A peer reviewed, open-access electronic journal: ISSN 1531-7714

Performance of Data Mining Methods for Propensity Score Analysis with Multilevel Data

Meng Fan, *HumRRO*
Henry May, *University of Delaware*
Zachary Collier, *University of Connecticut*

Abstract: Propensity score methods have been developed primarily for single-level data structures. In educational studies, however, data typically have a hierarchical structure, where the probability of receiving treatment is a function of both individual and cluster-level factors. Using Monte Carlo simulations, this study compared two tree-based data mining approaches to two parametric models for propensity score estimation. All four estimation methods with an appropriately specified outcome model provided unbiased treatment effect estimates. However, under non-additivity and non-linearity, the data mining approaches provided a better performance on predicting the propensity score. Also, hidden bias from unobserved covariates greatly impacted the estimates of causal effect - rendering all propensity score analysis approaches invalid. The generalized boosting modeling-multilevel logistic regression combination provided more accurate treatment effect estimates among the eight estimation and outcome model combinations.

Keywords: Causal inference; Propensity score analysis; Hierarchical structure; Data mining approaches, GBM

Introduction

In observational studies, since propensity scores are typically unknown and unobservable, they must be estimated. The success of propensity score analysis (PSA) depends on the propensity score model's adequate specification, such that the balancing property is satisfied. Therefore, it is crucial to know what methods produce the most accurate estimation of propensity scores under different scenarios. Although numerous educational studies use propensity score methods, very few studies incorporate data mining approaches to estimate propensity scores. This study seeks to provide guidelines for educational researchers and practitioners to draw better causal inferences using data mining approaches with multilevel data.

We compared two tree-based data mining approaches (i.e., generalized boosting modeling, generalized linear mixed-effects model trees) to two parametric models (i.e., multiple logistic regression, multilevel logistic regression) for propensity score estimation under different simulated multilevel settings. We investigated: 1) How do the propensity scores obtained by parametric and data mining techniques compare based on (a) mean squared error and (b) the correlation between the actual probability of being in the treatment group and the estimated propensity score?, 2) To what extent do the relative bias of treatment effect estimates, proportion bias reduction of treatment effects, and the standard error of the treatment effects vary across the method of propensity score estimation, propensity score model, and outcome model?, and 3) Do the relative bias of treatment effect estimates depend on the sample size, ratios of treatment exposure, treatment effects, and levels of intraclass correlations (ICC)?

Prior studies have examined tree-based propensity score estimation, multilevel propensity score analysis, and the effects of nonlinearity in separate contexts. However, we know less about how parametric and data mining approaches compare within a unified multilevel framework, especially when the goal is to both predict treatment assignment and obtain precise effect estimates. This study aims to fill that gap by comparing two parametric methods and two tree-based methods across multilevel simulation conditions. These conditions differ in functional-form complexity and covariate omission. We also include an empirical illustration using PISA 2018 data to complement the simulation results.

Literature Review

Rubin's Causal Model

Rubin's causal model (a.k.a., the potential outcomes model; Rubin, 1974, 1978) focuses on the formal mathematical and statistical perspective related to causal inference (West & Thoemmes, 2010). In Rubin's causal model, treatment effects are estimated by comparing the potential outcomes that would have been observed for an individual under different conditions. Since potential outcomes for all treatment conditions cannot be observed for all units, Rubin's causal model requires several key assumptions, including Strong Ignorability in Treatment Assignment (SITA) and Stable Unit Treatment Value Assumption (SUTVA) (Rosenbaum et al., 1983; Rubin, 2010).

SITA assumption refers to the process used to assign units to conditions. It requires the assignment condition to be independent and not associated with the outcome and other factors. SUTVA is defined as an "a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive" (Rubin, 1986, p. 961). In other words, SUTVA means the observation on one unit should not be affected by the assignment of treatments to the other units.

Propensity Score Analysis in the Multilevel Context

In a multilevel context, the observations within a cluster tend to be more similar than those in other clusters, which demonstrates dependencies among individuals within clusters (Raudenbush & Bryk, 2002). Additionally, treatment assignment may be dependent on cluster characteristics. For example, a student's decision to participate in an after-school program may partly depend on the compositional characteristics or policies of the student's school. Therefore, the treatment assignment can be viewed as a combination of individual and cluster-level factors. This implies the need for multilevel modeling in estimating propensity scores.

A random intercepts model allows each Level-2 unit to have different intercepts but has common slopes (Raudenbush et al., 2002). A propensity score for a unit i, $e(X_{ij})$, can be estimated from the treatment assignment T_i on the Level-1 covariate vector X_i and Level-2 covariate vector W_i:

$$\text{Ln}\left[\frac{e(X_{ij})}{1-e(X_{ij})}\right] = \tau_{0j} + \beta X_{ij} + \pi W_j + \mu_{0j} \quad (1)$$

In this logistic regression equation, τ_{0j} represents the mean log-odds of propensity scores across clusters. βX_i is a vector of regression coefficients and Level-1 covariates. πW_j is a vector of regression coefficients and Level-2 covariates. μ_{0j} refers to the cluster-level residual and is assumed to be normally distributed with mean zero and variance □. By fixing the slopes, this model assumes that the effects of X_{ij} on treatment assignment are constant across clusters (Hong & Raudenbush, 2006).

Hong and Raudenbush (2005) first extended the propensity score analysis to the multilevel context. They applied a multilevel propensity score stratification method to evaluate the effects of the kindergarten retention policy using the Early Childhood Longitudinal Study Kindergarten cohort (ECLS-K) data. Hong et al. (2006) explained the proposed approach in detail. This study provided a general framework for making causal inferences in multilevel observational studies. Later, the application of PSA in a multilevel context was further examined by other researchers (e.g., Kelcey, 2009; Arpino & Mealli, 2011, Li, 2013; Leite et al., 2015).

In general, multilevel modeling is widely used for PSA due to its familiarity among educational researchers and its feasibility in statistical software (e.g., R, SAS). However, some researchers addressed several limitations of multilevel modeling for PSA, including insufficient attention to non-linearity and non-additivity in the estimation model and low precision with large numbers of covariates (Keller, Kim & Steiner, 2015; Setoguchi et al., 2008). These limitations could result in biased estimates of treatment effects (Rubin, 2010).

