random/non-random selection process. The R and STATA codes for generating the populations, samples and estimating PATE are in Appendix A and B.

**Population size.** Two different populations are generated, corresponding to two different, potentially policy relevant, populations of interest. The first is a large population of \( H = 2000 \) schools. The school size, \( n_h \), follows a truncated normal distribution with mean of 200, standard deviation of 80, a minimum of 10 and maximum of 700 students. If the simulated cluster size is smaller or larger than the specified min or max, the number will be set equal to the min or max. This condition mimics the empirical distribution of public elementary schools in the State of Florida. The school size, \( n_h \), for each population of schools were chosen based on empirical distributions of the total number of grade 1 and grade 2 students in the State of Florida. The second population is a small population with \( H = 50 \) schools. This condition mimics the distribution of public elementary schools in a large school district in Florida. The school size \( n_h \) was also chosen based on the empirical distribution of total number of grade 1 and grade 2 students in those schools. The reason that I chose Florida schools and the within school populations is that these were relevant populations of interest in my case study, *Replicating the Cognitively Guided Instruction Experiment in Diverse Environments*. 
As described below, the simulations assume 12% of the population participates in the study. Therefore, I expect that all estimators will perform better in condition (1), because it has a larger population of clusters, and therefore a larger number of clusters in the study. For both conditions, I expect that the estimators that utilize random effects models to generate student IPP weights will perform better than those that utilize single level propensity score models, because the random effects model allows the small clusters to borrow information from large clusters which may stabilize the weights.

**Population treatment effect model.** In the population, treatment effect is moderated by the school-level covariates, student-level covariates and their interactions, as shown in equation (2.1). I specify 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. I set the value of $\pi_{30} = \pi_{31} = \pi_{32} = \pi_{40} = 1$ and $\pi_{41} = \pi_{42} = 0$ in (2.1), resulting in the following equation for the treatment effect .

Treatment effect $= Y_{hk}(1) - y_{hk}(0)$

$$= \pi_{30} + \pi_{31}V_{1,h} + \pi_{32}V_{2,h} + \pi_{40}X_{hk} + \pi_{41}V_{1,h}X_{hk} + \pi_{42}V_{2,h}X_{hk}$$

$$= 1 + V_{1,h} + V_{2,h} + X_{hk}$$

(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 characteristic. 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 $V_{1,h}$ and $V_{2,h}$. I set the value of $\pi_{30} = 0$ and $\pi_{31} = \pi_{32} = \pi_{40} = \pi_{41} = \pi_{42} = 1$ in (2.1), resulting in the following equation for the treatment effect.

Treatment effect $h_k = y_{hk}(1) - y_{hk}(0) = \phi_0 + \phi_1 X_{hk}$

$$= \pi_{30} + \pi_{31} V_{1,h} + \pi_{32} V_{2,h} + \pi_{40} X_{hk} + \pi_{41} V_{1,h} X_{hk} + \pi_{42} V_{2,h} X_{hk}$$

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

Holding other factors constant, the true PATE for the TE Main Effect condition and the TE Interaction condition are approximately the same. I expect that for the TE interaction condition, inclusion of student IPP weights reduce more bias than for the TE main effects condition.

**Sample selection processes.** To understand the impact of random or nonrandom selection to the generalization of RCTs in the multilevel context, I vary the randomness of the sample selection processes at the school and student levels. These conditions are achieved by varying the coefficient value of the school and student covariates in (2.2) and (2.3). The values of the intercepts are discussed in the next section regarding participation rates.

There are four sample selection conditions. First, the random school & student selection condition selects school and student randomly at both stages. The parameters in the selection models (2.2) and (2.3) are set to be zero except for the intercepts $\alpha_0$ and $\tau_{00}$. Second, the nonrandom school, random student condition sets $\alpha_1 = 1$ in (2.2). All parameters in (2.3) are set to be zero except the intercept $\tau_{00}$. Third, the nonrandom school, nonrandom student condition again sets $\alpha_1 = 1$ in (2.2). It also sets $\tau_{01} = \tau_{02} = 1, \tau_{10} = 2$ and $\tau_{11} = \tau_{12} = 0$ in (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 (2.2). It sets $\tau_{01} = \tau_{02} = \tau_{11} = \tau_{12} = 1, \tau_{10} = 2$ in
(2.3). Student selection into the within school sample are 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 he/she is in.

**School and within school participation rates.** Past studies found that participation rates vary vastly across schools in school-based interventions (Blom-Hoffman et al., 2009; Cadwell et al., 2010). I vary the within school participation rates to be 25% or 50% to represent small to medium within school participation rates. These conditions are designed to understand if within school participation rates interact with nonrandom within school selection to impact the generalizability of RCTs.

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 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 rate are set by varying the values of intercepts ($\alpha_0$ and $\tau_{00}$) in the selection models, equations (2.2) and (2.3). The specific values of these parameters are varied for the different simulations. As shown in Table 1 A-D, 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 $\alpha_0$ is selected so that in each simulation condition, approximately 12% of schools select into the sample across replications. The value of $\tau_{00}$ is 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%.

Even though the parameter values of the school selection model are selected so that on average, 12% schools will select into the sample over replications, for any replication,
there is a nonzero probability that zero school is selected into the sample. To solve this problem, in the large school population (H = 2000) conditions, the selection of schools is repeated until at least 10 schools are selected. In the small school population size (H = 50) conditions, the simulation procedure repeats the Bernoulli trials until exactly six schools are selected into the sample. In the treatment assignment step of the small population size condition, the program randomly assigns three schools to the treatment and three schools to the control condition.

**Summary of simulation conditions.** There are two population sizes and two population treatment effects models. At the sample selection stage, there are two participation rate levels and four sample selection processes. Therefore, there are a combination of $4 \times 8 = 32$ conditions to be studied. For each condition, the intercepts are calibrated to achieve the desired participation rate at each level. The population size and population treatment effect models will be referred to as *population conditions*. The sample selection processes and participation rates will be referred to as *sampling conditions*. A summary of the conditions is shown in Table 2.

**Evaluation criteria.** For each condition, 200 replications are simulated. For the methods that use multilevel logistic models, the percentage of non-convergent replications are calculated for the *melogit* and *gllamm* commands. The estimators are evaluated by the mean standardized bias and root standardized mean square error (RSMSE). The bias and root mean squared error are divided by the standard deviation of the real treatment effects in the population to make the magnitude of the bias and RMSE comparable across conditions.
Table 1.

Parameter values for (2.1), (2.2) & (2.3) in correspondence with simulation conditions.

A. Population size = 50, Population treatment effect model = TE Main Effect.

<table>
<thead>
<tr>
<th>H = 50 Parameters</th>
<th>Sample Selection Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random School &amp; Student</td>
</tr>
<tr>
<td>( \pi_{30} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{31} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{32} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{40} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{41} )</td>
<td>0</td>
</tr>
<tr>
<td>( \pi_{42} )</td>
<td>0</td>
</tr>
</tbody>
</table>

Population Treatment Effect Model

| \sigma_{0} | -2 | -2.5 | -2.5 | -2.5 |
| \sigma_{1} | 0  | 1    | 1    | 1    |
| \tau_{00}  | -1 | -1   | -6   | -9   |
| \tau_{01}  | 0  | 1    | 1    | 1    |
| \tau_{10}  | 2  | 2    | 2    | 2    |
| \tau_{11}  | 0  | 0    | 0    | 0    |
| \tau_{02}  | 0  | 1    | 1    | 1    |
| \tau_{12}  | 0  | 0    | 0    | 0    |

Participation Rates: School 12%, student 50%


<table>
<thead>
<tr>
<th>H = 50 Parameters</th>
<th>Sample Selection Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random School &amp; Student</td>
</tr>
<tr>
<td>( \pi_{30} )</td>
<td>0</td>
</tr>
<tr>
<td>( \pi_{31} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{32} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{40} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{41} )</td>
<td>1</td>
</tr>
<tr>
<td>( \pi_{42} )</td>
<td>1</td>
</tr>
</tbody>
</table>

Population Treatment Effect Model

| \sigma_{0} | -2 | -2.5 | -2.5 | -2.5 |
| \sigma_{1} | 0  | 1    | 1    | 1    |
| \tau_{00}  | -1 | -1   | -6   | -10  |
| \tau_{01}  | 0  | 1    | 1    | 1    |
| \tau_{10}  | 2  | 2    | 2    | 2    |
| \tau_{11}  | 0  | 0    | 0    | 0    |
| \tau_{02}  | 0  | 1    | 1    | 1    |
| \tau_{12}  | 0  | 0    | 0    | 0    |

Participation Rates: School 12%, student 50%

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Selection Process</th>
<th>Sample Selection Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random School &amp; Student</td>
<td>Nonrandom school, random student</td>
</tr>
<tr>
<td>$\pi_{30}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{31}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{32}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{40}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{41}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\pi_{42}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-2</td>
<td>-2.5</td>
</tr>
<tr>
<td>$\tau_{00}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{01}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{10}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{11}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{02}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{12}$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Participation Rates**
- School 12%, student 50%
- School 12%, student 25%


<table>
<thead>
<tr>
<th>Parameter</th>
<th>Sample Selection Process</th>
<th>Sample Selection Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random School &amp; Student</td>
<td>Nonrandom school, random student</td>
</tr>
<tr>
<td>$\pi_{30}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\pi_{31}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{32}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{40}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{41}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\pi_{42}$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>-2</td>
<td>-2.5</td>
</tr>
<tr>
<td>$\tau_{00}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{01}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{10}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{11}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{02}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tau_{12}$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Participation Rates**
- School 12%, student 50%
- School 12%, student 25%
Table 2.

