Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: http://www.tandfonline.com/loi/hmbr20

An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies

Walter L. Leite, Francisco Jimenez, Yasemin Kaya, Laura M. Stapleton, Jann W. MacInnes & Robert Sandbach

To cite this article: Walter L. Leite, Francisco Jimenez, Yasemin Kaya, Laura M. Stapleton, Jann W. MacInnes & Robert Sandbach (2015) An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies, Multivariate Behavioral Research, 50:3, 265-284, DOI: 10.1080/00273171.2014.991018

To link to this article: http://dx.doi.org/10.1080/00273171.2014.991018

Multivariate Behavioral Research, 50:265–284, 2015

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2014.991018

Routledge
Taylor & Francis Group

An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies

Walter L. Leite, Francisco Jimenez, and Yasemin Kaya
College of Education, University of Florida

Laura M. Stapleton
College of Education, University of Maryland

Jann W. MacInnes
College of Education, University of Florida

Robert Sandbach
Santa Fe College

Observational studies of multilevel data to estimate treatment effects must consider both the nonrandom treatment assignment mechanism and the clustered structure of the data. We present an approach for implementation of four propensity score (PS) methods with multilevel data involving creation of weights and three types of weight scaling (normalized, cluster-normalized and effective), followed by estimation of multilevel models with the multilevel pseudo-maximum likelihood estimation method. Using a Monte Carlo simulation study, we found that the multilevel model provided unbiased estimates of the Average Treatment Effect on the Treated (ATT) and its standard error across manipulated conditions and combinations of PS model, PS method, and type of weight scaling. Estimates of between-cluster variances of the ATT were biased, but improved as cluster sizes increased. We provide a step-by-step demonstration of how to combine PS methods and multilevel modeling to estimate treatment effects using multilevel data from the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K).

Researchers and program evaluators frequently conduct observational studies (i.e., quasi-experimental studies) to estimate the effects of educational interventions or policies using secondary data from large surveys, such as the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K; National Center for Education Statistics, 2010). Educational surveys produce multilevel data because of the organizational structure of education systems, where students are clustered into classrooms and schools. These surveys frequently take advantage of the natural clustering of individuals by using sampling methods such as cluster sampling and multistage sampling (Lohr, 1999). Analyses of these data must consider the clustering of individuals in order to obtain estimates with adequate standard errors (Van Landeghem, De Fraine, & Van Damme, 2005), using model-based approaches (e.g., multilevel regression models; Goldstein, 2003; multilevel structural equation models; Muthén, 1994), design-based approaches (e.g., balanced repeated replication, jackknife, bootstrap; Rodgers, 1999) or a combination of both (Rabe-Hesketh & Skrondal, 2006; Sterba, 2009). Furthermore, estimates of the effects of educational interventions or policies using observational survey data will be biased if the analysis does not account for the influence of individual-level or cluster-level confounding variables, which are covariates related to both the assignment to the intervention and the

Correspondence concerning this article should be addressed to Walter L. Leite, School of Human Development and Organizational Studies, University of Florida, 1215 Norman Hall, Gainesville, FL 32611. E-mail: walter.leite@coe.ufl.edu

outcome (Rubin, 1973; Winship & S. L. Morgan, 1999). This selection bias is due to the fact that, in observational studies, the treatment assignment mechanism is not random and therefore the treated and untreated groups being compared do not have equivalent potential outcomes in the absence of the intervention (Shadish, Cook, & Campbell, 2002). Although randomized experiments have increasingly been used for educational program evaluation (Shadish, 2002), including a focus on randomly assigning classrooms or schools to treatments instead of students (Schochet, 2012; Zhu, Jacob, Bloom, & Xu, 2011), observational studies using survey data are a very common approach for educational program evaluation, as indicated by a recent review by Thoemmes and Kim (2011).

Propensity score (PS) methods are a group of strategies that aim to reduce selection bias by balancing differences between treated and untreated individuals on observed covariates. Propensity scores, $p(T_i = 1|X)$, are the predicted probabilities of being assigned to treatment T given a vector of covariates X. Commonly used PS methods include one-to-one greedy matching, optimal full matching (OFM), stratification, and PS weighting (see Guo & Fraser, 2010, 2015; Holmes, 2014 for a review). Rosenbaum and Rubin (1983) have shown that reduction of selection bias can be obtained by balancing treatment and control groups with respect to the distribution of propensity scores. However, this result only holds if the PS model accounts for all sources of selection bias. Therefore, in multilevel research designs where individuals within clusters (i.e., individual level or Level 1) are exposed to a treatment, it is essential that the estimation of the propensity score accounts for both individual- and cluster-level effects on the probability of treatment assignment (Arpino & Mealli, 2011; Kelcey, 2011b; Li, Zaslavsky, & Landrum, 2013; Thoemmes & West, 2011). This design is the quasi-experimental version of a multisite experimental design. By contrast, the use of PS methods with a cluster-level treatment requires accounting for sources of selection bias at the cluster level or higher. This design is the quasi-experimental version of a cluster randomized trial (Moerbeek, 2005; Steiner, Kim, & Thoemmes, 2013). Because there has been a surge in applications of PS methods to estimate the effects of individual-level educational interventions with multilevel data (e.g., Anand, Mizala, & Repetto, 2009; Berends, 2010; Doyle, 2009; Lockheed, Harris, & Jayasundera, 2010; P. L. Morgan, Frisco, Farkas, & Hibel, 2010; Ou & Reynolds, 2010), we will focus on an application of PS methods to the multilevel data obtained from a quasi-experimental study that mimics the multisite experimental design.