Data Mining Approaches

Breiman, Friedman, Stone and Olshen (1984) introduced classification and regression trees algorithms (CART). CART are non-parametric classifiers that construct hierarchical decision trees by splitting data among classes of the criterion at a given step (node) accordingly to an "if-then" rule applied to a set of predictors, into two child nodes repeatedly, from a root node that contains the whole sample. Compared with traditional approaches (e.g., logistic regression), CART algorithms have several advantages in solving classification and prediction problems (e.g., propensity score estimation). However, as the stopping rule for partitioning is loose, the tree based on CART algorithm may have unstable classification and prediction results (Hastie, Tibshirani & Friedman, 2009). Several approaches have been proposed to remedy the limitations of the CART, including bagging, boosting, and random forests.

GBM keeps the primary features of CART algorithms, but it can produce a smoother fit by combining many simple trees (Burgette, McCaffrey, Griffin, 2015). GBM utilizes the "forward stagewise additive algorithm" when estimating propensity scores. Based on tree-based methods, a new algorithm named the generalized linear mixed-effects model tree (GLMERTREE) was developed to take into account for nested data structure (Fokkema, Smits, Zeileis, Hothorn, & Kelderman, 2018). This method has been found to outperform other tree-based methods and generalized linear models in simulation studies. Also, unlike other tree-based approaches with random effects such as mixed effects regression trees (Hajjem, Bellavance & Larocque, 2011), GLMERTREE allows continuous as well as non-continuous response variables.

With the continuing focus on examining the effectiveness of educational programs, it is critical that methodologists continue to evaluate the applicability of PSA in varied data structures. Data mining approaches have received substantial attention in recent decades and have been widely used in big data applications. Several studies have shown that data mining approaches are an improvement over traditional methods for estimating propensity scores, including a few studies that extended those algorithms to a

multilevel context. This study aims to investigate the performance of two data mining approaches in a two-level hierarchical context.

Methods

Simulation Design

The simulation design for this study is informed by earlier simulation studies (e.g., Arpino et al., 2011; Leite, et al., 2015; Li, et al., 2013; Thoemmes et al., 2011) and several nationwide studies (e.g. Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 [ECLS-K: 2011]). The main objective for the simulation design was to reflect common scenarios for sample size and structure, treatment assignment, and effects on outcomes from impact evaluations of education interventions.

Covariates. In this study, population treatment assignment was manipulated at the individual level. Specifically, the population treatment assignment has a multilevel nature with ten Level-1 ($X_1, X_2, \ldots, X_{10}$) and four Level-2 covariates ($W_1, W_2, \ldots, W_4$). Three different distributions were selected to represent three different types of variables. X_1, X_4, X_7 and W_1 were generated from a Bernoulli distribution (i.e., binary) with probability of success p=.50. X_2, X_5, X_8 and W_2 were generated using discrete uniform distributions from 0 to 1. X_3, X_6, X_9, X_{10}, W_3 and W_4 were generated using standard normal distributions with mean of zero and standard deviation (SD) of one. Additionally, X_1, X_2 and X_3 and W_1 were correlated with the treatment assignment only. X_7, X_8, X_9 and W_3 were correlated with the outcome only. X_4, X_5, X_6 and W_2 were regarded as confounders that were correlated with both the treatment assignment and outcome. X_{10} and W_4 were regarded as unmeasured confounders (Table 1).

Population Treatment Assignment Models. When evaluating the performance of propensity score estimation methods for multilevel data, existing studies have relied on multilevel logistic models (e.g., Arpino et al., 2011; Li et al., 2013). However, using such a model can make the estimation techniques that utilize logistic regression (i.e., multivariate logistic regression) perform better than they would otherwise, since data generated from a Bernoulli with a logit will favor analyses that employ a logit link and a Bernoulli

Table 1. Properties of Level-1 and Level-2 Covariates

	Correlated With			Unmeasured
	Treatment	**Outcome**	**Both**	
Level-1				
X_1	✓			
X_2	✓			
X_3	✓			
X_4			✓	
X_5			✓	
X_6			✓	
X_7		✓		
X_8		✓		
X_9		✓		
X_{10}			✓	✓
Level-2				
W_1	✓			
W_2			✓	
W_3		✓		
W_4			✓	✓

distribution. In the current study, this problem is avoided by using a beta-binomial mixed effects model simulation to generate the data. This design aims to reduce built-in alignment between the data-generating process and any single estimation method, thus providing a fair comparison of parametric and tree-based approaches.

A beta-binomial mixed effects model extends the beta-binomial generalized linear regression to the inclusion of random effects in the linear predictor of the model. Conditional on some random effects u, the response variable y (i.e., representing treatment assignment) follows a beta-binomial distribution of parameters m, p and ϕ

$$y \mid u \sim BB(m, p, \phi),\ u \sim Lognormal(0, D) \tag{2}$$

where m is maximum score number in each beta-binomial observation (which is fixed at 1); p indicates the probability parameter of occurrence; ϕ indicates the value of the dispersion parameter of the conditional beta-binomial distribution; and D is the variance of the random effects. The lognormal distribution was applied to simulate the random effect of school intercept.

Following the marginal beta-binomial regression approach, the probability parameter of the beta-binomial distribution is connected with some given covariates $X_1,\ldots, X_n$ and the random effects by means of a cauchit (tangent) link function (i.e., as opposed to a logit).

$$\text{cauchit}(p) = \tan\left(\pi p - \frac{\pi}{2}\right) = \beta * X + Z * u \tag{3}$$

here β is the matrix of fixed effects and Z is the matrix of random effects.

In this study, three scenarios (A, B, and C) were considered in the population treatment assignment, which differ with the degree of linearity and/or additivity of modeled associations between treatment and the covariates. Scenario A assumes linear associations between treatment assignment and covariates. The population treatment assignment model for Scenario A is shown as follows.

$$\text{cauchit}(T_{ij} = 1) = \tau_{0j} + \sum_{b=1}^{9} \beta_b X_{bij} + \beta_{10} X_{10ij} + \sum_{n=1}^{3} \eta_n W_{nij} + \eta_4 W_{4ij} + S_{0j} + \mu_{0j} \tag{4}$$

$$T_{ij} \sim BB(m, p, \phi)$$

In this equation, τ_{0j} represents the mean log-odds of propensity scores across clusters. S_{0j} was the cluster intercepts drawn from a normal distribution. The random slope of Level-1 covariates were set to zero which assumed the same effects of Level-1 covariates on treatment across clusters. μ_{0j} is a proxy for unmeasured cluster-level covariates (i.e., a cluster-level residual). μ_{0j} was drawn from a lognormal distribution as defined earlier.