**Simulation Conditions.**

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>H = 2000, ( n_h \sim N(200,80) ), ( \min = 10, \max = 700 )</td>
</tr>
<tr>
<td></td>
<td>H = 50, ( n_h \sim N(250,60) ), ( \min = 25, \max = 330 )</td>
</tr>
<tr>
<td>Population treatment effect distribution</td>
<td>TE Main Effect</td>
</tr>
<tr>
<td></td>
<td>TE Interaction</td>
</tr>
<tr>
<td>Sample selection process</td>
<td>Random school &amp; student</td>
</tr>
<tr>
<td></td>
<td>Nonrandom school, random student</td>
</tr>
<tr>
<td></td>
<td>Nonrandom school &amp; student</td>
</tr>
<tr>
<td></td>
<td>Nonrandom school, nonrandom student and interaction</td>
</tr>
<tr>
<td>Participation rate</td>
<td>School 12%, within school 50%</td>
</tr>
<tr>
<td></td>
<td>School 12%, within school 25%</td>
</tr>
</tbody>
</table>
Simulation Results

**Convergence rates.** The true SATE, IPP-School and IPPSSS had 100% convergence rates under all conditions. The proportion of convergent replications for the other two estimators, IPPSSM and IPPSSM (miss) are shown in Table 3. The IPPSSM and IPPSSM (miss) are two estimators that used student IPP weights estimated by the generalized multilevel random effects model, using all student-level observations in the participating schools. 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 misspecified multilevel propensity score model that omits one school-level covariate. The generalized multilevel random effects models were computed by both the `gllamm` command and `melogit` command in STATA 15. Results show that the `gllamm` had superior performance, with convergence rate close to 100% in all conditions. In comparison, the `melogit` had much lower convergence rates, ranging from zero to 88%.

The `melogit` had worse convergence rates when the population cluster size was small (H = 50) rather than large (H = 2000), and lower convergence rates when the within school participation rate was 25% rather than 50%. Both are because the former conditions yielded smaller sample sizes. In addition, the `melogit` command had higher convergence rate for the IPPSSM (miss) than the IPPSSM, because the former has a missing school-level covariate, thus one fewer cross-level interaction in the model.

Due to its clearly superior convergence rate, I present results estimated by the `gllamm` command for the IPPSSM and IPPSSM (miss) in the following sections.
Table 3.

Proportion of convergent replications by two STATA packages for two PATE estimators.

<table>
<thead>
<tr>
<th>Population Parameters</th>
<th>Within school participation rate</th>
<th>Estimator</th>
<th>Sample Selection Process</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Random School &amp; Student</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nonrandom School Random Student</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nonrandom School &amp; Student</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nonrandom School Nonrandom Student Interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>melogit</td>
<td>gllamm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gllamm</td>
<td>gllamm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>melogit</td>
<td>gllamm</td>
</tr>
<tr>
<td>H = 50 TE Main Effects</td>
<td>50%</td>
<td>IPPSSM</td>
<td>0.235 0.970 0.150 0.970 0.115 0.975 0.410 0.995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.320 1.000 0.195 1.000 0.880 0.995 0.690 1.000</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>IPPSSM</td>
<td>0.225 0.970 0.160 0.970 0.070 0.975 0.005 0.895</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.340 1.000 0.225 1.000 0.690 1.000 0.335 1.000</td>
</tr>
<tr>
<td>H = 50 TE Interaction</td>
<td>50%</td>
<td>IPPSSM</td>
<td>0.235 0.970 0.150 0.970 0.115 0.980 0.410 0.995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.320 1.000 0.195 1.000 0.880 1.000 0.690 1.000</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>IPPSSM</td>
<td>0.325 0.970 0.235 0.970 0.320 0.975 0.056 0.866</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.340 1.000 0.225 1.000 0.690 1.000 0.338 0.995</td>
</tr>
<tr>
<td>H = 2000 TE Main Effects</td>
<td>50%</td>
<td>IPPSSM</td>
<td>0.593 1.000 0.565 1.000 0.495 1.000 0.689 1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.635 1.000 0.585 1.000 1.000 1.000 1.000 1.000</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>IPPSSM</td>
<td>0.605 1.000 0.625 1.000 0.085 1.000 0.000 1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.625 1.000 0.625 1.000 1.000 1.000 1.000 1.000</td>
</tr>
<tr>
<td>H = 2000 TE Interaction</td>
<td>50%</td>
<td>IPPSSM</td>
<td>0.595 0.995 0.565 0.995 0.500 1.000 0.651 1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.635 0.995 0.585 0.995 1.000 1.000 1.000 1.000</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>IPPSSM</td>
<td>0.600 0.995 0.620 1.000 0.085 1.000 0.000 1.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IPPSSM (miss)</td>
<td>0.620 0.995 0.625 1.000 1.000 1.000 1.000 1.000</td>
</tr>
</tbody>
</table>

Note. 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.
**True sample average treatment effects.** First, I present the true sample average treatment effects (SATEs) under each condition (Table 4). The true SATE is the average of the real treatment effects of all students in the study sample, \( y_{hk}(1) \) minus \( y_{hk}(0) \). In practice, the true SATE is unobserved because one cannot simultaneously observe \( y_{hk}(1) \) and \( y_{hk}(0) \). The standardized bias and RSMSE of the true SATE are due to the study recruitment and consent processes alone, absent the additional random error induced by randomization.

Table 4.

**True sample average treatment effects.**

<table>
<thead>
<tr>
<th>Population Parameters</th>
<th>Within School Participation Rate</th>
<th>Random School Student</th>
<th>Nonrandom School Random Student</th>
<th>Nonrandom School Random Student Interaction</th>
<th>Nonrandom School Nonrandom Student Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Std. Bias</td>
<td>RSMSE</td>
<td>Std. Bias</td>
<td>RSMSE</td>
<td>Std. Bias</td>
</tr>
<tr>
<td>H = 50 TE Main Effects</td>
<td>50%</td>
<td>0.0009</td>
<td>0.361</td>
<td>0.765</td>
<td>0.820</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>0.0028</td>
<td>0.367</td>
<td>0.765</td>
<td>0.821</td>
</tr>
<tr>
<td>H = 50 TE Interaction</td>
<td>50%</td>
<td>-0.0236</td>
<td>0.334</td>
<td>0.751</td>
<td>0.830</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.0207</td>
<td>0.338</td>
<td>0.750</td>
<td>0.831</td>
</tr>
<tr>
<td>H = 2000 TE Main Effects</td>
<td>50%</td>
<td>-0.0026</td>
<td>0.055</td>
<td>0.735</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.0022</td>
<td>0.055</td>
<td>0.736</td>
<td>0.740</td>
</tr>
<tr>
<td>H = 2000 TE Interaction</td>
<td>50%</td>
<td>-0.0018</td>
<td>0.053</td>
<td>0.773</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.0019</td>
<td>0.053</td>
<td>0.771</td>
<td>0.775</td>
</tr>
</tbody>
</table>

*Note.* The table shows standardized bias and RSMSE of the true sample average treatment effects 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.

Several patterns emerge from the simulations. First, as the sample selection process becomes less random, the standardized bias and RSMSE of the SATE grow larger. Under the simulation design, the distribution of true treatment effects in the population are shown in (2.8) for the TE Main Effect conditions and (2.9) for the TE Interaction conditions.

\[
\text{Treatment effect}_{hk} = l + V_{1,h} + V_{2,h} + X_{hk} \quad (2.8)
\]
\[ \text{Treatment effect}_{hk} = V_{1,h} + V_{2,h} + X_{hk} + V_{1,h}X_{hk} + V_{2,h}X_{hk} \quad (2.9) \]

Random school & student selection produces SATEs with near-zero standardized bias and relatively small RSMSEs. Standardized bias and RSMSE increase monotonically across the following three conditions: (i) nonrandom school, random student selection, (ii) nonrandom school & student selection and (iv) nonrandom school, nonrandom student and interaction selection. This is because the variables that predict nonrandom selection in the sample also predict variation in the treatment effect. In the nonrandom selection conditions, the population parameters in the selection probability models (2.2 & 2.3) are positive, therefore, schools and students with larger values on the covariates \( V_{1,h}, V_{2,h} \) and \( X_{1,hk} \) are more likely to be selected into the study. In the population treatment effect model (2.1), \( V_{1,h}, V_{2,h} \) and \( X_{1,hk} \) positively predict treatment effects in the population. Therefore, study samples in the nonrandom selection conditions consist of individuals with higher treatment effects, producing larger standardized bias in the SATE. The less random the sampling process is, the more biased the SATE gets. In addition, these samples are more homogenous than the samples selected in the random selection conditions, producing smaller variance in the SATE. The increase in bias is faster than the reduction in variance, however, resulting in the increasing RSMSE.

Second, the SATE of 25% within school participation rate conditions have higher standardized bias and RSMSE than the SATE of 50% participation rate conditions, holding other conditions the same. This result shows that the smaller within school samples are more biased than the larger within school samples. This result is counterintuitive at first, as the participation rates of the within school samples are calibrated by changing the intercepts of the student selection probability models (2.3). The 25% participation rate conditions have smaller intercepts than the 50% participation rate conditions. The parameters of the student
and school variables remain the same. The resulting sample, therefore, should be equally biased regardless of sample size.

Figure 1.

*Logistics function: The impact of changing the intercept.*

Note. This figure illustrates the impact of changing intercept in the logistics function on the probability curves using 1000 data points. X follows $N(0, 1)$. $P_1$ is the predicted probabilities under the logistics function with intercept = 1 and slope = 2. $P_2$ is the predicted probabilities under the logistics function with intercept = -1 and slope = 2. The reduction in intercept causes a horizontal shift of the curve to the right. As a result, data points with X values around the mean have larger drops in probabilities than the data points with X values on the tails.

The reason for the larger bias in the smaller samples is that the population probability of selection models S-shaped logistics functions. Figure 1. is an example of two logistic functions. The y-axis is the selection probability predicted by the function, and the x-axis is an independent variable X. A reduction in the intercept causes the S-shape to move parallel to the right while the shape remains the same. Due to the S-shape of the function, probability reduction for individuals with X values in the middle is larger than individuals with X values on the end of the spectrum. In other words, when the within school participation rate drops
from 50% to 25%, students with moderate $X_{1,hk}$ values in schools with moderate $V_{1,h}$ and $V_{2,h}$ values experience drops in selection probabilities and become unlikely to be selected into the study. Students on the top of the spectrum experience little change in selection probabilities and are still highly likely to be selected. Students on the bottom of the spectrum remains unlikely to be selected. Therefore, the sample becomes more biased toward students on the top. Another way to interpret this is that as the probability curve shifts to the right, the students with larger $X$ values are more likely to be selected into the sample, which yields a more biased sample. Since $V_{1,h}, V_{2,h}$ and $X_{1,hk}$ also predict treatment effects in the population, the resulting SATEs are larger when the sample has students with larger values of $V_{1,h}, V_{2,h}$ and $X_{1,hk}$.