There have been few methodological studies of PS methods for estimating individual-level treatment effects with multilevel data. This study has three objectives that refer to issues that either have not been addressed or have been insufficiently investigated in previous research: Our first objective is to present a strategy for the implementation of four PS methods (i.e., one to many greedy matching, OFM, stratification, and PS weighting) to estimate treatment effects with multilevel models and multilevel structural equation models through the creation of weights, scaling of weights, and estimation of weighted models with the multilevel pseudo-maximum likelihood estimation method (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006). Our second objective is to compare models to estimate propensity scores for multilevel data, given that previous studies reached divergent conclusions: Arpino and Mealli (2011) recommended logistic regression with fixed-cluster effects, while Thoemmes and West (2011) recommended multilevel logistic regression with random cluster effects. The third objective is to evaluate the adequacy of this approach to provide estimates of treatment effects, their standard errors, and estimates of between-cluster variability of treatment effects. The first objective is addressed in the next four sections. Then, Objectives 2 and 3 are addressed in a Monte Carlo simulation study manipulating cluster sizes, size of the cluster effects on treatment assignment, intra-class correlation, and variability of covariate effects on treatment assignment between clusters. We follow the simulation study with an example of estimating treatment effects with multilevel observational data using PS methods: the effect of children receiving special educational services in third grade on reading achievement using data from the ECLS-K.

SELECTION BIAS IN MULTILEVEL DESIGNS

For multilevel data where each individual i in cluster j is non-randomly assigned to either a treatment group ($T = 1$) or a control group ($T = 0$), Rubin's potential outcomes framework (Holland, 1986; Rubin, 1974) defines the average treatment effect (ATE) as $E[Y_{ij}^t] - E[Y_{ij}^c]$, where $E[Y_{ij}^t]$ is the expected value of the outcome for all individuals under the treated condition, and $E[Y_{ij}^c]$ is the expected value of the outcome for all individuals under the control condition (S. L. Morgan & Harding, 2006; Winship & S. L. Morgan, 1999). The average treatment effect on the treated (ATT) is $E[Y_{ij}^t|T = 1] - E[Y_{ij}^c|T = 1]$, which is the difference between the expected value of the observed outcomes of the treated individuals and the expected value of the potential outcomes of the treated individuals. In this study we focus on using PS methods with multilevel data to estimate the ATT. The required assumptions are the strong ignorability of treatment assignment, adequate common support (i.e., overlap of PS distributions), and stable unit treatment value assumption (SUTVA; Rubin, 1986). Strong ignorability of treatment assignment requires that the potential outcomes are independent of treatment assignment conditional on covariates (i.e., $[Y^t, Y^c \perp T|X]$). For estimating the ATE, strong ignorability implies both $E[Y_{ij}^c|T = 0] = E[Y_{ij}^c|T = 1]$ and $E[Y_{ij}^t|T = 1] = E[Y_{ij}^t|T = 0]$, but only the first condition is needed for the ATT (S. L. Morgan & Winship, 2015).

This difference has consequences in the required area of common support, which is a region of the distribution of propensity scores where values exist for both treated and untreated individuals, and defines the set of individuals for which inferences about the treatment effect can be made from the available data. While ATE requires adequate common support for both treated and untreated individuals, the ATT only requires that the distribution of propensity scores of the treated is contained within the distribution of propensity scores of untreated individuals. SUTVA requires that the potential outcomes of the study participants are independent of both the assignment mechanism and the treatment status of other individuals in the sample. Multilevel data may lead to violations of SUTVA because individuals in the same cluster may affect the potential outcomes of others through mechanisms such as personal communication and sharing of resources. In this paper we discuss PS methods assuming that SUTVA holds, but Hong and Raudenbush (2006) provide an example of how to account for violation of SUTVA.

With multilevel observational data, a nonrandom treatment assignment mechanism at the individual level may depend on individual-level and cluster-level covariates, as well as their same-level or cross-level interactions. Furthermore, the effects of individual-level covariates on the probability of treatment assignment may vary from cluster to cluster (i.e., the individual-level covariates may have random slopes). In applications of PS methods to multilevel data, these cluster effects on treatment assignment can be dealt with using three distinct approaches (Griswold, Localio, & Mulrow, 2010): (1) Ignore cluster effects by using a PS model with individual-level covariates only, (2) Pool cluster-specific ATTs, which consists of using a PS method with potentially different PS models separately for each cluster and then combine the ATT estimates across clusters weighting by cluster size, and (3) Obtain a marginal ATT, which consists of using a PS method with a single PS model for the entire data to obtain the ATT. Arpino and Mealli (2011) found that when there are cluster-level confounders of treatment assignment, Option 1 results in substantially more bias than Option 3. Options 2 and 3 for obtaining the ATT across multiple clusters are identical to the methods of pooling conditional mean differences and computing marginal mean differences, respectively, discussed by Hong (2010) for obtaining the treatment effect across multiple strata.

Pooling cluster-specific ATTs has the advantage of removing concern about selection bias due to both observed and unobserved cluster-level confounders because they do not bias estimates of cluster-specific ATTs. The goal of PS methods in this case would be to balance the within-cluster distributions of individual-level confounding variables across treated and untreated groups. This strategy should be preferred when the cluster sizes are large enough to allow stable estimation of PS models and an adequate area of common support for the treated within each cluster (Kim & Seltzer, 2007). However, in survey data obtained through multistage sampling, cluster sizes are typically small, and this may complicate the estimation of cluster-specific PS models and/or result in poor common support and inadequate covariate balance within clusters (Arpino & Mealli, 2011; Steiner, et al., 2013). For these reasons, estimating a marginal ATT from the entire data is amenable to more general use with multilevel observational data with a variety of cluster sizes, including small clusters, and thus will be the focus of this study. In this case, the goal of PS methods would be to balance the marginal distributions of individual-level and cluster-level confounding variables across treated and untreated groups. Steiner et al. (2013) found that using PS methods across clusters is able to remove selection bias if the PS model is correctly specified. We discuss challenges and options of PS model specification for multilevel data in the next section.