Scenario B describes the moderate non-linearity and non-additivity by adding three quadratic terms and one interaction term for Level-1 covariates. The population treatment assignment model for Scenario B is shown as follows.

$$\begin{aligned}\text{cauchit}(T_{ij} = 1) = \tau_0 + \sum_{b=1}^{9} \beta_b X_{bij} + \beta_{10} X_{10ij} + \beta_{11} X_{3ij}^2 + \beta_{12} X_{5ij}^2 + \beta_{13} X_{6ij}^2 + \\ \beta_{14} X_{1ij} * X_{2ij} + \sum_{n=1}^{3} \eta_n W_{nij} + \eta_4 W_{4ij} + S_{0j} + \mu_{0j}\end{aligned} \tag{5}$$

$$T_{ij} \sim BB(m, p, \phi)$$

In Equation 5, β_{11}, β_{12}, and β_{13} represent the regression coefficients for the three quadratic terms. β_{14} represents the regression coefficient for the interaction term. The remaining terms were defined similarly as in Equation 4.

Scenario C describes a more complex situation than Scenario B by adding a cross-level interaction term (X_4 and W_1). The population treatment assignment model for Scenario B is shown as follows.

$$\text{cauchit}(T_{ij} = 1) = \tau_0 + \sum_{b=1}^{9} \beta_b X_{bij} + \beta_{10} X_{10ij} + \beta_{11} X_{3ij}^2 + \beta_{12} X_{5ij}^2 + \beta_{13} X_{6ij}^2 + \beta_{14} X_{1ij*} X_{2ij} + \sum_{n=1}^{3} \eta_n W_{nij} + \eta_4 W_{4ij} + \beta_{15} X_{4ij*} W_{1ij} + S_{0j} + \mu_{0j} \quad (6)$$

$T_{ij} \sim \text{BB}(m, p, \phi)$

In Equation 6, β_{15} represents the regression coefficient for the cross-level interaction term. The remaining terms were defined similarly as in Equation 4.

In sum, the above three scenarios were designed to represent a progression from simple to increasingly complex treatment-assignment structures. Then we can evaluate whether data mining methods offered practical advantages as nonlinearity and interaction effects increased.

Population Outcome Models. Secondary to the population treatment model above, the population outcome model was specified as follows:

$$Y_{ij} = \alpha_0 + \sum_{b=4}^{9} \psi_b X_{bij} + \psi_{10} X_{10ij} + \sum_{n=2}^{3} \kappa_n W_{nij} + \kappa_4 W_{4ij} + T_{ij} + \theta_{0j} + \varepsilon_{ij} \quad (7)$$

$\theta_{0j} \sim N(0, \xi)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$

In the above equation, Y_{ij} are the observed outcomes and α_0 is an intercept. ψ_b are the fixed effects coefficients of seven Level-1 covariates X_4 through X_{10} and κ are the fixed effects coefficients of Level-2 covariates W_2 through W_4. γ is the treatment effect. T_{ij} is the treatment indicator, θ_{0j} represents the influence of unobserved cluster-level factors on the individual-level outcome. ε_{ij} is the level-1 residual. The sum of ξ and σ^2 is constrained to equal 1, with the ratio of $\xi/(\xi+\sigma^2)$ constrained to equal an assumed intraclass correlation (ICC, defined below).

We specified treatment-assignment and outcome models separately rather than using the same functional form. This was done to avoid giving artificial advantage to methods aligned with the data-generating treatment model and to better reflect applied educational observational studies, in which treatment selection and outcome generation may follow different processes.

Sample Size. Sample characteristics from previous simulation studies under multilevel frameworks are varied. Arpino et al. (2011) applied sample sizes ranging from 500 to 4,000. Thoemmes et al. (2011) simulated 20 clusters with 50 units each to represent a realistic large sample size and 200 clusters with 500 people each to represent an asymptotic sample size. To capture the multilevel data structure, this study consisted of three sizes. First, 500 schools with 10 students each; second, 200 schools with 30 students each; and 60 schools with 150 students each. The overall sample size was 5,000, 6,000, and 9,000. These cluster conditions were selected since they represent a range of sample sizes (from small to large) based on educational data (Maas & Hox, 2005).

Treatment Exposure Level. Three levels of treatment exposure were manipulated in this study based on previous simulation and applied literature. The ratios of 1:3 and 1:1 (treated units vs. control units) were chosen to represent ranges of exposure levels in the simulation literature on the PSA (e.g., Rubin, 1979; Lingle, 2009). Besides, a 1:9 ratio was included since it is similar to the pioneer PSA study of Hong et al. (2005).

Treatment Effects. In existing propensity score simulation studies, the population treatment effects were normally set to 1.0 (e.g., Leite et al., 2015, Gurel, 2015) or 1.5 (e.g., Chen, 2014). Education interventions frequently have small effects or no effect at all (Kraft, 2018). Mark Lipsey and his colleagues (2012) found an average effect size of only 0.28 SD among a sample of 124 randomized trials. The current study used 0.3 and 0.5 as the population treatment effects, which were set as the regression coefficient for treatment in the population outcome model.

Intraclass Correlation Coefficient (ICC). The ICC represents the proportion of variance in the outcome variable that is explained by the grouping structure of the hierarchical model. It is calculated as a ratio of Level-2 error variance over the total error variance (Gelman & Hill, 2007; Snijders & Bosker, 1999):

$$\text{ICC} = \frac{\xi}{\sigma^2+\xi} \tag{8}$$

In educational research, ICC levels typically range from .05 to .30 (Raudenbush et al., 2002). As a result, the values of ICC included in this study were .10 and .30, both of which are plausible in educational clustered data (Niehaus, Campbell, & Inkelas, 2014).