Third, when student selection is nonrandom, i.e. under nonrandom school & student conditions and nonrandom school, nonrandom student and interaction conditions, the SATE of TE Main Effect conditions have smaller standardized bias and RSMSE than the SATE of TE Interaction conditions. This can be attributed to the impact of interaction terms $X_{1,hk}V_{1,h}$ and $X_{1,hk}V_{2,h}$ on individual treatment effects in the TE Interaction condition. These extra terms make the student and school characteristics more influential on the individual treatment effects, and thus the same sampling process produces larger bias, variance and RMSE in the SATE for the TE Interaction condition than for the TE Main Effects condition. When within school selection is random (i.e., under random school & student condition and nonrandom school, random student condition), the difference in the distribution of treatment effects within schools do not matter. As long as the within school samples are random, the SATE of each school is unbiased.

Fourth, the patterns in the small school population and large school population conditions ($H = 50$ vs. $H = 2000$) are the same. This is expected because populations under these conditions were generated with the same population parameters except for the school
population size. The sample selection probability models were also the same. All differences in the values between the two conditions should be attributed to the random data generation process in R.
<table>
<thead>
<tr>
<th>Population Parameters</th>
<th>Estimators</th>
<th>Random School &amp; Student</th>
<th>Nonrandom School Random Student</th>
<th>Nonrandom School &amp; Student</th>
<th>Nonrandom School Nonrandom Student Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Std. Bias</td>
<td>RSMSE</td>
<td>Std. Bias</td>
<td>RSMSE</td>
</tr>
<tr>
<td>Within school participation rate</td>
<td>True SATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H = 50 TE Main Effects</td>
<td>Unadjusted ATE</td>
<td>0.0009</td>
<td>0.361</td>
<td>0.765</td>
<td>0.820</td>
</tr>
<tr>
<td></td>
<td>IPP School</td>
<td>-0.0176</td>
<td>0.779</td>
<td>0.549</td>
<td>1.294</td>
</tr>
<tr>
<td></td>
<td>IPPSS</td>
<td>-0.0185</td>
<td>0.778</td>
<td>0.548</td>
<td>1.292</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0217</td>
<td>0.785</td>
<td>0.538</td>
<td>1.295</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0176</td>
<td>0.779</td>
<td>0.549</td>
<td>1.294</td>
</tr>
<tr>
<td></td>
<td>H = 50 TE Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>True SATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unadjusted ATE</td>
<td>-0.0236</td>
<td>0.334</td>
<td>0.751</td>
<td>0.830</td>
</tr>
<tr>
<td></td>
<td>IPP School</td>
<td>-0.0179</td>
<td>0.750</td>
<td>0.832</td>
<td>1.189</td>
</tr>
<tr>
<td></td>
<td>IPPSS</td>
<td>-0.0926</td>
<td>0.719</td>
<td>0.434</td>
<td>1.156</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0943</td>
<td>0.718</td>
<td>0.433</td>
<td>1.152</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0918</td>
<td>0.723</td>
<td>0.426</td>
<td>1.160</td>
</tr>
<tr>
<td></td>
<td>H = 2000 TE Main Effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>True SATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unadjusted ATE</td>
<td>-0.0206</td>
<td>0.055</td>
<td>0.737</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>IPP School</td>
<td>-0.0040</td>
<td>0.133</td>
<td>0.010</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>IPPSS</td>
<td>-0.0039</td>
<td>0.133</td>
<td>0.010</td>
<td>0.207</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0039</td>
<td>0.133</td>
<td>0.010</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0054</td>
<td>0.722</td>
<td>0.437</td>
<td>1.159</td>
</tr>
<tr>
<td></td>
<td>H = 2000 TE Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>True SATE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unadjusted ATE</td>
<td>-0.0018</td>
<td>0.053</td>
<td>0.773</td>
<td>0.778</td>
</tr>
<tr>
<td></td>
<td>IPP School</td>
<td>-0.0046</td>
<td>0.117</td>
<td>0.725</td>
<td>0.759</td>
</tr>
<tr>
<td></td>
<td>IPPSS</td>
<td>-0.0095</td>
<td>0.117</td>
<td>0.013</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0096</td>
<td>0.117</td>
<td>0.013</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0091</td>
<td>0.117</td>
<td>0.013</td>
<td>0.143</td>
</tr>
<tr>
<td></td>
<td>IPPSSM (miss)</td>
<td>-0.0090</td>
<td>0.117</td>
<td>0.013</td>
<td>0.143</td>
</tr>
</tbody>
</table>

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. 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. IPPSSS refers to the IPP-School+Student-muti 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-muti 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.
Performance of estimators. The standardized bias and RSMSE of the estimators are shown in Table 5. These numbers are influenced by both nonrandom selections of students and schools in to the study and error induced by the randomization process within the study. The randomization error exists due to the random assignment of sampled schools into the treatment and control conditions and the estimation of the average treatment effect. The performance of the IPPSSSM and IPPSSSM (miss) are based on results from the `gllamm` command, due to its superior rate of convergent replicates compared to the `melogit` command.

Sample selection conditions. The average standardized bias and RSMSE of the true SATE and all PATE estimators by sample selection conditions are shown in Figure 2.

Figure 2.

Mean standardized bias and RSMSE for true SATE and PATE estimators by sample selection process.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Estimator</th>
<th>Random School &amp; Student</th>
<th>Nonrandom School Random Student</th>
<th>Nonrandom School Nonrandom School &amp; Student</th>
<th>Nonrandom School Nonrandom School Student Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Bias</td>
<td>True SATE</td>
<td>-0.006</td>
<td>0.706</td>
<td>1.829</td>
<td>1.886</td>
</tr>
<tr>
<td></td>
<td>Unadjusted ATE</td>
<td>0.006</td>
<td>0.700</td>
<td>1.777</td>
<td>1.383</td>
</tr>
<tr>
<td></td>
<td>IPP-School</td>
<td>-0.024</td>
<td>0.333</td>
<td>0.817</td>
<td>1.366</td>
</tr>
<tr>
<td></td>
<td>IPPSSS</td>
<td>-0.024</td>
<td>0.250</td>
<td>0.846</td>
<td>1.351</td>
</tr>
<tr>
<td></td>
<td>IPPSSSM</td>
<td>-0.022</td>
<td>0.248</td>
<td>0.844</td>
<td>0.739</td>
</tr>
<tr>
<td></td>
<td>IPPSSSM (miss)</td>
<td>-0.025</td>
<td>0.252</td>
<td>0.500</td>
<td>0.755</td>
</tr>
<tr>
<td>RSMSE</td>
<td>True SATE</td>
<td>0.202</td>
<td>0.791</td>
<td>1.850</td>
<td>1.891</td>
</tr>
<tr>
<td></td>
<td>Unadjusted ATE</td>
<td>0.451</td>
<td>0.931</td>
<td>1.347</td>
<td>1.617</td>
</tr>
<tr>
<td></td>
<td>IPP-School</td>
<td>0.437</td>
<td>0.998</td>
<td>1.316</td>
<td>1.421</td>
</tr>
<tr>
<td></td>
<td>IPPSSS</td>
<td>0.486</td>
<td>0.695</td>
<td>0.850</td>
<td>1.567</td>
</tr>
<tr>
<td></td>
<td>IPPSSSM</td>
<td>0.440</td>
<td>0.669</td>
<td>0.846</td>
<td>1.151</td>
</tr>
<tr>
<td></td>
<td>IPPSSSM (miss)</td>
<td>0.438</td>
<td>0.698</td>
<td>0.877</td>
<td>1.160</td>
</tr>
</tbody>
</table>

Note. The mean standard bias and RSMSE are averaged over population size, population treatment effect distribution and participation rate.
Random school & student. When the selection processes were random at both the school and the student level, all estimators performed similarly well. The standardized bias of all estimators was less than 0.1 standard deviations away from the true PATE when school population size was \( H = 50 \) and less than 0.01 standard deviations away from the true SATE when school population size was \( H = 2000 \). The RSMSEs of all estimators were less than one standard deviation away from the true PATE. All estimators had bigger RSMSEs than the SATEs because of the estimation error. The standardized bias and RSMSE of the unadjusted ATE and the IPP weighted estimators were similar. The unadjusted ATE had slightly smaller standardized bias than the IPP weighted estimators, but the differences were 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 processes were nonrandom at the school level and random at the student level, the IPP-School estimator had smaller standardized bias than the unadjusted ATE in both population sizes. The three IPP-School+Student estimators showed similar performance to the IPP-School estimator. The RSMSE of the IPP-School compared to the unadjusted ATE, on the other hand, show different patterns in the two population sizes. When the population size is large (\( H = 2000 \)), the RSMSE of the IPP-School is 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 adjust 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%, the sample only consists of 6 schools. The small number of clusters combined with the addition of sampling weights, likely increased the variance of the estimators and inflated the RSMSE. When the school population size is 2000 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, as a result most of the mean squared error is due to bias, and the reduction in bias that results from using the IPP-School estimator (or some version of an IPP-School+Student estimator) substantially reduces RSMSE compared to the unadjusted ATE.

Nonrandom school & student. When the selection processes were nonrandom at both the school and student level, and the selection probabilities are predicted by the main effects only, the standardized bias of the three IPP-School+Student estimators outperformed the IPP-School and the unadjusted ATE, because the IPP-School+Student estimators corrected for nonrandom selections at both levels. Amongst the three IPP-School+Student estimators, performance was similar. The IPPSSM (miss) always underperformed compared to the IPPSSM, due to the reason that the former is missing a school-level covariate in the model for estimating the student IPP weight. The difference in performance between the two estimators, however, was small. When school population size is 50, the difference in standardized bias between the two estimators was less than 0.02 standard deviations. When school population size is 2000, the performance of these two estimators was almost exactly the same. This result confirms our hypothesis that the misspecification of lacking a level-2 covariate in the model for estimating the student IPP weight is offset by the 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 had smaller standardized bias than the IPPSSS. The IPPSSM also had smaller RSMSE than the IPPSSS.
when the within school participation rate was 25%. When school population size is 2000, the IPPSSS had 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 was small. The difference in standardized bias and RSMSE between these two estimators was less than 0.04 standard deviations in all population size, population treatment effect and participation rate conditions.