PROPENSITY SCORE MODELS FOR MULTILEVEL DATA

To estimate propensity scores with a single model for the total sample, a critical difficulty is that the effect of confounders on treatment selection may vary across clusters. This variation could be modeled with a multilevel logistic regression model (i.e., a generalized mixed effects model) with a random intercept and random slopes of covariates (Kelcey, 2011b; Kim & Seltzer, 2007):

$$\text{logit}\,(T_{ij}=1)=\beta_0+\sum_{m=1}^{M}\beta_m X_{mij}+\sum_{n=1}^{N}\pi_n Z_{nj}+s_{0j}+\sum_{m=1}^{M}s_{mj}X_{mij}\;s_{0j},s_{mj}\sim N(0,\Sigma) \quad (1)$$

where β_0 is the intercept, β_m are the fixed effects of the M individual-level covariates X_{mij}, π_n are the fixed effects of N cluster-level covariates Z_{nj}, s_{0j} is a normally distributed random intercept, and s_{mj} are normally distributed random slopes of the individual-level covariates with mean of zero and covariance matrix Σ. This model could also include polynomial terms and interaction terms involving X_{mij} and Z_{nj}.

Alternatively, a researcher could use a logistic regression model with fixed-cluster effects (Allison, 2009; Arpino & Mealli, 2011), and include covariate-by-cluster interactions to capture the variation of covariate effects between clusters:

$$\text{logit}\,(T_{ij}=1)=\beta_0+\sum_{m=1}^{M}\beta_m X_{mij}+\sum_{j=1}^{J-1}\delta_j g_j+\sum_{m=1}^{M}\sum_{j=1}^{J-1}\upsilon_j g_j X_{mij} \quad (2)$$

where δ_j are the effects of the *J-1* dummy-coded cluster indicators g_j, and υ_j are the two-way interaction effects be-

tween g_j and the individual-level covariates X_{mij}. This model does not include cluster-level covariates because their effects are captured by the dummy-coded cluster indicators, but the model could be extended to include polynomial terms and other interaction terms involving X_{mij} and g_j.

The logistic regression model with fixed-cluster effects has the advantage of controlling all effects of both observed and unobserved cluster-level covariates (Allison, 2009; Arpino & Mealli, 2011), whereas the multilevel logistic regression model relies on the researchers' ability to identify and include observed cluster-level confounders in the model. Arpino and Mealli (2011) found that when a cluster-level confounder that correlates with individual-level confounders is omitted from the multilevel logistic regression model, this PS model can be improved by adding the means of the individual-level covariates to the model. However, when there is strong suspicion of unobserved cluster-level confounders, the use of a logistic regression model with fixed-cluster effects is recommended. A common difficulty of using the logistic regression model with fixed-cluster effects is that with a large numbers of small clusters, unstable propensity scores as well as predicted probabilities equal to zero or one may occur (Li, et al., 2013).

A major complication of using the multilevel logistic regression model and the logistic regression model with fixed-cluster effects is that applications of PS methods using multilevel observational data typically aim to account for confounding due to a large number of individual-level and cluster-level covariates. Furthermore, effects of individual-level covariates may vary from cluster to cluster and there could be multiple cross-level interactions. Including a large number of random slopes and/or cross-level interactions in Equations (1) or a large number of covariate-by-cluster interactions in Equation (2) is likely to result in model convergence problems. On the other hand, Arpino and Mealli (2011) found that even when interaction effects are omitted from the models in Equations (1) and (2), the increase in bias of treatment effect estimates may be negligible and there is a substantial improvement over a PS model omitting cluster effects.

Thoemmes and West (2011) compared a single-level logistic regression model that ignored clustering, a logistic regression model with fixed-cluster effects, and two multilevel logistic regression models with random intercepts and slopes (i.e., one including random effects in the estimated propensity scores and another including only fixed effects). Arpino and Mealli (2011) compared a single-level logistic regression model, a logistic regression model with fixed-cluster effects, and a multilevel logistic regression model with random intercepts. Although these authors compared similar models, their choices of manipulated conditions and PS method were quite different. Thoemmes and West (2011) simulated data with large cluster sizes (i.e., 50 and 500) and estimated the treatment effect with PS stratification within clusters or across clusters. Arpino and Mealli (2011) simulated conditions with balanced cluster sizes varying from small to large (i.e., 5, 10, 20, 25, 50, and 100), as well as unbalanced cluster sizes and focused on PS matching across clusters. Regardless of their dissimilar conditions, these studies reached similar conclusions with respect to the adequacy of PS models for analyses across clusters: PS analyses using either the logistic regression model with fixed-cluster effects or the multilevel logistic regression model result in unbiased treatment effect estimates, while the logistic regression model ignoring cluster effects performs poorly. However, Thoemmes and West (2011) found that the multilevel logistic regression model with random intercepts and slopes performed best in most conditions, while Arpino and Mealli (2011) concluded that the fixed-effects model performed best under almost all conditions simulated. Additionally, Arpino and Mealli (2011) found that the fixed-effects model performed acceptably even with very small cluster sizes (e.g., 5 or 10), but increasing the sample size at either the individual- or cluster-level improved the results with all PS models. The fixed-effects model outperformed the other two models in the presence of cross-level interaction effects, but there were no differences between the fixed-effects and multilevel models due to the hierarchical data being balanced or unbalanced (Arpino & Mealli, 2011).

PROPENSITY SCORE METHODS FOR MULTILEVEL DATA

Although there are many PS methods (e.g., one-to-one greedy matching, one-to-many greedy matching, optimal full matching, stratification, PS weighting), they can all be used to define observation weights that adjust the distributions of covariates so that they are similar for treated and untreated groups (i.e., obtain covariate balance) and reduce selection bias. These weights can be used for estimation in the same manner as sampling weights (Heeringa, West, & Berglund, 2010; Kish, 1965; Lohr, 1999). A key difference between PS methods is how coarse the weight is: one-to-one greedy matching can be viewed as a strategy to define sample weights where treated and untreated but matched individuals receive weights of one and untreated and unmatched individuals receive weights of zero; PS stratification combines both treated and untreated individuals into k strata, and as many as $k \times 2$ different weights are defined depending on the number of untreated observations in each stratum; PS weighting produces potentially as many different weights as there are observations in the sample. In essence, weights based on propensity scores adjust the observed sample to represent a pseudo-population where treated and untreated groups have similar distributions of covariates and therefore the treatment is unconfounded by measured covariates (Robins, Hernan, & Brumback, 2000). Consequently, much of the theory and methods developed for estimation with sampling weights in survey research (Heeringa, et al., 2010;

Kish, 1965; Lohr, 1999) apply to PS methods. Below we will describe the calculation of weights and implementation variations for common PS methods that can be applied to multilevel data. Then, we will discuss a multilevel model to estimate the ATT with multilevel pseudo-maximum likelihood estimation (Asparouhov, 2006), which can be used with the weights obtained from any PS method.