In sum, we organized the simulation into three primary methodological contrasts. First, the treatment assignment model varied in functional-form complexity, ranging from linear/additive relations to nonlinear, interaction, and cross-level interaction structures. Second, the propensity score model varied in whether key confounders were fully represented or omitted. Third, treatment effects were estimated using both single-level and random-effects outcome models to evaluate the role of accounting for clustering. The remaining factors, including sample size, treatment exposure level, treatment effect size, and ICC, were included to assess the robustness of these patterns across plausible multilevel educational settings. The summary of simulation design factors is presented in Table 2. A total of 108 conditions were simulated to generate the data, and each condition was replicated 500 times. All the data generation procedures were performed in R 3.6.1 (R Core Team, 2018).

Table 2. Summary of Simulation Design Factors

Factor	Level
Scenario	A; B; C
Sample size (clusters * units)	500 * 10; 200 * 30; 60 * 150
Treatment exposure level	50%; 25%; 10%
Treatment Effects	.30; .50
ICC	.10; .30

Note. All factor levels will be fully crossed in the simulations

Analytic Procedures

Propensity Score Estimation Methods. Four estimation methods were compared in this study. First, multiple logistic regression (MLR) was used to estimate propensity scores using all Level-1 and Level-2 covariates while ignoring clustering. Second, a random effects model (RE), specified as a multilevel logistic regression model, was used to account for clustering by allowing the intercept to vary across clusters. Third, GBM was implemented in R using the WeightIt package, with the cluster indicator included alongside the Level-1 and Level-2 covariates. The final GBM model was selected using the es.max stopping rule, which optimized covariate balance based on standardized effect sizes. Fourth, GLMERTREE was implemented in R using the glmertree package, with clustering incorporated through the cluster identifier in the tree specification. In GLMERTREE, each terminal node is associated with a different set of fixed-effects regression coefficients while adjusting for global random effects, allowing the detection of subgroups with different fixed-effects parameter estimates while maintaining a common random-effects structure throughout the tree (Fokkema et al., 2018).

*Estimating Treatment Effects - Propensity score weighting***.** Propensity score weighting is similar to weighting in survey sampling that accounts for over- or under-represented subgroups (Lee, et al., 2010; Morgan & Todd, 2008). The logic of using propensity scores as weights is to control the influence of units by weighting their responses based on their propensity scores (McCaffrey et al., 2004). One example of propensity score weighting is the inverse probability of treatment weighting (IPTW; Rosenbaum, 1987), which is defined as:

$$W_i = \frac{Z_i}{e(X_i)} + \frac{1-Z_i}{1-e(X_i)} \quad (9)$$

In this equation, Z_i is an indicator variable denoting whether or not the *i*th unit is treated, and $e(X_i)$ is the propensity score for the *i*th unit. Each unit weight is equal to the inverse of the probability of receiving the treatment that the subject received.

The IPTW method for estimating the treatment effects was used to weight each observation to make it representative of the entire population. For each estimation method, two separate treatment models were investigated. The first model (Model A1; B1; C1) included the same variables that were used to generate the data. Besides, a misspecified model (Model A2; B2; C2) was also implemented. Model A2; B2; C2 excluded the two unmeasured covariates (X_{10}, W_4). Two outcome models, including the single level model (MLR) and the random effects (RE) model, were investigated separately. The outcome variable was regressed on the treatment indicator and covariates, and the weights were specified as sampling weights. In total, 16 combinations were compared.

Evaluation Criteria

First, the mean squared error (MSE) was calculated to evaluate the performance of different propensity score estimation methods. Second, both relative bias of treatment effects and proportion bias reduction were calculated. Relative bias measures the average tendency that the simulated treatment effect is below or above the treatment effect, which is represented by a percentage. The following equation was used to calculate the relative bias:

$$B(\hat{\theta}) = \frac{\hat{\theta}-\theta}{\theta} \quad (10)$$

In this equation, $\hat{\theta}$ is the average treatment effect estimate in the baseline with propensity scores, and θ is the population treatment effect estimate. Hoogland and Boomsma (1998) suggested that relative bias within -.05 and .05 is acceptable. Additionally, the proportion bias reduction of treatment effects was obtained via:

$$PRB(\hat{\theta}) = \frac{\theta_{null}-\theta_{ps}}{\theta_{null}} \quad (11)$$

where θ_{null} is the relative bias in the null model where no propensity score weighting was performed to obtain treatment effects and θ_{ps} is the relative bias of treatment effects after propensity score weighting was added to the outcome model. It is suggested that a propensity score estimation procedure performs well if it removes at least 90% of the selection bias (Gurel, 2015; Cochran, 1968). In addition, the standard error of the treatment effects was calculated to assess the precision of the resulting treatment effect estimates.

A series of analyses of variance (ANOVA) were used to investigate the effects of manipulated conditions on the relative bias of treatment effects. The post hoc tests using Tukey's method for significant comparisons are also applied. In these analyses, the relative bias of treatment effects was used as the dependent variable. The independent variables include propensity score estimation methods, propensity score models, outcome models, sample size, treatment exposure level, treatment effects, and ICC. The eta squared was used as the measure of effect size on the relative bias of treatment effects. Meanwhile, since ICC determines the nested

structure of data, its interaction effects with estimation methods and outcome models were investigated as well.

Results

A total of 288 combinations of conditions were investigated on the MSE and correlation coefficients between the actual probability of being in the treatment group and estimated propensity score, and 576 combinations of conditions were on the relative biases of treatment effects, percent bias reduction of propensity score estimation, and standard errors. To facilitate interpretation, the results are summarized with emphasis on three broad issues: how the four methods compared in propensity score prediction, how they compared in treatment effect estimation in terms of bias, bias reduction, and precision, and how these patterns changed across treatment-model complexity and other simulated design factors.

Across the simulated conditions, four general patterns emerged. First, when confounders were adequately represented in the propensity score and outcome models, all four estimation methods yielded very small relative bias in treatment effect estimates. Second, under nonlinear and non-additive treatment assignment, the data mining methods generally improved propensity score prediction, but these gains did not consistently translate into meaningfully lower treatment effect bias. Third, omission of important covariates had a much larger impact on bias than the choice of propensity score estimation method. Fourth, random effects outcome models generally improved precision, with the GBM-RE combination showing the most consistent advantage in standard errors.

Scenario A

Under the linear and additive treatment-assignment condition, RE performed best in propensity score prediction, but all four methods performed similarly in treatment effect estimation when the confounders were adequately represented.