*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 performed better than the IPP-school and the unadjusted ATE. The reason lies in the ability of the IPP-School+Student estimators to adjust away the bias caused by nonrandom selection at both school and student levels. The performance of the IPP-School compared to unadjusted ATE depended on the school population size and population treatment effect distribution. When school population size is 50, the IPP-School always had smaller standardized bias than the unadjusted ATE. The IPP-School estimators had larger RSMSE than the unadjusted ATEs under TE Main Effects. When the school population size is 2000, the IPP-School always performed better than the unadjusted ATE, having smaller standardized bias and RSMSE. Amongst the three estimators and adjusted for within school selection, the performances were similar. The IPPSSM (miss) always underperforms compared to the IPPSSSM, due to the obvious reason that the former is misspecified due to a missing school-level covariate in the model for estimating the IPP student weight. The relative performance of the IPPSSS and IPPSSM depended on the school population size. When the school population size is 50, the IPPSSM had 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 had lower RSMSE than IPPSSS when the population size is 50. The difference in the performance of these three estimators were small - within 0.1 standard deviations of the treatment effects in their respective populations. When school population size is 2000, the IPPSSS performed slightly better than the IPPSSM across population treatment effect models and within school participation rates, but the difference was less than 0.02 standard deviations of treatment effects in their respective populations.

Figure 3.

*Average standardized bias and RSMSE for PATE estimators by sample selection process and participation rate.*

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Within School Participation Rate</th>
<th>Random School &amp; Student</th>
<th>Nonrandom School Random Student</th>
<th>Nonrandom School Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Bias</td>
<td>50%</td>
<td>-0.020</td>
<td>0.352</td>
<td>0.590</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>-0.020</td>
<td>0.353</td>
<td>0.843</td>
</tr>
<tr>
<td>RSMSE</td>
<td>50%</td>
<td>0.440</td>
<td>0.751</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td>25%</td>
<td>0.440</td>
<td>0.746</td>
<td>1.162</td>
</tr>
</tbody>
</table>

*Note.* The mean standardized bias and RSMSE are averaged over all PATE estimators (excluding true SATE), population size and population treatment effect distribution.

*Within school participation rate.* The average standardized bias and RSMSE of all PATE estimators by within school participation rate are shown in Figure 3. When the sample selection processes were random at the student level, each estimator had similar standardized bias and RSMSE under the 50% and 25% within school participation rate. As long as within school selection is random, smaller within school participation rates have little impact on the bias of the estimators. Smaller within school samples may increase the variance of the estimators if the schools are small. In this simulation, the school sizes are relatively big - average school size is 250 students with standard deviation of 60.
When the sample selection processes were nonrandom at student level, each estimator performed better when the within school participation rate was 50% than when it was 25%. This is because the 25% within school participation rate selected not only fewer students per schools, but also more biased student samples (see explanation associated with Figure 1).

Figure 4.

**Average standardized bias and RSMSE for PATE estimators by sample selection process and population treatment effect distribution.**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>TE</th>
<th>Random School &amp; Student</th>
<th>Nonrandom School Random Student</th>
<th>Nonrandom School &amp; Student</th>
<th>Nonrandom School Nonrandom Student Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Std. Bias</td>
<td>Main Effect</td>
<td>-0.004</td>
<td>0.316</td>
<td>0.778</td>
<td>0.888</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>-0.005</td>
<td>0.159</td>
<td>0.556</td>
<td>0.990</td>
</tr>
<tr>
<td>RSMSE</td>
<td>Main Effect</td>
<td>0.459</td>
<td>0.189</td>
<td>1.164</td>
<td>1.324</td>
</tr>
<tr>
<td></td>
<td>Interaction</td>
<td>0.432</td>
<td>0.708</td>
<td>0.962</td>
<td>1.283</td>
</tr>
</tbody>
</table>

*Note.* The mean standardized bias and RSMSE are averaged over all PATE estimators (excluding true SATE), population size and participation rate.

**TE Main Effect vs. TE Interaction.** The average standardized bias and RSMSE of all PATE estimators by population treatment effect distribution is shown in Figure 4. Averaging across the other conditions, the TE Interaction conditions had 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 performed better in the TE Main Effects condition and the IPP weighted estimators generally performed 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 had 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 were 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 effect in the population is 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 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 had smaller standardized bias in the TE Main Effect than in the TE Interaction conditions. This anomaly of this condition may be associated with the fact that the selection probabilities in this condition were highly variable and the average participation rate were low, thus yielding schools with very small within school sample sizes. The combination of small school sample size (only six when the school population is 50), small within school sample size and high standardized bias in the SATE in the TE Interaction condition may have made it more difficult for the estimation of appropriate weights and thus their bias reduction properties become less effective.

**School population size.** The pattern of estimator performances in the $H = 2000$ and $H = 50$ conditions are for the most part the same, with a few aforementioned exceptions. All estimators performed better in the $H = 2000$ than in the $H = 50$, with smaller standardized bias and RSMSE. This is the result of larger sample sizes in the $H = 2000$ condition.

**Summary of Simulation Results**

This simulation study found that applying IPP weights that account for the level of nonrandom selection in the sampling process generally improved the performance of the estimators of PATE. When schools were nonrandomly selected, applying the school IPP weight reduced the standardized bias of the estimator compared to the unadjusted ATE. However, when the number of schools in the study is small (the $H = 50$ case in the
simulations) the increased variance due to reweighting means that no improvements in RSMSE is observed. The RSMSE of the reweighted estimators was smaller than that of the unadjusted ATE only when the number of schools in the experimental sample was large. When schools and students were both nonrandomly selected, applying both the school IPP weight and the student IPP weight improved the performance of the estimator relative to the IPP-School and the unadjusted ATE estimators. Therefore, applying IPP weights always reduced 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 may increase the RSMSE of the estimator through the inflation of estimator variance if the sample size is small.

The model for estimating the student IPP weights had little impact on the performance of these weights. The IPPPSSM performed better than the IPPSSS when the school population size was small and the reverse was true when the school population size was large. The IPPSSM (miss) showed slightly worse performance compared to the IPPSSM, but the difference was small. This supports the hypothesis 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 estimators also protects against missing school-level covariates because it does not need school level covariate in the model.

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

Under the TE Interaction conditions, the IPP weighted estimators generally performed better than under TE Main Effect conditions, and had 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 were more impacted by school and student characteristics than in the TE main effect conditions. Consequently, adjusting away bias caused by nonrandom selection was more effective in improving the accuracy of the estimator for PATE. The distribution of treatment effects in the population did not affect the order of the performance rankings among the estimators.
Application of IPP Weights to an Example

We utilize a unique dataset connected with the IES-funded experiment, *Replicating the Cognitively Guided Instruction Experiment in Diverse Environments* (Award #: R305A120781), to apply the methods tested in the simulation study. Cognitively Guided Instruction (CGI) is a widely adopted professional development program for elementary math education. Volunteer first and second grade math teachers from twenty-two schools in two school districts in the State of Florida participated in this study. Seven schools were from District A and 15 were from District B. Within each district, schools were randomized to the treatment and control conditions. Teachers in the treatment schools participated in the intervention program in 2013 and 2014. Teachers in the control condition participated in the business-as-usual professional development activities. The experiment collected measures on volunteer first and second grade teachers and their participating students over three years, including but not limited to variables on participant demographics, achievement scores on State Mathematics Assessments and achievement scores on researcher-designed mathematics assessments.

The CGI research team and District A provided de-identified student-level information on participating and non-participating students in the seven participating schools, providing a unique opportunity to explore the within school selection process and the impact of IPP weighting on the estimates of average treatment effect. This study focuses on generalizing the results of the experiment to estimate an ATE for all students and schools in District A.

We explore the following research questions. First, to what extent do the participating and non-participating first and second grade students differ with respect to observed pre-treatment variables? Second, how well do adjustment methods that account for both the within- and between-school selection processes balance observed covariates between
participating and non-participating students? Third, how do adjusted ATE estimates that use only school-level data compare with adjusted estimates that also utilize within-school information about non-participants?

**Participants and Target Population**

The study sample consists of 805 participating students of teachers who volunteered to participate in the CGI study in seven schools in District A. The intervention was carried out from Fall 2013 to Spring 2014. Twenty students had missing values on the school they attended, resulting in a study sample of 785 participating students. The non-participating sample consists 1,219 non-participating first and second grade students who attended these seven schools. A combination of these data results in 2,004 students whose individual-level demographic information and pretreatment State assessment scores are available. The within school participation rates among eligible first and second graders vary from 22% to 81%, and the aggregate participation rate is 60.83% (Table 6). The available student-level pretreatment variables are gender, grade level, free/reduced lunch status, English Language Learner status, disability status, gifted status and achievement scores on District Math Assessment at the first quarter of the academic year 2013-2014.

The target population consists of all first and second grade students in public schools in District A. I obtained pretreatment school-level information from the Florida Department of Education website (FLDOE) and the Common Core of Data (CCD) website. I collected school-level information for the 2012-2013 academic year to ensure school characteristics were unaffected by the CGI intervention. There were 75 public schools in District A according to the CCD. I excluded 37 schools that had no students in the target population or no information about school-level achievement scores (Figure 2). The resulting target population consists of 38 schools, seven of which have students and teachers who participated in the CGI intervention.
Figure 2.

Flowchart of school target school population determination.

Table 6.
Number and Percentages of Participants and Non-participants for Each School in the Sample.