Greedy Matching

Greedy matching is one of the most widely used matching algorithms and includes nearest neighbor matching and caliper matching, among other variations. The greedy matching algorithm seeks to minimize the distance between each pair, yet does not minimize the total distance between all matched pairs (Austin, 2011). In addition, greedy matching is most commonly done without replacement, but matching with replacement performs better when the number of available matches is small (Rosenbaum, 1989). Greedy matching can be performed with a one-to-one nearest neighbor strategy or with a one-to-many strategy, where each treated individual is matched to multiple untreated individuals (Guo & Fraser, 2010). One-to-many matching is known to outperform one-to-one matching for estimating the ATT in general conditions (Cepeda, Boston, Farrar, & Strom, 2003; Gu & Rosenbaum, 1993). The use of a caliper, which is a maximum distance within which matches are allowed, has also been shown to improve matching performance, as well as enforce common support. Therefore, in this study we focus on an implementation of one-to-many greedy matching with replacement within a caliper. The weights for one-to-many greedy matching with replacement are (Ho, Imai, King, & Stuart, 2014):

$$w_i = T_i + (1 - T_i)\,\frac{n_M}{n_T}\sum_{1}^{n_{Ti}}\frac{1}{n_{Mi}} \tag{3}$$

where T_i is a binary treatment indicator equal to one for treated cases and zero for untreated cases, n_{Ti} is the number of treated cases that untreated case i was matched to, and $n_{Mi}$is the number of untreated cases matched to the same treated case as case i. n_M is the total number of matched cases and n_T is the total number of treated cases. The ratio n_M/n_T scales the weights of the matched individuals to sum to n_M. If matching is conducted without replacement, n_{Ti} becomes one. The use of this weight requires that untreated cases not within a caliper of a treated case and treated cases with no untreated cases within their calipers are dropped from the sample before calculation of weights to guarantee that $n_{Mi} \neq 0$. Because this weight is for estimating the ATT, treated units receive a weight of one.

Propensity Score Stratification

Stratification based on propensity scores consists of dividing the sample into strata that are similar with respect to propensity scores. Cochran (1968) showed that stratifying a single covariate into quintiles removes about 90% of selection bias in the treatment effect estimate. PS stratification has become a popular method for adjusting treatment effect estimates for selection bias and a review of applications of PS stratification by Thoemmes and Kim (2011) showed that researchers typically use between 5 and 20 strata, with 5 being the most common choice.

Estimating treatment effects with PS stratification can be accomplished by weighting the observations to account for disproportionate distributions of treated and untreated observations within strata. The weights for estimating the ATT with PS stratification were derived by Hong (2010):

$$w_i = T_i + (1 - T_i)\,\frac{n_{ts}n_u}{n_{us}n_t}\,, \tag{4}$$

where T_i is a binary indicator of treatment, $n_{ts}$is the number treated and $n_{us}$is the number untreated within stratum s, while n_u and n_t are the total number of untreated and treated in the sample, respectively. The weight for untreated units $n_{ts}n_u/n_{us}n_t$is equal to the ratio of the proportion treated to the proportion untreated in the stratum, and therefore is a nonparametric estimate of the odds of treatment within stratum and will sum to the sample size of untreated units, n_u. PS stratification weights for estimating the ATE and for multiple treatments are presented in Hong (2010, 2012).

Optimal Full Matching (OFM)

With optimal matching, treated individuals are matched with untreated individuals by minimizing the total distance between treated and untreated matched pairs (Austin, 2011) using network flow theory (Hansen, 2007; Rosenbaum, 1989). OFM attempts to match all untreated individuals in the dataset to a treated counterpart, resulting in no loss of sample size as long as there is an adequate area of common support. OFM to estimate the ATT results in the creation of strata where each stratum contains at least one treated individual and at least one untreated individual, minimizing both the within-strata and between-strata PS distances (Rosenbaum, 2010). OFM can be viewed as a generalization of PS stratification where the number of strata is optimized to reduce the distance between treated and untreated individuals, rather than defined a priori. Therefore, the formula for calculation of weights to estimate the ATT with OFM is the same as in PS stratification, as shown in Equation (4).

Propensity Score Weighting

PS weighting is simpler to implement than PS matching and stratification because weights are obtained directly from propensity scores, while matching and stratification requires the creation of groups of observations based on similarity in propensity scores before weights are calculated. However, because it uses propensity scores directly, PS weighting is less robust to misspecifications of the PS model and more

prone to bias due to extreme weights than PS matching and stratification (Hong, 2010, 2012). For estimation of the ATT, PS weights are defined as:

$$w_i = T_i + (1 - T_i)\frac{p(T_i = 1|X)}{1 - p(T_i = 1|X)}, \quad (5)$$

which equals one for treated individuals and the odds of treatment for untreated individuals. Equation (5) differs from inverse probability-of-treatment weighting (IPTW; Robins, et al., 2000), which is used to estimate the ATE. With PS weighting, the sum of the weights of the untreated individuals is not expected to sum to the untreated sample size.

MULTILEVEL MODELS TO ESTIMATE TREATMENT EFFECTS

Estimation of treatment effects with PS methods can be performed with nonparametric estimators, which can be differences between weighted means, or with parametric models. For a multilevel context where treatment was assigned at the individual level and pooling of cluster-specific ATTs is of interest, Li et al. (2013) presented a nonparametric clustered estimator, which is the weighted mean of the within-cluster treatment effects. A doubly robust version of this estimator could be obtained by including important covariates to make this estimator more robust to misspecifications of the PS model and to increase efficiency. Nonparametric doubly robust estimators with PS weighting are reviewed by Lunceford and Davidian (2004), Kang and Schafer (2007a, 2007b), and Bang and Robins (2005).