The first two columns of Table 3 show that the MSE values for Model A1 were lower than those for Model A2. For Model A1, the MSE values for MLR, RE, GBM, and GLMERTREE were .079, .050, .062, and .052, respectively, indicating that RE had the lowest MSE and MLR had the largest. A similar pattern was observed for Model A2, where RE (.055) again had the lowest MSE and MLR (.086) had the largest one. Consistent with these results, the first two columns of Table 4 show that RE and GLMERTREE had the highest correlation with the true treatment probability in Model A1 (.770), followed closely by GBM (.769), whereas MLR had the lowest correlation (.585). In Model A2, RE again had the highest correlation (.745), followed by GBM (.726), whereas MLR had the lowest (.531).

Despite these differences in propensity score prediction, treatment effect estimation was highly similar across methods when the models were appropriately specified. As shown in Table 5, the relative biases for Model A1 were extremely small regardless of whether the outcome model accounted for clustering, with the GBM-RE combination producing the smallest value, very close to zero. Table 6 further shows that all eight

Table 3. Average MSE across Simulated Conditions

	Scenario A		Scenario B		Scenario C	
	Model A1	**Model A2**	**Model B1**	**Model B2**	**Model C1**	**Model C2**
MLR	.079	.086	.093	.098	.110	.116
RE	.050	.055	.067	.070	.087	.090
GBM	.062	.068	.060	.067	.072	.079
GLMERTREE	.052	.066	.046	.064	.051	.069

Table 4. Average Correlation between Actual Probability of Receiving the Treatment and Estimated Propensity Scores across Conditions

	Scenario A		Scenario B		Scenario C	
	Model A1	**Model A2**	**Model B1**	**Model B2**	**Model C1**	**Model C2**
MLR	.585	.531	.480	.439	.463	.417
RE	.770	.745	.668	.652	.615	.594
GBM	.769	.726	.766	.718	.761	.714
GLMERTREE	.770	.683	.797	.707	.807	.722

Table 5. Average Relative Bias across Simulated Conditions for Scenario A

	Model A1		Model A2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	-.002	-.002	-1.071	-1.105
RE	-.003	-.001	-1.037	-1.027
GBM	.001	.000	-.962	-.957
GLMERTREE	.005	.001	-1.042	-1.077

Table 6. Percent Bias Reduction across Simulated Conditions for Scenario A

	Model A1		Model A2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	94.5%	95.2%	75.2%	75.7%
RE	93.8%	94.8%	76.7%	78.2%
GBM	95.2%	96.0%	75.7%	76.4%
GLMERTREE	93.6%	94.8%	73.7%	74.2%

combinations removed more than 90% of the initial bias in Model A1. In contrast, when the two unmeasured covariates were omitted in Model A2, relative bias became very large across all methods, and none of the methods removed more than 80% of the initial bias. These results indicate that correct covariate representation mattered much more than the choice of estimation method.

Table 7 shows that accounting for clustering in the outcome model improved precision. In Model A1, the GBM-RE combination yielded the smallest standard error (.003), followed by MLR-RE (.004) and GLMERTREE-RE (.004). A similar pattern was observed in Model A2, where the GBM-RE combination again produced the smallest standard error (.005). Table 8 shows that omission of unmeasured covariates had by far the largest effect on relative bias in Scenario A (G η^2 = .482), whereas the effect of propensity score estimation method was statistically significant but substantively very small (G η^2 = .001).

Scenario B

Under moderate nonlinearity and non-additivity, the data mining methods improved propensity score prediction relative to the parametric models, but treatment effect estimates remained very similar across methods when confounders were adequately represented.

The third and fourth columns of Table 3 show that GLMERTREE had the lowest MSE in both Model B1 and Model B2, whereas MLR had the largest. The third and fourth columns of Table 4 show a similar

Table 7. Standard Errors across Simulated Conditions for Scenario A

	Model A1		Model A2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	.005	.004	.010	.006
RE	.007	.005	.010	.007
GBM	.004	.003	.008	.005
GLMERTREE	.006	.004	.009	.006

Table 8. Effect Size on the Relative Bias of Treatment Effects

	Effect Size (G η2)		
	Scenario A	**Scenario B**	**Scenario C**
Sample Size	.002 **	.001 **	.001 **
ICC	.002 **	.001 **	.001 **
Ratio of Treatment Exposure	.056 **	.044 **	.040 **
Treatment Effect Size	.001 **	.001 **	.003 **
Propensity Score Estimation Method	.001 **	.001 **	.001 **
Inclusion of Unmeasured Covariates	.482	.523 **	.596 **
Outcome Model	.000	.000	.000
ICC * Propensity Score Estimation Method	.000	.000	.001 **
ICC * Outcome Model	.001 **	.001 **	.001 **

Note. **p<.01, *<.05

pattern for the correlations with the true treatment probability: GLMERTREE had the highest correlation in Model B1 (.797), followed by GBM (.766), whereas GBM had the highest correlation in Model B2 (.718), followed by GLMERTREE (.707). MLR had the lowest correlations in both models. These results suggest that the data mining methods were more successful in recovering treatment probabilities when the treatment-assignment process included nonlinear and interaction terms.

However, these gains in propensity score prediction did not translate into meaningful differences in treatment effect bias when covariates were adequately represented. Table 9 shows that the relative biases for Model B1 were again extremely small across all methods and outcome models, with MLR-MLR producing the value closest to zero. Table 10 shows that all methods removed more than 90% of the initial bias in Model B1. By comparison, when the two unmeasured covariates were omitted in Model B2, the relative biases became very large across all methods, and bias reduction dropped to roughly 70%. Thus, as in Scenario A, omission of important covariates had a much larger effect on treatment effect estimation than the choice of propensity score estimation method.

Table 9. Average Relative Bias across Simulated Conditions for Scenario B

	Model B1		Model B2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	.000	.001	-1.006	-1.007
RE	.005	.003	-.992	-1.001
GBM	-.003	-.002	-.964	-.956
GLMERTREE	-.005	-.005	-.982	-.983

Table 10. Percent Bias Reduction across Simulated Conditions for Scenario B

	Model B1		Model B2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	92.9%	94.3%	69.6%	71.2%
RE	92.9%	93.9%	71.0%	71.3%
GBM	93.6%	95.2%	70.8%	72.5%
GLMERTREE	91.1%	93.3%	69.8%	71.5%

Table 11 shows that standard errors were generally smaller when the random effects outcome model was used. For Model B1, the GBM-RE and MLR-RE combinations had the smallest standard errors (.003). For Model B2, all four estimation methods combined with the RE outcome model yielded similar standard errors of approximately .005. Table 8 again shows that omission of unmeasured covariates was the dominant factor affecting relative bias in Scenario B (G η^2 = .523), whereas the effect of method choice remained very small.