<table>
<thead>
<tr>
<th>School #</th>
<th>Total</th>
<th>Participant</th>
<th>%</th>
<th>Non-participant</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>296</td>
<td>64</td>
<td>21.62</td>
<td>232</td>
<td>78.38</td>
</tr>
<tr>
<td>2</td>
<td>339</td>
<td>274</td>
<td>80.83</td>
<td>65</td>
<td>19.17</td>
</tr>
<tr>
<td>3</td>
<td>223</td>
<td>60</td>
<td>26.91</td>
<td>163</td>
<td>73.09</td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>99</td>
<td>41.25</td>
<td>141</td>
<td>58.75</td>
</tr>
<tr>
<td>5</td>
<td>297</td>
<td>86</td>
<td>28.96</td>
<td>211</td>
<td>71.04</td>
</tr>
<tr>
<td>6</td>
<td>292</td>
<td>92</td>
<td>31.51</td>
<td>200</td>
<td>68.49</td>
</tr>
<tr>
<td>7</td>
<td>317</td>
<td>110</td>
<td>34.70</td>
<td>207</td>
<td>65.30</td>
</tr>
<tr>
<td>Total</td>
<td>2,004</td>
<td>785</td>
<td>39.17</td>
<td>1,219</td>
<td>60.83</td>
</tr>
</tbody>
</table>

Data Analysis Procedures

First, I compute aggregate descriptive statistics of the participating and non-participating students from the seven participating schools. Effect sizes for the standardized difference between the participating and non-participating students for each measure are reported, along with the results of corresponding hypothesis tests. For dichotomous student demographic variables, i.e. male, grade level, free/reduced priced lunch, ELL, disability and
gifted status, I ran one multilevel logistic regression model per variable, with the
dichotomous student demographic variable as the outcome variable and student participation
status as the predictor, controlling for school membership. Reported effect sizes (and
corresponding p-values) are the coefficients of the predictor and should be interpreted as the
log odds ratio of participants to non-participants: \( \ln \left( \frac{\text{Odds}_{\text{participant}_i}}{\text{Odds}_{\text{participant}_j}} \right) \).
For the continuous variable, District Math Assessment, I ran one multilevel regression model
with achievement score as the outcome and participation rate as the predictor, controlling for
school membership. I report the Hedges’ g effect size and corresponding p-value.

Second, I seek to understand the within-school selection process. The IPP student
weights were estimated using the single-level propensity score model for each school, and the
multilevel propensity score model using all students in the seven schools. The effectiveness
of the weights in balancing the sample and the target population are evaluated by the absolute
standardized difference - the absolute value of the weighted sample mean subtracted from the
target population mean, divided by the standard deviation of the variable in the target
population. A commonly used cut off for sufficient covariate balance in the propensity score
literature is 0.2 (Stuart, 2010). First, student-level IPP weights were estimated separately
within each school. Student-level covariates were entered into the propensity score model,
and selection probability for each participating student was predicted. The inverse of the
selection probability was the student’s IPP weight. In one participating school, the IPP-
student weights (IPPS) could not be estimated because there was insufficient overlap between
the propensity scores of the participating and non-participating students. For the remaining
six schools, I evaluated the balance between the weighted within school sample and the
within school population within each school, and then together for all schools.

To estimate the second type of IPP student weights (IPPSM), one multilevel
propensity score model using the \textit{gllamm} package in STATA 15 was run for all students in
the seven schools with random intercepts for the schools. The random effects model has the ability to accommodate the lack of overlap between the participating and non-participating students within any particular school by pooling observations from all participating schools. The balance between the weighted student sample and the student population in all seven schools was evaluated. The IPPSM weights were chosen as the final IPP student weights due to their ability to utilize all student-level information.

Next, I seek to understand the between-school selection process. Observed school-level covariates for the seven participating schools and the 31 non-participating schools were compared using the independent sample t-test and the Chi-square test for independence. To estimate IPP school weights, one logistic regression was run with participating status as the dependent variable and all available school-level covariates as the independent variable. The school-level IPP weights were the inverse of the estimated selection probability for each school in the sample. To evaluate whether the weights are effective in balancing the sample and the target population, I calculated the absolute standardized difference between weighted school means and the population mean.

Finally, three sets of models were run to estimate the ATE. The dependent variables were student achievement scores on three math assessments in Spring 2014: the Mathematics Performance and Cognition Interview (MPAC), the Iowa Test for Basic Skills (ITBS) Math Problems test and the Iowa Test for Basic Skills Math Computation test. All achievement scores were vertically scaled for grade 1 and grade 2 students. The model in (2.4) was used to estimate the ATE. The independent variable was the indicator variable of school treatment status and all models included random intercepts for schools. The first estimator was the unadjusted internally valid ATE, estimated by a “standard” model that doesn’t incorporate adjustments to account for external validity bias. The second estimator is the IPP-School, estimated by incorporating the IPP school weights into the second level of the “standard”
model. This estimator was proposed in the existing literature. The third estimator is IPP-School+Student multi (IPPSSSM), estimated by incorporating both the IPP student weights and the IPP school weights into the first and second levels of the “standard” model. Reported effect sizes are the coefficient of the indicator variable of school treatment status divided by student-level variance, $\Delta = \frac{y_{01}}{\sqrt{\sigma^2}}$ (Tymms, 2004).

$$y_{hk} = \beta_{0h} + \varepsilon_{hk}, \varepsilon_{hk} \sim N(0, \sigma^2)$$ (2.4)

$$\beta_{0h} = \gamma_{00} + \gamma_{01}Z_h + u_{0h}, u_{0h} \sim N(0, \tau)$$

**Results**

Direct comparison between participating and non-participating students in all seven schools showed that participants were more “advantaged” students. There was no significant difference between the two groups on % male, grade level and % students with disabilities. The participants were less likely to be eligible for free/reduced lunch, to be English Language Learners, more likely to be gifted and had higher scores on the District assessment of Math Achievement (Table 7).

In one of the seven schools, all participating students were grade 2 students, making it impossible to balance the within school sample and the within school target population of all eligible first and second grade students using a single level propensity score model for this school. In each of the other six schools, the IPP student weight was estimated using a single level propensity score model per school and the covariate balance was checked after applying student weights. Results showed that the within school balance between the weighted student sample and the target within school population was sufficient - the weighted |SMD|s were below 0.2 for each covariate in all six schools (Appendix C). I evaluated the balance for all six schools by pooling all students together, applying the IPPSS weights, and checking balance. Table 8 presents the results after pooling all schools together, and shows that the
weighted $|SMD|$s were close to zero for all student-level covariates, indicating sufficient within school balance.

The second type of IPP student weights, IPPSM, were estimated as the inverse of student participation probability predicted by a multilevel propensity score model estimated using all seven schools (rather than separately within each school). The observations for building this model included all participating and non-participating students in seven schools. After applying these IPP student weights to participating students, the weighted $|SMD|$s were close to zero for all covariates, indicating sufficient within school balance. Table 9 presents the results when student-level IPP weights are estimated with a multilevel model, pooling across the seven schools.

Comparing the seven participating schools with the target population of 38 schools in District A, the participating schools had higher numbers of full time teachers and total students, higher percentage of students eligible for free/reduced price lunch, and a higher percentage of students who achieved “satisfactory” or higher on State reading, math, writing and science assessments. The differences, however, are not significant (Table 10). The absolute standardized mean difference between the participating schools and the target population ranged from 0.03 to 0.31. The largest $|SMD|$ was with respect to percent male in the student population, however, the difference was very small (52% for participating schools and 51% for the target population). After applying the IPP weights to participating schools, the $|SMD|$s on total number of full time teacher, total number of students, percent male students, student proficiency on reading and math reduced. On the other hand, the $|SMD|$ on percent minority students, percent free/reduced priced lunch, title I school status and student proficiency on writing increased. All $|SMD|$s were below 0.2 except for title I status (Table 11).
Table 7.
Comparisons of participants and non-participants on student-level covariates - Aggregates of students in seven participating schools.

<table>
<thead>
<tr>
<th></th>
<th>Participants (N = 805)</th>
<th>Non-participants (N = 1,219)</th>
<th>p-value</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>% male</td>
<td>50.65%</td>
<td>49.06%</td>
<td>0.901</td>
<td>0.01</td>
</tr>
<tr>
<td>Grade level (% Grade 1)</td>
<td>47.90%</td>
<td>51.11%</td>
<td>0.125</td>
<td>-0.15</td>
</tr>
<tr>
<td>% Free/reduced Lunch</td>
<td>41.45%</td>
<td>54.80%</td>
<td>0.001</td>
<td>-0.38</td>
</tr>
<tr>
<td>% ELL</td>
<td>6.87%</td>
<td>10.75%</td>
<td>0.014</td>
<td>-0.45</td>
</tr>
<tr>
<td>% Disability</td>
<td>11.92%</td>
<td>14.36%</td>
<td>0.066</td>
<td>-0.27</td>
</tr>
<tr>
<td>% Gifted</td>
<td>7.12%</td>
<td>3.77%</td>
<td>&lt;0.001</td>
<td>0.77</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1307.29</td>
<td>1281.938</td>
<td>&lt;0.001</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Note. For male, grade level, free/reduced priced lunch, ELL, disability and gifted status, p-values were obtained from multilevel logistics regression models with dichotomous student demographics variable as the outcome variable and student participation status as the predictor, controlling for school membership. Effect sizes are the coefficient of the predictor and should be interpreted as the ln($Odd_{participate} - Odd_{participate}$). For District Math Assessment, p-value was obtained from a multilevel regression model with achievement score as the outcome and participation rate as the predictor, controlling for school membership. Effect size is Hedge’s g.
Table 8.

**Results of inverse probability of participation weighting on participating students - Single level propensity score approach, averaging across six schools.**

|                          | Mean in Sample | Mean in Population | Mean in Weighted Sample | SD in Population | Unweighted | Weighted | | SMD | | SMD |
|--------------------------|----------------|--------------------|-------------------------|-----------------|------------|----------| |     | |     |
| % male                   | 0.50           | 0.50               | 0.50                    | 0.50            | 0.01       | 0.00     | |     | |     |
| Grade level (% grade 1)  | 0.56           | 0.51               | 0.51                    | 0.50            | 0.09       | 0.01     | |     | |     |
| % free/reduced lunch     | 0.43           | 0.55               | 0.55                    | 0.50            | 0.26       | 0.02     | |     | |     |
| % ELL status             | 0.06           | 0.10               | 0.10                    | 0.30            | 0.12       | 0.01     | |     | |     |
| % Disability             | 0.12           | 0.14               | 0.13                    | 0.35            | 0.07       | 0.02     | |     | |     |
| % Gifted                 | 0.05           | 0.04               | 0.04                    | 0.19            | 0.08       | 0.02     | |     | |     |
| District Math Assessment (SY 2013-2014 Quarter 1) | 1300.25        | 1288.56            | 1288.49                | 81.32           | 0.14       | 0.00     | |     | |     |

*Note.* This table shows the aggregate balance between the within school study sample and within school target population before and after weighting for six out of the seven participating schools. One participating school was excluded from this analysis due to the lack of overlap between participating and non-participating students in that school. Student weights were calculated for each participating student using formula \( \hat{w}_{hk} = 1/p(S_hk = 1) \). Probability of participation was estimated using logistics regression \( \ln \left( \frac{p(S_{hk} = 1)}{1-p(S_{hk} = 1)} \right) = X\beta + \epsilon_{hk} \), where \( S_{hk} = 1 \) when student is a participant. Independent variables Xs' are observed student-level covariates listed in this table. Six logistics regression models were run, one for each school. Within each school, population of interest is defined as all Grade 1 and Grade 2 students in this school, which includes participants and non-participants. [SMD] is absolute standardized mean difference, calculated by the absolute difference between (weighted) sample and population means divided by the standard deviation of population.
Table 9.