In this study we will focus on multilevel models (Goldstein, 2003; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012) for treatment effect estimation, because they can be expanded to include latent variables (Muthén, 1994), interactions (Leite & Zuo, 2011), and mediation (Preacher, Zyphur, & Zhang, 2010). This focus agrees with Ho et al.'s (2006) argument that PS methods can be viewed as a preprocessing phase to remove selection bias prior to treatment effect estimation with a wide variety of parametric models. The simplest specification of a multilevel model to estimate the treatment effect (in the absence of covariates) is:

$$\begin{aligned} y_{ij} &= \gamma_0 + \gamma_1 T_{ij} + u_{0j} + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma^2), u_{0j} \sim N(0, \tau) \end{aligned} \quad (6)$$

where y_{ij} is the outcome of individual i in cluster j, γ_0 is an intercept, γ_1 is the treatment effect, T_{ij} is the treatment indicator, u_{0j} is the random intercept of cluster j with variance τ and ε_{ij} is an individual-level residual with variance σ^2. In this random-intercepts model (Raudenbush & Bryk, 2002), the treatment effect is assumed to be constant across clusters. Relaxing this assumption leads to the random intercepts and slopes model:

$$\begin{aligned} y_{ij} &= \gamma_0 + \gamma_1 T_{ij} + u_{0j} + u_{1j} T_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma^2), u_{0j}, u_{1j} \sim N(0, \Phi) \end{aligned} \quad (7)$$

This model allows for both a random intercept $u_{0j}$and a cluster-specific component u_{1j} of the treatment effect, with covariance matrix Φ.

In a multilevel research design, reduction of selection bias due to clustering can be performed by accounting for cluster effects either in the PS model, as shown earlier, or in the multilevel model of the outcome, but that maximum reduction of bias is obtained by combining both strategies (Li, et al., 2013; Su & Cortina, 2009). For example, Thoemmes and West (2011) used multilevel models for both propensity scores and the outcomes. In contrast, Arpino and Mealli (2011) accounted for the multilevel structure of the data in the PS model, but estimated the ATT with a matching estimator (Abadie & Imbens, 2006) that does not account for clustering. Su and Cortina (2009) found that using multilevel models both to estimate propensity scores and the treatment effect resulted in the least biased parameter estimates and smallest root mean squared error. Li et al. (2013) obtained results that agree with Su and Cortina (2009), but also found that accounting for cluster effects in the outcome model was more effective than in the PS model in terms of reducing bias.

Combining PS methods with the multilevel models defined in Equations (6) and (7), as well as more complex multilevel models, requires including the weights in Equations (3), (4), or (5) in the model estimator as sampling weights. Snijders and Bosker's (2012) review multilevel modeling estimation methods with sampling weights, which also apply to weights from PS methods. They identify two estimation methods that have been developed for multilevel modeling with sampling weights: The probability-weighted iterative generalized least squares (PWIGLS) estimation method (Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998) implemented in the MLwin, HLM 6, and LISREL 8.7 statistical programs and the multilevel pseudo-maximum likelihood (MPML) estimation method (Asparouhov, 2006; Rabe-Hesketh & Skrondal, 2006) implemented in the Mplus and Stata programs. In this study we will use the MPML method because it can be applied to more general models (e.g., generalized linear mixed models, multilevel structural equation models) than the PWIGLS method (Snijders & Bosker, 2012). In the *R* statistical software, the *lme4* (Bates, Maechler, & Bolker, 2011) and *nlme* (Pinheiro & Bates, 2000) packages for multilevel modeling accept precision weights but not sampling weights. While sampling weights are inverse probabilities of selection, precision weights are inverse variances. Treating weights from PS methods as sampling weights or precision weights will not affect estimates of the treatment effect, but standard error estimates will differ (Snijders & Bosker, 2012). Similar to the use of sampling weights, the objective of using weights from PS methods is to avoid

bias in parameter estimates and standard errors, but not to improve precision. Therefore, when weights from PS methods are used as sampling weights, the precision of treatment effect estimates is not being optimized and the standard errors could be smaller or larger than when weights are not used (Snijders & Bosker, 2012). However, treating weights from PS methods as precision weights could adversely affect standard errors and Type I error rates, because larger precision weights reflect smaller standard errors, but larger sampling weights reflect larger uncertainty given that cases with larger sampling weights are representing a larger group in the pseudo-population (Snijders & Bosker, 2012).

Asparouhov (2006) presented the weighted likelihood function that is maximized in the MPML method, and demonstrated that the estimates obtained are approximately unbiased if three conditions are met: The first of these conditions is that both the cluster size and the sum of scaled weights within clusters are sufficiently large, but there are no guidelines in the literature of what magnitude of cluster sizes and sum of scaled weights would produce acceptable levels of bias. Weights are scaled by multiplying by a constant so that the sum of the weights is equal to a target characteristic of the sample, such as the total sample size. The second necessary condition is that the scaling of Level 2 weights is conditionally independent of the cluster sizes given covariates. The third condition is that the ratio of the within-cluster sum of scaled weights and cluster sizes are independent of cluster sizes given covariates (Asparouhov, 2006).