Table 11. Standard Errors across Simulated Conditions for Scenario B

	Model B1		Model B2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	.004	.003	.008	.005
RE	.004	.004	.007	.005
GBM	.004	.003	.008	.005
GLMERTREE	.006	.004	.008	.005

Scenario C

Under the most complex treatment-assignment condition, which included a cross-level interaction, the tree-based methods again showed better propensity score prediction, but the overall conclusions for treatment effect estimation remained unchanged.

The fifth and sixth columns of Table 3 show that GLMERTREE had the lowest MSE in both Model C1 and Model C2, whereas MLR had the largest. The fifth and sixth columns of Table 4 likewise show that GLMERTREE had the highest correlation with the true treatment probability in both Model C1 (.807) and Model C2 (.722), whereas MLR had the lowest. These results are consistent with the expectation that more flexible methods perform better in recovering treatment probabilities when treatment assignment becomes more complex.

As in the previous scenarios, however, the treatment effect results were very similar across methods when confounders were adequately represented. Table 12 shows that the relative biases for Model C1 were again extremely small across all methods and outcome models, with the MLR-MLR and GBM-MLR combinations producing the smallest values. Table 13 shows that all eight combinations removed more than 90% of the initial bias in Model C1. When the unmeasured covariates were omitted in Model C2, all methods showed substantial bias and removed less than 70% of the initial bias. This pattern again indicates that omitted confounding was far more consequential than the choice of estimation method.

Table 14 shows that the random-effects outcome model again improved precision. In Model C1, the MLR-RE, RE-RE, and GBM-RE combinations yielded similarly small standard errors of approximately .003. In Model C2, the MLR-RE, RE-RE, and GBM-RE combinations again had very similar standard errors of about .004. According to Table 8, omission of unmeasured covariates had the largest effect on relative

bias in Scenario C (G η^2 = .596), whereas the influence of propensity score estimation method remained very small.

Table 12. Average Relative Bias across Simulated Conditions for Scenario C

	Model C1		Model C2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	.001	.002	-1.060	-1.065
RE	.003	.004	-.976	-.972
GBM	-.001	.002	-.959	-.944
GLMERTREE	.008	.005	-.987	-.991

Table 13. Percent Bias Reduction across Simulated Conditions for Scenario C.

	Model C1		Model C2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	93.1%	94.1%	62.8%	63.5%
RE	93.4%	93.8%	66.0%	66.4%
GBM	93.5%	94.9%	65.6%	67.3%
GLMERTREE	90.7%	92.7%	64.1%	65.6%

Table 14. Standard Errors across Simulated Conditions for Scenario C

	Model C1		Model C2	
	Outcome Model 1 (MLR)	**Outcome Model 2 (RE)**	**Outcome Model 1 (MLR)**	**Outcome Model 2 (RE)**
MLR	.004	.003	.007	.004
RE	.003	.003	.005	.004
GBM	.003	.003	.007	.004
GLMERTREE	.005	.004	.007	.005

Empirical Example Using PISA 2018

In this section, we applied the four propensity score estimation methods to data from the U.S. sample of the Programme for International Student Assessment (PISA) 2018 study. This example study estimated the effect of chronic bullying victimization on students' mathematics achievement using four propensity score estimation methods. PISA collected data for 15-year-old students every three years. For the U.S. sample, this included all students attending school in grade 7 or higher. The PISA 2018 U.S. sample was selected by a two-stage sample design, including a sample of schools and a sample of students within those selected schools (Organisation for Economic Co-operation and Development [OECD], 2019). The purpose of this empirical example was to provide a practical illustration of how the four propensity score methods can be implemented and compared in a real multilevel educational dataset.

The binary treatment variable represented whether students experienced chronic bullying victimization during the past 12 months. A student was coded as treated (1) if the student reported once a week or more on at least one of the six bullying items. A student was coded as control (0) if the student reported never or almost never on all six items. We controlled seven student-level variables, including gender, mother's education, home wealth resources, books at home, food insecurity, subjective socioeconomic status, and

immigrant status. We additionally included four school-level covariates, including school type, school size, school economic disadvantage, and school resource hindrance. The covariates were chosen based on prior literature showing that student demographic, family, and school-context characteristics are associated with bullying victimization and academic performance (OECD, 2019; Kljakovic & Hunt, 2016).

The U.S. PISA 2018 sample consists of 4,552 students from 154 schools. Students with missing treatment status or covariate values included in the primary propensity score model were excluded from the final analytic sample. The final sample included 1,788 students from 119 schools.

After estimating the propensity scores, inverse probability of treatment weights were constructed for each method. We evaluated the covariate balance after weighting using standardized mean differences. We fit the random effects model to estimate treatment effects separately for each of the 10 plausible values of mathematics. The final treatment effect estimate was obtained by averaging the 10 coefficients, and the final standard error combined within- and between-plausible-value variability (OECD, 2019). Since the purpose of the empirical example was methodological illustration rather than population inference, the PISA student sampling weights were not incorporated.

Figure 1 shows the covariate balance before and after weighting. Before weighting, the treatment and control groups showed imbalances on several covariates, with the greatest imbalance observed for subjective SES, school resource hindrance and food insecurity. After applying propensity score weights, the MLR, RE, and GBM methods achieved acceptable covariate balance, as all absolute standardized mean differences fell below the .10 criterion. The GLMERTREE approach produced a weaker balance, with several covariates remaining above the threshold (e.g., subjective SES, school size).

Table 15 shows the treatment effect across the four propensity score estimation methods. The estimated treatment effects were similar, ranging from -17.376 to -18.897. Among the four methods, GBM yielded the smallest standard error, whereas the other three methods produced slightly larger but comparable standard errors. Overall, the results suggest a consistent negative effect of chronic bullying victimization on mathematics achievement across methods.