Results of inverse probability of participation weighting on participating students - Multilevel propensity score approach, averaging across seven schools.

<table>
<thead>
<tr>
<th></th>
<th>Mean in Sample</th>
<th>Mean in Population</th>
<th>Mean in Weighted Sample</th>
<th>SD in Population</th>
<th>Unweighted [SMD]</th>
<th>Weighted [SMD]</th>
</tr>
</thead>
<tbody>
<tr>
<td>% male</td>
<td>0.50</td>
<td>0.50</td>
<td>0.49</td>
<td>0.50</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Grade level (% grade 1)</td>
<td>0.48</td>
<td>0.50</td>
<td>0.51</td>
<td>0.50</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>% free/reduced lunch</td>
<td>0.39</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.21</td>
<td>0.00</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.06</td>
<td>0.09</td>
<td>0.10</td>
<td>0.29</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.12</td>
<td>0.13</td>
<td>0.14</td>
<td>0.34</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.08</td>
<td>0.05</td>
<td>0.06</td>
<td>0.22</td>
<td>0.11</td>
<td>0.03</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1307.66</td>
<td>1292.50</td>
<td>1289.73</td>
<td>80.94</td>
<td>0.19</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note. This table shows the aggregate balance between the within school study sample and within school target population before and after weighting for the seven participating schools. Student weights were calculated for each participating student using formula $w_{hk} = 1/\hat{p}(S_{hk} = 1)$. Probability of participation was estimated using one multilevel logistics regression that includes all students in seven participating schools, with one random intercept per school: $\ln\left(\frac{p(S_{hk} = 1)}{1-p(S_{hk} = 1)}\right) = X\beta + u_h + \epsilon_{hk}$, where $S_{hk} = 1$ when student is a participant. Independent variables $X$s’ are observed student-level covariates listed in this table. Within each school, population of interest is defined as all Grade 1 and Grade 2 students in this school, which includes participants and non-participants. [SMD] is absolute standardized mean difference, calculated by the absolute difference between (weighted) sample and population means divided by the standard deviation of population.
Table 10.

Comparisons of participating and non-participating schools in District A.

<table>
<thead>
<tr>
<th></th>
<th>Participating schools</th>
<th>Non-participating schools</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>7</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Total # Full Time Teachers</td>
<td>56.57</td>
<td>53.55</td>
<td>0.36</td>
</tr>
<tr>
<td>Total # students</td>
<td>781.43</td>
<td>737.19</td>
<td>0.39</td>
</tr>
<tr>
<td>% minority students</td>
<td>46.57</td>
<td>46.06</td>
<td>0.94</td>
</tr>
<tr>
<td>% Free/reduced priced lunch</td>
<td>48.57</td>
<td>50.77</td>
<td>0.82</td>
</tr>
<tr>
<td>% male</td>
<td>51.99</td>
<td>50.92</td>
<td>0.35</td>
</tr>
<tr>
<td>Title 1 school</td>
<td>4</td>
<td>17</td>
<td>0.91</td>
</tr>
<tr>
<td>% Students Achieve &quot;Satisfactory&quot; or higher on State assessments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>74.71</td>
<td>71.61</td>
<td>0.55</td>
</tr>
<tr>
<td>Math</td>
<td>72.43</td>
<td>70.13</td>
<td>0.68</td>
</tr>
<tr>
<td>Writing</td>
<td>58.29</td>
<td>56.48</td>
<td>0.72</td>
</tr>
<tr>
<td>Science</td>
<td>65.86</td>
<td>64.58</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Note. Chi-square test for independence was used to test hypotheses of for title 1. Independent sample t-tests were used to test hypotheses for the other variables.
Table 11.
*Results of Inverse Probability of Participation (IPP) Weighting on Participating Schools.*

<table>
<thead>
<tr>
<th></th>
<th>Mean in school sample</th>
<th>Mean in school population</th>
<th>Mean in weighted school sample</th>
<th>Target population Std.</th>
<th>Unweighted</th>
<th>Weighted</th>
<th>[SMD]</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>7</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total # Full Time Teachers</td>
<td>56.57</td>
<td>54.11</td>
<td>55.81</td>
<td>11.17</td>
<td>0.22</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Total # students</td>
<td>781.43</td>
<td>745.34</td>
<td>772.23</td>
<td>154.52</td>
<td>0.23</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>% minority students</td>
<td>46.57</td>
<td>46.16</td>
<td>48.58</td>
<td>15.72</td>
<td>0.03</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>% Free/reduced priced lunch</td>
<td>48.57</td>
<td>50.37</td>
<td>52.51</td>
<td>22.48</td>
<td>0.08</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>51.99</td>
<td>51.12</td>
<td>51.04</td>
<td>2.83</td>
<td>0.31</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Title 1 School</td>
<td>0.57</td>
<td>0.55</td>
<td>0.72</td>
<td>0.50</td>
<td>0.04</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>% Students Achieve “Satisfactory” or higher on State assessments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>74.71</td>
<td>72.18</td>
<td>72.64</td>
<td>13.34</td>
<td>0.19</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>72.43</td>
<td>70.55</td>
<td>70.08</td>
<td>12.08</td>
<td>0.16</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>58.29</td>
<td>56.82</td>
<td>58.92</td>
<td>13.94</td>
<td>0.11</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>65.86</td>
<td>64.82</td>
<td>63.77</td>
<td>14.42</td>
<td>0.07</td>
<td>0.07</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Sample is defined as 7 schools in district A that participated in CGI. Population of interest is defined as 38 public elementary schools in District A. School weights were the inverse probability of participation weights calculated using the formula \( \hat{w}_h = 1/p(S_h = 1) \). Probability of participation was estimated using logistic regression \( \ln \left( \frac{p(S_h = 1)}{1 - p(S_h = 1)} \right) = X\beta + \epsilon_h \), where \( S_h = 1 \) when school is participating. Independent variables Xs’ are school-level covariates listed in this table. |SMD| is the absolute standardized mean difference, calculated by the absolute difference between (weighted) sample and population means divided by the standard deviation of target population.
Lastly, the unadjusted internally valid ATE, the IPP-School and IPP-School+Student multi (IPPSSM) were estimated for three student outcome variables. Table 12 shows that the magnitude of estimates for the ATEs differed depending on the application of weights, but the significance levels did not. For MPAC, the IPP-school adjusted estimators were larger in magnitude than the unadjusted ATE and statistically significant. The IPPSSM estimators, however, were smaller than the unadjusted ATE and not significant. For the two ITBS tests, adding school and student-level IPP weights both led to increases in the magnitude of the estimate compared to the unadjusted ATE. The significance level did not change. The standard errors of the IPP-School and IPPSSM were always larger than that of the unadjusted ATE. One plausible reason for the increase in magnitude of the estimated population ATE is that the treatment effect is higher for disadvantaged schools and students, who are underrepresented in the sample. Applying IPP weights increased the weights for these schools and students. The lack of change in significance level of the estimates can be attributed to the increased in standard errors.

**Discussion of School and Student-Level Reweighting Applied to CGI Study**

Through empirical investigation of a unique data set related to the *Replicating the Cognitively Guided Instruction Experiment in Diverse Environments* study, I found that the within school student samples were more advantaged, higher performing than students not in the sample. This result is consistent from findings in large scale international assessments, where less capable students are more likely to be absent from assessments (Rust, 2013). It shows that the within school sample was not necessarily representative of the within school population, so external validity adjustments based only on school-level characteristics may still be biased. Adding student-level IPP weights in addition to school-level weights changed the magnitude of the estimated population ATE, but not the significance.
Table 12.

*Estimated Average Treatment Effects on Student Mathematics using “Standard” Two-level model and with School-level and Student-level Inverse Probability of Participation weights.*

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Unadjusted Internally Valid ATE</th>
<th></th>
<th></th>
<th></th>
<th>IPP-School</th>
<th></th>
<th></th>
<th></th>
<th>IPPSSM</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MPAC Interview</td>
<td>0.375 0.30 0.16 0.06 0.455 0.35 0.17 0.04 0.371 0.29 0.20 0.139</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITBS Math Problems</td>
<td>0.190 3.93 5.76 0.50 0.212 4.90 7.25 0.54 0.243 5.11 7.41 0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ITBS Math Computation</td>
<td>0.074 1.21 3.69 0.74 0.120 1.89 4.00 0.64 0.142 2.11 4.19 0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* For each outcome, three separate models were fit. All models were fit using STATA 15 mixed function with random intercepts for school clusters and an indicator variable for treatment condition as the predictor. For weighted models, school weights were added to level 2 and student weights were added to level 1. All weights were treated as sampling weights.
General Discussion

This dissertation explored methodological approaches of handling multilevel selection of samples into randomized controlled trials for the purpose of generalizing treatment effect estimates to a target population. The simulation study showed that when the within school sample is not randomly selected and the unconfounded sample selection assumption holds, ignoring the within school selection process leads to bias in the estimated population average treatment effect. When the assumption holds, the two estimators that involve student IPP weights (IPPSSS and IPPSSM), applied in addition to the school IPP weights, significantly reduce bias in the estimated population average treatment effect compared to applying the school IPP weights alone. In addition, these estimators protect against 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, as shown by the similarity in performance between IPPSSM and IPPSSM (miss).

The simulation study also showed that small sample size creates 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 = 2000, school sample size ≈120) clearly showed large reduction in both standardized bias and RSMSE, the small school population condition (H = 50, school sample size = 6) had reduction in standardized bias, and much smaller reduction in RSMSE. In one particular condition, the RSMSE of the IPP-School was larger than the unadjusted RSMSE, suggesting that not applying retrospective adjustment would be the best choice for the specific condition. The sample size needed for the IPP weights to effectively reduce both bias and RMSE is a topic for future research.
Analysis of data from the CGI study showed that the within school samples were systematically different from the within school populations, with the participating within school students showing “advantage” over non-participants. The application of student and school IPP weights showed some difference between the re-weighted and original ATE estimates. To the author’s knowledge, this is the first study that shows empirical evidence of within-school selection bias in randomized controlled trials.