In multilevel modeling with sampling weights, the scaling of the Level 1 weights has been shown to affect the estimation of standard errors (Asparouhov, 2006; Pfeffermann et al., 1998; Rabe-Hesketh & Skrondal, 2006; Stapleton, 2002) and no clear method of scaling has been shown to be preferable in all situations. Three possible scaling methods for Level 1 weights that have been discussed in the literature are of interest here: (1) to scale weights to sum to the total sample size, which requires multiplying the weights by the scaling factor $\lambda = n/\sum w_{ij}$ (i.e., the inverse of the grand mean of the weights); (2) to scale weights so that the within cluster weights sum to the cluster size, which entails multiplying weights by $\lambda_{1j} = n_j/\sum_j w_{ij}$ (i.e., the inverse of the cluster mean of the weights); and (3) scale weights to make the within cluster weights sum to the effective sample size (Pfeffermann et al., 1998) with multiplying by the scaling factor $\lambda_{1j} = \sum_j w_{ij}/\sum_j w_{ij}^2$(Asparouhov, 2006). We will refer to these three types of weights as normalized, cluster-normalized, and effective, respectively. These weight-scaling options for multilevel models have been investigated by several researchers: Pfeffermann and colleagues reviewed their performance using PWIGLS estimation while Asparouhov (2006) and Rabe-Hesketh and Skrondal (2006) examined them in simple univariate and bivariate multilevel models using MPML estimation and Stapleton (2002) evaluated their performance with latent variable models under MPML. Pfeffermann concluded that under PWIGLS, when weights were non-informative, effective scaling of weights resulted in less bias in estimates of standard errors and of Level 2 variance. However, when the weights were informative, this effective weight-scaling method resulted in more bias in the variance estimates than just using weights that summed to the actual cluster size (i.e., the cluster-normalized weights) and therefore the recommendation was made to use cluster normalized weights (Pfeffermann et al., 1998). In both weight-scaling cases, the Level 2 variance tended to be underestimated and this underestimation decreased with increases in the sample size within clusters Conversely, with MPML estimation, it has been found that between-cluster variances were overestimated and within-cluster variance was underestimated with cluster sizes of 5 and 20 (Asparouhov, 2006) and with 10 and 20 (Stapleton, 2002). From these studies with MPML, estimates of fixed effects were less biased with cluster-normalized weight scaling than with effective weight scaling but the latter was preferred if interest was in the variance components (Asparouhov, 2006; Stapleton, 2002), but the preference depends on the informativeness of the weights and on the cluster size.

In this study we will focus on multisite quasi-experimental studies where clusters were selected with equal probability and treatment assignment occurs within cluster. In this scenario, cluster-level weights are equal to one and therefore the second condition given by Asparouhov (2006) for approximately unbiased estimates is met. However, in designs where clusters are selected with probability proportional to size or another sampling method based on unequal probabilities, or when the treatment assignment is at the cluster level, an examination of whether there is a correlation between Level 2 weights and cluster sizes should be performed.

Asparouhov's (2006) third condition is always met with cluster-normalized weights, because the ratio of the within-cluster sum of scaled weights and cluster sizes is always one. In balanced designs, the third condition is also always met with normalized weights or effective weights, because the cluster size is constant. In unbalanced designs, researchers planning to use normalized weights or effective weights should examine whether there is a correlation between cluster sizes and the ratio of the within-cluster sum of scaled weights and cluster sizes.

In the Monte Carlo simulation study presented in the next section, we will address the performance of treatment effect estimators under different scaling methods of weights, as well as cluster sizes, because there is no definitive evidence in the literature about which scaling method works best, and what qualifies as a sufficiently large cluster size. Furthermore, the selection bias inherent in observational studies would lead to weights that are informative, which suggests that the effective weighting approach may yield more accurate estimates of standard errors and variance components under MPML estimation. The research questions addressed in the simulation study are: Do estimates of the ATT, its standard error, and estimates of the between-cluster variance of the ATT obtained

with multilevel pseudo-maximum likelihood estimation depend on the weight-scaling method, PS method, cluster size, magnitude of cluster effects on treatment assignment, and intra-class correlation? Do estimates of the ATT depend on how the PS model accounts for cluster effects on the probability of treatment assignment and variability of covariate effects across clusters?

METHOD

We generated conditions where the probability of treatment assignment depended on five individual-level and five cluster-level confounders as well as on the random effects of clusters. The population treatment assignment model was:

$$\text{logit}\,(T_{ij}=1)=\beta_0+\sum_{m=1}^{5}\beta_m X_{mij}+\sum_{n=1}^{5}\pi_n Z_{nj}+s_{0j}+\sum_{m=1}^{5}s_{mj}X_{mij}+r_{ij} \quad (8)$$

and

$$(T_{ij}=1)\,if\ \text{logit}\,(T_{ij}=1)>0, else\,(T_{ij}=0) \quad (9)$$

The terms in Equation (8) were defined as in Equation (1). Five individual-level covariates X_{mij} and five cluster-level covariates Z_{nj} were simulated from standard multivariate normal distributions. Population correlation matrices for individual-level and cluster-level covariates were based on an analysis of a subset of the variables from the Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K)[1].[1] The coefficients β_m, π_n, and s_{mj} were defined based on condition manipulations discussed later. The cluster effects s_{0j} were drawn from a standard normal distribution, and the vector of residuals $r_{ij}$was drawn from a logistic distribution with mean of zero and variance equal to $\pi^2/3$. The presence of the residual r_{ij} in Equation (8) enables Equation (9) to create treated and untreated groups whose estimated probability distributions overlap. With this population treatment assignment model, the degree of common support for treated units will vary randomly with each sample drawn. Because in this study we examined both methods that discard treated units with no common support (i.e., one-to-many greedy matching within calipers) and methods that do not (i.e., PS weighting), we made the ATT constant across the entire range of propensity scores so estimates of ATT obtained with different PS methods can be compared to the same population ATT.

[1]The population parameters used can be obtained by contacting the first author.