Table 15. Treatment Effect Estimates by Propensity Score Estimation Method

Estimation Method	Estimate	Standard Error
MLR	-17.376	5.298
RE	-17.723	5.388
GBM	-17.507	3.995
GLMERTREE	-18.897	5.305

Discussion

Previous studies have shown that data mining approaches outperform the traditional parametric method (e.g., multiple logistic regression) in estimating the propensity score (e.g., McCaffrey et al., 2004; Setoguchi, et al., 2008) in single-level settings. As an extension of that work, this study used Monte Carlo simulation to compare two tree-based data mining approaches (i.e., generalized boosting modeling, generalized linear mixed-effects model trees) to two parametric models (i.e., multiple logistic regression, multilevel logistic regression) for propensity score estimation in multilevel settings.

Summary of Findings

In Scenario A, RE provided a lower MSE and a higher correlation coefficient between the true and estimated propensity scores. These results matched some of the previous studies (e.g., Thoemmes et al., 2011; Gurel, 2015; Su & Cortina, 2009). Besides, when non-additivity and non-linearity exist, GBM and

GLMERTREE provided better performance in predicting the propensity score. The poor performance of prediction by MLR and RE resulted from the violation of linearity or additivity assumption. This will produce erroneous predictions of the propensity scores. Similarly, Lin, Zhu, and Chen (2019) found that data mining approaches including GBM provided lower MSE than MLR when the relationship between the treatment assignment and the covariates was non-linear and non-additive.

Meanwhile, although GBM and GLMERTREE had better performance in predicting the true propensity scores when non-additivity and non-linearity exist, they did not outperform MLR and RE in the relative bias of treatment effects. Furthermore, the overall performance on the relative bias among the four estimation methods did not degrade as the increased complexities of treatment assignment model in Scenario B and C. These results indicated that the goal of the propensity score model is to efficiently control for confounders instead of predicting treatment assignment. The correctly specified outcome models can provide another chance to control for confounders and variables only correlated with the outcome. These results support the findings presented by Setoguchi et al. (2008) in a comparison of MLR, neural network, CART, and pruned CART in the propensity score model. Approximately unbiased estimates of the treatment effect estimates can be obtained from both traditional parametric modeling and data mining propensity score methods in a variety of scenarios differing by additivity and linearity.

GBM produced the lowest standard errors of treatment effects across the three scenarios. One explanation of the decreased sampling variability of the GBM is that it can produce fewer extreme values of the weights (McCaffrey et al., 2004). The use of the piecewise constants in GBM has the effect of flattening the estimated propensity scores at the extreme values of the predictors. This minimizes the chance of obtaining predicted probabilities near 0 or 1, thus preventing the high variability in weights that can be problematic for propensity score weighting.

The current study also showed that the MLR and RE outcome models performed similarly on the relative bias of treatment effects. However, RE is more effective to improve precision by producing lower standard errors of treatment effects. One explanation is that MLR treats the units of analysis as independent observations. One consequence of failing to recognize nested structures is that standard errors of individual-level regression coefficients will be larger. This can result in an understatement of statistical significance. As Li et al. (2013) suggested, considering cluster effects in at least one stage greatly improves the estimates in bias and standard errors of treatment effects.

ANOVA results suggested that the inclusions of unmeasured covariates played a larger role than the propensity score estimation and outcome models. If the unmeasured covariates are confounders, which are not included in the propensity score and outcome models, the SITA assumption is not met. Not all the covariates that impact the treatment assignment have been properly accounted for. To this end, the bias in treatment effect estimates are exacerbated. This finding was in agreement with studies conducted by Kelcey (2009) and Arpino et al. (2011). Kelcey (2009) claimed that the inclusions of appropriate variables in a model were more crucial than the use of a specific propensity score model.

In summary, there are several primary findings in this study. First, missing covariates render all PSA approaches invalid. Second, all of the four estimation methods with an appropriately specified outcome model can provide unbiased treatment effect estimates. Third, although the MLR and RE outcome models performed similarly on the relative bias of treatment effects, RE offers improved precision by producing smaller standard errors of treatment effects. Fourth, among the eight estimation and outcome model combinations, GBM-RE combination consistently provided more precise treatment effect estimates across the simulated conditions. This pattern was broadly consistent with the PISA 2018 empirical illustration, in which MLR, RE, and GBM achieved acceptable covariate balance and yielded highly similar treatment effect estimates, while GBM produced the smallest standard error.

Implications for Educational Researchers

Findings from this study confirm some earlier results, while also illustrating some new findings, all of which have substantial implications for educational researchers who conduct PSA in multilevel settings. Five key recommendations based on this study's results are as follows:

1. Researchers should develop a thorough understanding of the treatment assignment mechanism to identify whether there are critical unmeasured covariates that are related to both treatment assignment and the outcome. If the potential for important confounders exists, then no modeling strategy is likely to produce valid results.
2. Under the most realistic conditions (e.g., those with missing covariates, nonlinearity, and/or interactions), data mining techniques (i.e., GBM, GLMERTREE) tend to outperform parametric models in terms of precision, but only slightly so.
3. When researchers find the variation at the cluster-level in the form of random intercepts but no variation in the form of a random slope, employing multilevel modeling in the outcome model is adequate to provide valid and reliable treatment effects.
4. Once the multilevel structure in the data has been captured by the multilevel outcome model, it matters little if some cluster level information is omitted from the propensity score estimation model.
5. Researchers should allow for nonlinear patterns and interactions when estimating propensity scores.

GBM has been confirmed to perform well in the single-level setting (McCaffrey et al., 2004). The current study showed that it does provide relatively reliable and stable performance in multilevel settings. As discussed by McCaffrey et al., the GBM algorithm has some features that improve propensity score estimation performance. In R, the twang package is well established to fit the GMB to estimate propensity scores (Ridgeway, et al., 2017). This package can also provide diagnostics statistics (e.g., covariate balance) for PSA. Alternatively, WeightIt package was newly developed to conduct propensity score weighting (Greifer, 2019).

GLMERTREE, as one of the tree-based methods, was newly developed to take account of nested data structures (Fokkema et al., 2018). Despite this, the performance of GLMERTREE was not superior to GBM in this study. Meanwhile, the computation time for GLMERTREE is much longer than GBM. In addition, other results suggest that the propensity score estimation model can ignore clustering with little to no detriment, as long as the outcome impact model recognizes clustering. To this end, if applied researchers would like to use a data mining approach to perform the PSA, GBM is still preferable.