This study has several implications for future directions. First, there needs to be more studies on with within school information about non-participants to know if a general pattern will emerge. These studies can help researchers understand the extent to which certain within school populations are overrepresented or underrepresented and which variables are associated with nonrandom selection. If the variables that are associated with nonrandom within school selection are also associated with variation in treatment effects, researchers need to pay particular attention to selecting a student sample that is representative with respect to these variables and/or adjust their effects away in data analysis. Second, given the difficulty of random selection of samples, methods for purposive within school sampling, similar to those for purposive school sampling (e.g. Tipton, 2013; Tipton et al., 2014), may be needed to select samples that are as representative as possible in order to produce generalizable results. The Generalizability Index (Tipton, 2014) may need to be revised to reflect the representativeness of both the between school and within school samples.

Third, novel propensity score methods that yield propensity scores with more favorable properties may be applied to retrospectively adjust away biasing effects of moderating variables. The result of the simulation study showed that, while IPP school and student weights substantially decreased standardized bias, they had less of an effect on the RSMSE, indicating a trade-off between bias and variance. This shows the possible effectiveness of the stable weights developed by Zubizarreta (2015), which should, in theory,
optimally balance the bias-variance trade-off. In the case of limited coverage between a sample and a target population, stable weights may reduce extreme weights and the inflation of standard errors for the estimate. In addition, recent research on propensity score methods show that machine learning methods outperform logistic regression models in terms of bias reduction and mean squared error under conditions of nonlinearity and non-additivity (Lee, Lessler & Stuart, 2010). By extension, such propensity score models using machine learning methods can also be applied to the generalizability setting (e.g. Kern et al., 2016). These methods, however, have to be adapted for the multilevel setting.

This study has several limitations. First, the methods explored in this study rely on the strongly ignorable sample selection assumption, which cannot be verified empirically. Nguyen, Ebnesajjad, Cole & Stuart (2016) did a sensitivity analysis for an unobserved school-level moderator, and future research should investigate the impact an unobserved student-level moderator. Second, 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 be linear or nonlinear predictors, and more 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). Third, all existing methods and the methods proposed here assume that the selection probability and the corresponding IPP weight are fixed. However, the weights themselves are random variables estimated from the data and carry uncertain in themselves. 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. 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.
Reference


http://doi.org/10.2307/2335942


StataCorp (2017b). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.


Appendix A

R Codes for generating population

## These codes generates population data - potential outcomes for treatment and control conditions for all students in the population, sample selection probabilities for schools, sample selection probabilities for students ##

## The simulation condition is: population size H = 2000, TE interaction, participation rates: school = 12%, student = 25% ##

## To generate other conditions, change parameter values according to Table 1 A-D in the main text ##

setwd("path to save generated datasets")

library(foreign)

## Setting parameter values ##

H=2000 # Number of schools in the population
pi30_l=rep(0,4) # intercept of treatment effect (TE) model
pi31_l=rep(1,4) # impact of v1 on TE
pi32_l=rep(1,4) # impact of v2 on TE
pi40_l=rep(1,4) # impact of x1 on TE
pi41_l=rep(1,4) # impact of x1v1 on TE
pi42_l=rep(1,4) # impact of x1v2 on TE

alpha0_l=c(-2,-2.5,-2.5,-2.5) # intercept of school selection prob
alpha1_l=c(0,1,1,1) # impact of v1 on school selection prob

tau00_l=c(-1,-1,-5.5,-8) # intercept of student selection prob
tau01_l=c(0,0,1,1) # impact of v1 on student selection
tau10_l=c(0,0,2,2) # impact of v2 on student selection
tau11_l=c(0,0,0,1) # impact of x1 on student selection
tau02_l=c(0,0,1,1) # impact of x1v1 on student selection
tau12_l=c(0,0,0,1) # impact of x1v2 on student selection

### Start of simulation ####

for (t in 1:4){
  seed=123456+t
  set.seed(seed)

  # Predictors of treatment effects in the population
  pi30=pi30_l[t]
  pi31=pi31_l[t]
  pi32=pi32_l[t]
  pi40=pi40_l[t]
  pi41=pi41_l[t]
  pi42=pi42_l[t]

  # School selection probability parameters
  alpha0=alpha0_l[t]
  alpha1=alpha1_l[t]

  # Student selection probability
\(\tau_{00}=\tau_{00\_l[t]}\)
\(\tau_{01}=\tau_{01\_l[t]}\)
\(\tau_{10}=\tau_{10\_l[t]}\)
\(\tau_{11}=\tau_{11\_l[t]}\)
\(\tau_{02}=\tau_{02\_l[t]}\)
\(\tau_{12}=\tau_{12\_l[t]}\)

# Setting up intermediate parameters

\[K=\text{rep}(\text{NA},H)\]  # number of student per school
\[v_1=\text{rep}(\text{NA},H)\]  # school characteristic
\[v_2=\text{rep}(\text{NA},H)\]  # school characteristic
\[x_1=\text{NULL}\]  # student characteristic
\[y=\text{NULL}\]  # student outcome in the population
\[y_0=\text{NULL}\]  # potential outcome for control units
\[y_1=\text{NULL}\]  # potential outcome for treated units
\[TE=\text{NULL}\]  # true treatment effect of every unit in the population
\[TE_S=\text{NULL}\]  # treatment effect of units in the sample
\[PATE=\text{NULL}\]  # true ATE of the population
\[SATE=\text{NULL}\]  # true SATE
\[\phi_0=\text{NULL}\]  # intermediate coefficient
\[\phi_1=\text{NULL}\]  # intermediate coefficient
\[\eta_0=\text{NULL}\]  # intermediate coefficient
\[\eta_1=\text{NULL}\]  # intermediate coefficient
\[p_2=\text{NULL}\]  # school selection probability
\[p_1=\text{NULL}\]  # student selection probability

### Start of the simulation ###

for (h in 1:H){

# Step 1. Generate population of schools, and school and student characteristics

\[K[h]=\text{as.integer}(\text{rnorm}(1,200,80))\]  # Generates K students per school
if(K[h]<10) {
    K[h]=10
} if(K[h]>700) {
    K[h]=700
}
\[v_1[h]=\text{rnorm}(1,0,1)\]  # v1~N(0,1)
\[v_2[h]=\text{rbinom}(1,1,0.5)\]  # v2~Bernoulli (0.5)
\[x_1[h]=\text{rep}(\text{NA},K[h])\]  # student level variable x1
\[x_1[h]=\text{rnorm}(K[h],v_1[h],1)\]  # x1~N(v1,1), so v1 is the mean of x1

# Step 2. Simulate potential outcomes for all students in all schools

\[y_0[h]=\text{rep}(\text{NA},K[h])\]
\[y_1[h]=\text{rep}(\text{NA},K[h])\]
\[TE[h]=\text{rep}(\text{NA},K[h])\]
\[\phi_0[h]=\pi_30+\pi_31*v_1[h]+\pi_32*v_2[h]\]  # level 2 intercept as outcome of v1
\[\phi_1[h]=\pi_40+\pi_41*v_1[h]+\pi_42*v_2[h]\]  # level 2 slope as outcome of v1

for (k in 1:K[h]){  # for each student
    \[y_0[h][k]=v_1[h]+x_1[h][k]*(v_1[h])\]  # potential outcome for treatment
    \[y_1[h][k]=y_0[h][k]+\phi_0[h]+\phi_1[h]*x_1[h][k]\]  # potential outcome for control
    \[TE[h][k]=\phi_0[h]+\phi_1[h]*x_1[h][k]\]  # true individual treatment effect
}
# Step 3. Generate selection probabilities

# Generate school selection probability

\[ p_2[h] = \frac{1}{1 + \exp(-\alpha_0 + \alpha_1 v_1[h])} \]

# Generate student selection probability

\[ p_1[h][k] = \frac{1}{1 + \exp(-\eta_0[h] + \eta_1[h] x_1[h][k])} \]

# Step 4. Compile data and write out to .dta files

schid=rep(seq(1:H),K)  # School IDs
studentid=sequence(K)  # Student IDs

# flatten all "lists" of variables to make them single variables
y1_l=unlist(y1)
y0_l=unlist(y0)
TE_l=unlist(TE)
x1_l=unlist(x1)
p1_l=unlist(p1)
v1_l=rep(v1,K)
v2_l=rep(v2,K)
p2_l=rep(p2,K)
size=unlist(K)

# Write out the data sets
dataset_5i=data.frame(cbind(schid,size,studentid,y1_l,y0_l,TE_l,v1_l,v2_l,x1_l,p2_l,p1_l))
filename=paste0("Data_5M_",t,".dta")
write.dta(dataset_5i,filename)
Appendix B

STATA codes for sample selection and estimating population average treatment effects

** Codes in this file conducts analysis using the data generated from R **
** Analysis are saved in the excel spreadsheet in the same folder as the datasets **
** These codes apply to all conditions with H = 2000 **
** Codes for conditions with H = 50 have small variations; see notes in the codes below **
** High performance cluster is recommended due to the long computation time **

quietly capture cd "path of datasets to be analyzed"

ssc install gllamm

forvalues t=1(1)4{
putexcel set myresults_gl`t'

putexcel A1="True PATE"
putexcel B1="True SATE"
putexcel C1="Internally Valid ATE"
putexcel D1="IPP-school"
putexcel E1="IPPSSS"
putexcel F1="IPPSSM"
putexcel G1="IPPSSM-miss"

forvalues i=2(1)201{
set seed 123456`i'
use Data_5M_`t'.dta

*** Step 1. Generate indicators for school and student selection ***

** Generate the school sample for H = 2000 conditions **
save Data_5M_`t'_school.dta
keep schid v1_l v2_l p2_l
duplicates drop
gen s2=.
replace s2=rbinomial(1,p2_l)
total s2
matrix define A=r(table)
while A[1,1]<10{
    matrix drop A
    replace s2=rbinomial(1,p2_l)
    quietly total s2
    matrix define A=r(table)
}
matrix drop A

** Below are the codes for H = 50 conditions:
* replace s2=rbinomial(1,p2_l)
* quietly total s2
* matrix define A=r(table)
* while A[1,1]!=6{
  * matrix drop A
  * replace s2=rbinomial(1,p2_l)
  * quietly total s2
  * matrix define A=r(table)
* }