The population models for the treated and untreated outcomes were:

$$\begin{aligned} y^c_{ij}&=\gamma_0+\sum_{m=1}^{5}\eta_m \mathrm{X}_{mij}+\sum_{n=1}^{5}\kappa_n \mathrm{Z}_{nj}+u_{0j}+\varepsilon_{ij},\\ y^t_{ij}&=\gamma_0+\gamma_1 T_{ij}+\sum_{m=1}^{5}\eta_m \mathrm{X}_{mij}+\sum_{n=1}^{5}\kappa_n \mathrm{Z}_{nj}+u_{0j}\\ &\quad+u_{1j}T_{ij}+\varepsilon_{ij},\\ y_{ij}&=(1-T_{ij})y^c_{ij}+T_{ij}y^t_{ij}, \end{aligned} \quad (10)$$

Where y^c_{ij} are the potential outcomes in the control condition, y^t_{ij} are the potential outcomes in the treated condition, and y_{ij} are the observed outcomes. η_m are the fixed effects of five individual-level covariates X_{mij}, κ_n are the fixed effects of cluster-level covariates Z_{nj}, and the other terms are defined as in Equations (6) and (7). To simplify the simulation, the population potential outcome models do not include random slopes of individual-level covariates. The population intercept γ_0 was set to 110 and the treatment effect γ_1 was set to –10, while the population variance of u_{0j} and u_{1j} were set to 24 and 9, respectively, and their covariance was set to zero. These values are based on the ECLS-K, and are similar to the estimates reported later in the applied example. Although the treatment effect varied across clusters, it was defined as independent of covariates, so that the ATT would be the same for all simulated datasets despite random variation of overlap of covariate distributions between treated and untreated groups. The population within-cluster variance of the outcome (i.e., the variance of ε_{ij}) was set based on the desired levels of intra-class correlation (ICC), which are described later.

We used an extension of the McKelvey and Zavoina pseudo-R^2 (McKelvey & Zavoina, 1975) for multilevel logistic and probit regression models (Snijders & Bosker, 1999) to set the amount of variability in the treatment assignment related to observed individual-level and cluster-level covariates and manipulate the strength of selection bias due to cluster effects. The McKelvey and Zavoina pseudo-$R^2=\sigma^2_F/(\sigma^2_F+\sigma^2_R)$ is the proportion of the variance of a continuous variable underlying the categorical outcome that is accounted for by the covariates in the model. Therefore, σ^2_F is the variance explained covariates, and $\sigma^2_R=\pi^2/3$ is the variance of a logistic distribution with a scale parameter equal to 1. Snijders and Bosker (1999) proposed a multilevel pseudo-$R^2=\sigma^2_F/(\sigma^2_F+\tau^2_0+\sigma^2_R)$ to quantify the proportion of variance explained by covariates in a multilevel logistic regression, where $\tau^2_0$is the variance of the random cluster effect s_{0j}. In the simulation of the treatment assignment, we first defined the regression coefficients β_m and $\pi_n$in Equation (8) such that, in the absence of the random cluster effect s_{0j}, the population pseudo-R^2 is 0.30, and the cluster-level covariates together account for as much variability in the treatment assignment as the individual-level covariates. Then, we

manipulated four levels of the variance τ_0^2 of the random cluster effect such that the population multilevel pseudo-R^2 was 0.10, 0.20, 0.25, or 0.30. In other words, the presence of the random cluster effect corresponds to a reduction of the pseudo-R^2 of 0.20, 0.10, 0.05, and 0.00 (i.e., no random cluster effect). In this study, we refer to this condition as the size of the cluster effect on treatment assignment, and identify the manipulated levels by the amount of reduction in pseudo-R^2 caused by the random cluster effect.

We manipulated whether the effects of individual-level covariates on treatment assignment varied across clusters by creating two conditions: (1) Same effects of individual-level covariates on the probability of treatment assignment for all clusters, which was accomplished by setting the variances of all s_{mj} in Equation (8) to zero; (2) Effects of individual-level covariates on the probability of treatment assignment vary randomly between clusters, which was accomplished by sampling s_{mj} from normal distributions with standard deviations of β_m. This implies that we simulated a substantial amount of variability in the effects of individual-level covariates between clusters, because the between-cluster standard deviations of the coefficients were as large as the coefficients themselves and therefore the coefficient of variation (cv) was 1.0.

In the data simulation, we also manipulated the ICC of the outcome. While keeping the population between-cluster variance of the outcome constant, we selected two levels of the within-cluster residual variance such that the ICC of the outcome was either 0.25 or 0.05. We also manipulated the cluster sizes but kept the number of clusters constant at 100. We simulated cluster sizes of 5, 10, and 20, which are typical of large educational datasets obtained through multistage sampling. For example, the 2007–2008 Schools and Staffing Survey (SASS) sampled schools and then teachers within schools such that the maximum number of teachers per school was 20 (National Center for Education Statistics, 2011).

The data simulation design had a total of 48 conditions, resulting of four sizes of the cluster effect on treatment assignment (i.e., reduction of the pseudo-R^2 of 0.20, 0.10, 0.05, and 0.00), two levels of between-cluster variability of individual-level covariate effects on treatment assignment (i.e., cv = 0.0 and 1.0), two levels of ICC of the outcome (i.e., 0.25 or 0.05), and three cluster sizes (i.e., 5, 10, and 20). One thousand datasets were simulated per condition. The multilevel datasets were simulated with the R 2.13.1 statistical software (R Development Core Team, 2012).

Across simulated datasets, the proportion treated varied randomly from 0.17 to 0.51, with mean of 0.30, median of 0.29, and standard deviation of 0.05. The level of covariate imbalance between treated and untreated groups were affected by the population multilevel pseudo-R^2, as explained previously, but varied randomly with each dataset. We measured the maximum observed covariate imbalance in each dataset using standardized mean differences, and found that it ranged from 0.12 to 1.37 with a mean of 0.59, median of 0.60, and standard deviation of 0.13. Given that standardized mean differences below 0.1 standard deviations can be considered to indicate adequate covariate balance (Austin, 2011), our simulation method succeeded in generating covariate imbalance that was on average 6 times larger than the recommended threshold for adequate balance. We also measured the observed level of lack of common support between PS distributions as the proportion of treated observations with propensity scores above the maximum PS of untreated observations plus 0.2 standard deviations. The proportion of lack of common support had a minimum of 0.0, maximum of 0.77, mean of 0.05, median of 0.01, and standard deviation of 0.10 across simulated datasets. Poor common support may reduce covariate balance with PS methods and the generalizability of estimated treatment effects (Crump, Hotz, Imbens, & Mitnik, 2009). In this study, we did not discard observations with poor common support. Considerations on strategies to deal with common support with multilevel data and directions for future research are presented in the Discussion and Conclusion section.