In practice, applied researchers may consider multiple PSA techniques to evaluate treatment effect estimates. In observational studies, it would be ideal to minimize both bias and variance. While bias is the primary concern in observational studies, low variance is an appealing feature as well. If different PSA techniques produce similar point estimates of the treatment effect, the one that yields a smaller variance of that treatment effect (i.e., a more precise estimate) may be preferred (McCaffrey et al., 2004; Stuart & Green, 2008).

Limitations and Future Research

While this is a merit to simulation research, it also limits the generalizability of the findings. First, this study did not consider varied correlation between covariates. Future research can be done to incorporate varied correlations among covariates and examine how the PS methods discussed in the current study perform on estimation of the effect on the outcome variable. Second, this study included only propensity score weighting as the conditioning method. Future research can assess the performance of data mining approaches to estimate the propensity score using matching and stratification conditioning methods. Third,

when using GBM to generate the propensity score in this study, only one algorithm specification was specified. Further research should include different algorithm specifications for GBM with multilevel data. Meanwhile, other promising data mining methods (e.g., random forest, neural networks) should be extended to deal with selection bias in multilevel observational studies.

Received: 8/4/2025. **Accepted:** 6/17/2026. **Published:** 7/13/2026.

Citation: Fan, M., May, H., & Collier, Z. (2026). Performance of data mining methods for propensity score analysis with multilevel data. *Practical Assessment, Research, & Evaluation*, 31(1), Article 15. Available online: https://doi.org/10.7275/pare.3322**.**

Corresponding Author: Meng Fan, HumRRO. Email: fanmeng2046@gmail.com

References

Arpino, B., & Cannas, M. (2015). *Comparing different approaches for propensity score matching with clustered data: a simulation study (*Working Paper No. 43). RECSM.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*, 1770-1780.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman and Hall.

Burgette, L. F., McCaffrey, D. F., & Griffin, B. A. (2015). Propensity score estimation with boosted regression. In W. Pan, H. Bai (Eds). *Propensity score analysis: Fundamentals, developments, and extensions* (pp. 49-73). New York: Guilford Press.

Chen, J. (2014). *A Bayesian propensity score approach for multilevel observational studies.* (Unpublished doctoral dissertation). University of Wisconsin Madison, Madison, WI.

Fokkema, M., Smits, N, Zeileis, A., Hothorn, T., Kelderman, H. (2018). Detecting treatment subgroup Interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods, 50*(5), 2016-2034.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press.

Greifer, N. (2019). Weighting for covariate balance in observational studies. Retrieved from https://cran.r-project.org/web/packages/WeightIt/WeightIt.pdf.

Gurel, S. (2015). *Dealing with selection bias in multilevel observational studies: An evaluation of propensity score and direct estimation procedures.* (Unpublished doctoral dissertation). University of Florida, Gainesville, FL.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451-459.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of statistical learning: Data mining, inference, and prediction*, Stanford, CA: Springer.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Evaluation and Policy Analysis, 27*(3),205-224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. *Journal of the American Statistical Association, 101*, 901-910.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, *26*, 329-367.

Kelcey, B. (2009). *Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings*. (Unpublished doctoral dissertation). University of Michigan, Ann Arbor, MI.

Keller, B., Kim, J. S., & Steiner, P. M. (2015). Neural networks for propensity score estimation: Simulation results and recommendations. In L. A. van der Ark, D. M. Bolt, S.-M. Chow, J. A. Douglas, & W.-C. Wang (Eds.), Quantitative psychology research (pp. 279-291). New York, NY: Springer.

Kljakovic, M., & Hunt, C. (2016). A meta-analysis of predictors of bullying and victimisation in adolescence. *Journal of Adolescence, 49*, 134–145.

Kraft, M. A. (2018). *Interpreting effect sizes of education interventions* (Brown University Working Papers). Providence. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_2018_interpreting_effect_sizes.pdf.

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

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., MacInnes, J. W., & Sandbach, R. (2015). An evaluation of weighting methods based on propensity scores to reduce selection bias in multilevel observational studies. *Multivariate Behavioral Research, 50*, 265-284.

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013), Propensity score weighting with multilevel data. *Statistics in Medicine, 32*, 3373–3387.

Lin, L., Zhu, Y. & Chen, L. (2019). Causal inference for multi-level treatments with machine-learned propensity scores. *Health Services and Outcomes Research Methodology, 19*, 106-126.

Lingle, J. A. (2009). *Evaluating the performance of propensity scores to address selection bias in a multilevel context: A Monte Carlo simulation study and application using a national data set.* (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA.

Lipsey, M. W. & Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. National Center for Special Education Research. (NCSER 2013-3000). Washington, DC: U.S. Department of Education.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86−92.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403-425.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology, 38*(1), 231-281.

Niehaus, E., Campbell, C. M., & Inkelas, K. K. (2014). HLM behind the curtain: unveiling decisions behind the use and interpretation of HLM in higher education research. *Research in Higher Education*, *55*(1), 101-122.

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2017). Toolkit for weighting and analysis of nonequivalent groups. Retrieved from https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf.

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

Rosenbaum P.R. (1987). Model-based direct adjustment. *Journal of the American Statistician, 82*, 387-394.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of American Statistical Association, 81*(396), 961-962.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318- 328.

Rubin, D. B. (2010). Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods, 15*(1), 38-46.

Setodji, C. M., McCaffrey, D. F., Burgette, L. F., Almirall, D., and Griffin, B. A. (2017). The right tool for the job: Choosing between covariate balancing and generalized boosted model propensity scores. *Epidemiology, 28*(6), 802–811.

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R. J., Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546-555.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London; Thousand Oaks, Calif.: Sage Publications.

Stuart, E. & Green, K. (2008). Using full matching to estimate causal effects in non-experimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology, 44*(2), 395-406.

Thoemmes, F., & West, S. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*, 514-543.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Hagedorn, M. C., Leggitt, J., & Najarian, M. (2015). *Early childhood longitudinal study, kindergarten class of 2010-11 (ECLS-K: 2011), user's manual for the ECLS-K: 2011 kindergarten first grade data file and electronic codebook public version (NCES 2015-078).* Washington, DC: U.S. Department of Education.

West, S., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on Causal Inference. *Psychological Methods, 15*(1), 18-37.