* Assign half schools to treatment and half schools to control

gsample 50, percent wor strata(s2) generate(z2)
replace z2=. if s2==0

* Estimate school weights w_schl
capture logit s2 v1_l
predict prob_schl,p
gen w_schl=1/prob_schl
save Data_5M_'t'_school.dta, replace

* Merge school data back into student dataset
use Data_5M_'t'.dta
quietly merge m:1 schid using Data_5M_'t'_school.dta, nogenerate
save Data_5M_'t'.dta, replace

* Delete school level data file
erase Data_5M_'t'_school.dta

** Generate student sample **
gen s1=.
gen y_l=.
replace s1=rbinomial(1,p1_l) if s2==1
replace y_l=y1_l if (s1==1 & z2==1)
replace y_l=y0_l if (s1==1 & z2==0)

*** Step 2. Estimation ***

* True PATE
quietly mean(TE)
matrix define A1=r(table)
putexcel A`i' = A1[1,1]

* True SATE
quietly mean(TE) if s1==1
matrix define A2=r(table)
putexcel B`i' = matrix(A2[1,1])

* Estimator 1. Unadjusted Internally valid ATE
quietly mixed y_l z2|| schid:
matrix define B1 = e(b)
putexcel C`i' = matrix(B1[1,1])

* Estimator 2. IPP-School
quietly mixed y_l z2|| schid:,pweight(w_schl)
matrix define B2 = e(b)
putexcel D`i' = matrix(B2[1,1])

* Estimator 3. IPP-School+Student separate (IPPSSS)
gen p_stud1=0
forvalues j=1(2000){
capture logit s1 x1 if(schid=="j")
if c(rc)==0{ predict p`j',p 
    quietly replace p`j'=0 if (schid!=`j')
    quietly replace p_stud1=p_stud1+p`j'
    quietly replace p`j'=. if (schid!=`j')
    drop p`j'
} else if !inlist(c(rc), 2000, 2001, 430) {
    exit c(rc)
}

replace p_stud1=. if (p_stud1==0)
gen w_stud1=1/p_stud1
* Apply student and school weights to the model

capture mixed y_l z2[pweight=w_stud1]|| schid:,pweight(w_schl)
if c(rc)==0{
    matrix define B3 = e(b)
    *display B3[1,1]
    putexcel E`i'=matrix(B3[1,1])
}

* Estimator 4. IPP-School+Student multi (IPPSSM)
* Generate student weights from MLM

gen x1_lv1_l=x1_l*v1_l
gen x1_lv2_l=x1_l*v2_l

gen cons = 1
eq sch_c: cons
eq sch_m3: x1_l

corr s1 v2_l
matrix define C=r(C)
if C[1,2]!=. {
    capture gllamm s1 x1_l v1_l v2_l x1_lv1_l x1_lv2_l , i( schid ) nrf(2)
    eqs(sch_c sch_m3)family(binom) `link(logit)` adapt iterate(100)
    if c(rc)==0{
        gllapred p_11, mu
gen w_11=1/p_11
        quietly mixed y_l z2[pweight=w_11]|| schid:,pweight(w_schl)
        matrix define B4 = e(b)
        putexcel F`i'=matrix(B4[1,1])
    }
}

* Estimator 5. IPPW-School+Student multi-misspecified (IPPSSM-miss)

capture gllamm s1 x1_l v1_l v2_l , i( schid ) nrf(2) eqs(sch_c sch_m3)family(binom) `link(logit)` adapt iterate(100)
if c(rc)==0{
    gllapred p_12, mu
gen w_12=1/p_12
    quietly mixed y_l z2[pweight=w_12]|| schid:,pweight(w_schl)
    matrix define B5 = e(b)
    putexcel G`i'=matrix(B5[1,1])
}

* Remove all generated data and save the original dataset
keep schid- p1_l
save Data_5M_`t'.dta, replace
Appendix C

Table 1.

Results of inverse probability of participation weighting on participating students.

<table>
<thead>
<tr>
<th></th>
<th>Mean in Sample</th>
<th>Mean in Weighted Sample</th>
<th>SD in Population</th>
<th>Unweighted [SMD]</th>
<th>Weighted [SMD]</th>
</tr>
</thead>
<tbody>
<tr>
<td>School 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>0.52</td>
<td>0.47</td>
<td>0.55</td>
<td>0.50</td>
<td>0.11</td>
</tr>
<tr>
<td>Grade level</td>
<td>0.81</td>
<td>0.55</td>
<td>0.53</td>
<td>0.50</td>
<td>0.53</td>
</tr>
<tr>
<td>% free/reduced</td>
<td>0.63</td>
<td>0.73</td>
<td>0.71</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.10</td>
<td>0.15</td>
<td>0.14</td>
<td>0.36</td>
<td>0.16</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.10</td>
<td>0.09</td>
<td>0.13</td>
<td>0.28</td>
<td>0.03</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1289.61</td>
<td>1285.28</td>
<td>1273.96</td>
<td>84.67</td>
<td>0.05</td>
</tr>
<tr>
<td>School 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>0.53</td>
<td>0.55</td>
<td>0.53</td>
<td>0.50</td>
<td>0.03</td>
</tr>
<tr>
<td>Grade level</td>
<td>0.54</td>
<td>0.53</td>
<td>0.52</td>
<td>0.50</td>
<td>0.03</td>
</tr>
<tr>
<td>% free/reduced</td>
<td>0.25</td>
<td>0.32</td>
<td>0.31</td>
<td>0.47</td>
<td>0.16</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.24</td>
<td>0.05</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.11</td>
<td>0.14</td>
<td>0.14</td>
<td>0.35</td>
<td>0.09</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.21</td>
<td>0.02</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1315.03</td>
<td>1308.72</td>
<td>1310.20</td>
<td>87.13</td>
<td>0.07</td>
</tr>
<tr>
<td>School 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>0.56</td>
<td>0.49</td>
<td>0.47</td>
<td>0.50</td>
<td>0.15</td>
</tr>
<tr>
<td>Grade level</td>
<td>0.72</td>
<td>0.52</td>
<td>0.50</td>
<td>0.50</td>
<td>0.39</td>
</tr>
<tr>
<td>% free/reduced</td>
<td>0.51</td>
<td>0.62</td>
<td>0.63</td>
<td>0.49</td>
<td>0.23</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.05</td>
<td>0.07</td>
<td>0.06</td>
<td>0.26</td>
<td>0.08</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.12</td>
<td>0.16</td>
<td>0.16</td>
<td>0.37</td>
<td>0.10</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.02</td>
<td>0.04</td>
<td>0.05</td>
<td>0.20</td>
<td>0.12</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1264.36</td>
<td>1280.26</td>
<td>1284.11</td>
<td>78.36</td>
<td>0.21</td>
</tr>
<tr>
<td>School 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>0.63</td>
<td>0.59</td>
<td>0.56</td>
<td>0.49</td>
<td>0.09</td>
</tr>
<tr>
<td>Grade level</td>
<td>0.48</td>
<td>0.49</td>
<td>0.51</td>
<td>0.50</td>
<td>0.01</td>
</tr>
<tr>
<td>% free/reduced</td>
<td>0.82</td>
<td>0.79</td>
<td>0.82</td>
<td>0.41</td>
<td>0.06</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.17</td>
<td>0.19</td>
<td>0.22</td>
<td>0.40</td>
<td>0.05</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.17</td>
<td>0.15</td>
<td>0.12</td>
<td>0.35</td>
<td>0.08</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.19</td>
<td>0.02</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1266.52</td>
<td>1266.83</td>
<td>1266.64</td>
<td>78.94</td>
<td>0.00</td>
</tr>
<tr>
<td>School 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>0.37</td>
<td>0.46</td>
<td>0.45</td>
<td>0.50</td>
<td>0.19</td>
</tr>
<tr>
<td>Grade level</td>
<td>0.66</td>
<td>0.47</td>
<td>0.47</td>
<td>0.50</td>
<td>0.38</td>
</tr>
<tr>
<td>% free/reduced</td>
<td>0.31</td>
<td>0.38</td>
<td>0.34</td>
<td>0.49</td>
<td>0.14</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.05</td>
<td>0.07</td>
<td>0.09</td>
<td>0.25</td>
<td>0.08</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.09</td>
<td>0.13</td>
<td>0.10</td>
<td>0.34</td>
<td>0.11</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.17</td>
<td>0.09</td>
<td>0.09</td>
<td>0.28</td>
<td>0.31</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1310.31</td>
<td>1300.07</td>
<td>1306.14</td>
<td>77.05</td>
<td>0.13</td>
</tr>
<tr>
<td>School 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% male</td>
<td>0.41</td>
<td>0.48</td>
<td>0.46</td>
<td>0.50</td>
<td>0.13</td>
</tr>
<tr>
<td>Grade level</td>
<td>0.30</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.38</td>
</tr>
<tr>
<td>% free/reduced</td>
<td>0.62</td>
<td>0.58</td>
<td>0.61</td>
<td>0.49</td>
<td>0.08</td>
</tr>
<tr>
<td>% ELL status</td>
<td>0.05</td>
<td>0.07</td>
<td>0.08</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>% Disability</td>
<td>0.14</td>
<td>0.19</td>
<td>0.15</td>
<td>0.39</td>
<td>0.12</td>
</tr>
<tr>
<td>% Gifted</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>District Math Assessment (SY 2013-2014 Quarter 1)</td>
<td>1305.79</td>
<td>1277.84</td>
<td>1278.55</td>
<td>73.50</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Note. This table shows the balance between the within school study sample and within school target population before and after weighting for six out of the seven participating schools. Student weights were calculated for each participating student using formula \( \hat{w}_{hk} = 1/\bar{p}(S_{hk} = 1) \). Probability of participation was estimated using logistic regression \( \ln \left( \frac{\hat{p}(S_{hk} = 1)}{1 - \hat{p}(S_{hk} = 1)} \right) = X \beta + e_{hk} \), where \( S_{hk} = 1 \) when student is a participant. Independent variables Xs’ are observed student-level covariates listed in this table. Six logistics regression models were run,
one for each school. Within each school, population of interest is defined as all Grade 1 and Grade 2 students in this school, which includes participants and non-participants. [SMD] is absolute standardized mean difference, calculated by the absolute difference between (weighted) sample and population means divided by the standard deviation of population.