Analysis

For each simulated dataset, propensity scores were estimated with three different models. The first model was a logistic regression model including individual-level and cluster-level covariates, but ignoring random cluster effects:

$$\text{logit}\,(T_{ij} = 1) = \beta_0 + \sum_{m=1}^{5} \beta_m X_{mij} + \sum_{n=1}^{5} \pi_n Z_{nj} \qquad (11)$$

The second model was a fixed-effects logistic regression model including individual-level covariates as well as 99 dummy-coded indicators of cluster membership, $g_1 ... g_{99}$:

$$\text{logit}(T_{ij} = 1) = \beta_0 + \sum_{m=1}^{5} \beta_m X_{mij} + \sum_{j=1}^{99} \delta_j g_j. \qquad (12)$$

The third model was a multilevel logistic regression model with a normally distributed random intercept s_{0j} with variance ς and including both individual-level and cluster-level covariates:

$$\text{logit}\,(T_{ij} = 1) = \beta_0 + \sum_{m=1}^{5} \beta_m X_{mij} + \sum_{n=1}^{5} \pi_n Z_{nj} + s_{0j}$$
$$s_{0j} \sim N(0, \varsigma) \qquad (13)$$

It is important to notice that the two models that include cluster effects did not allow for the effects of covariates to vary between clusters: there were no covariate-by-cluster interactions in the fixed-effects logistic regression model and no random slopes in the multilevel logistic regression model. Given that we simulated data with and without variation of covariate effects between clusters, we were able to examine whether ignoring this variation had undesirable consequences.

We estimated the logistic regression models with the *glm* function of the base package, and the multilevel logistic regression models with the *glmer* function of the *lme4* package (Bates, et al., 2011) of the R statistical software.

We used four PS methods implemented across clusters (i.e., with the entire sample) to reduce selection bias in the treatment effect estimates: one-to-many greedy matching, stratification with 10 strata, OFM, and weighting. We implemented one-to-many greedy matching with the *MatchIt* package of R (Ho, Imai, King, & Stuart, 2011) by requesting all matches with replacement within a caliper of 0.25 standard deviations of the propensity scores of treated individuals. This caliper size is often used by applied researchers (Thoemmes & Kim, 2011). Treated observations without any untreated observation within their calipers were dropped. We conducted PS stratification by dividing the sample into 10 strata based on equal ranges of the PS distribution of the treated. Lunceford and Davidian (2004) have shown that using 10 strata leads to greater reduction of bias than the more commonly used 5 strata. OFM was implemented with the *MatchIt* package (Ho, et al., 2011), which uses algorithms implemented in the *Optmatch* package (Hansen, 2007). For both PS stratification and OFM, any stratum containing only treated cases was dropped. Weights were obtained with Equation (3) for one-to-many greedy matching, Equation (4) for stratification and OFM, and Equation (5) for weighting.

We estimated the ATT with a multilevel model for the outcome that allowed random intercepts and treatment effects, as shown in Equation (7). Parameter estimation was performed with the MPML estimation method (Asparouhov, 2006) implemented in Mplus 7.0 (Muthén & Muthén, 2013) using weights from PS methods as sampling weights. We used three different scaling methods for the weights: normalized, cluster-normalized, and effective weights by utilizing the "WTSCALE = " option. To obtain a measure of initial levels of bias, we also estimated treatment effects with the multilevel model in Equation (7) without using weights; we refer to this model as the "baseline" model throughout the Results section.

In summary, each simulated dataset was analyzed with 37 different approaches, resulting from the combination of three PS models (i.e., logistic regression, fixed-effects logistic regression, and multilevel logistic regression), four PS methods (i.e., one-to-many greedy matching, stratification with 10 strata, OFM, and PS weighting), and three weight-scaling methods (i.e., normalized, cluster-normalized, and effective), plus a baseline model. Consequently, the total simulation design had 48 data simulation conditions analyzed with 37 approaches, resulting in 1,776 conditions.

We compared the data analysis approaches across data simulation conditions with respect to the relative biases of the parameter estimate of the ATT and its standard error estimate. In addition, we examined bias in the estimate of the between-cluster variance of ATT. The relative parameter bias was calculated using $B(\hat{\theta}) = (\overline{\hat{\theta}} - \theta)/\theta$, where $\overline{\hat{\theta}}$ is the mean of the estimates across all replications of a condition and $\theta$is the population parameter. The relative standard error bias was calculated using $B(S_{\hat{\theta}}) = [\overline{S_{\hat{\theta}}} - SD(\hat{\theta})]/SD(\hat{\theta})$, where $\overline{S_{\hat{\theta}}}$ is the mean of the estimated standard errors of $\hat{\theta}$, and $SD(\hat{\theta})$is the empirical standard error, which is the standard deviation of $\hat{\theta}$ across all iterations of a condition. Following recommendations by Hoogland and Boomsma (1998) and Bandalos and Leite (2013), we considered parameter bias and standard error bias acceptable if their absolute values were smaller than 0.05 and 0.1, respectively.

Because our simulation design was very large, which complicates visual inspection of effects, we used mixed-design analysis of variance (ANOVA) and the generalized eta squared (G η^2) (Olejnik & Algina, 2003) measure of effect size to sort manipulated conditions with respect to the magnitude of their effect on relative parameter bias and relative standard error bias (see Bandalos & Leite, 2013, for a discussion of this methodology). In these mixed-design ANOVAs, the between-dataset factors were cluster effect on treatment assignment, between-cluster variability of individual-level covariate effects on treatment assignment, ICC, and cluster sizes. The within-dataset factors were PS models, PS methods, and weight-scaling methods. The dependent variables were the relative bias estimates defined above. Outcomes of conditions with the baseline model were not included in these analyses. The mixed-design ANOVA models included all interactions between the manipulated factors.

RESULTS

For the simulated data, estimates of ATT obtained with the multilevel model without use of propensity score weights were positively biased in all conditions, with relative bias that ranged from 0.096 to 0.144, with a mean of 0.113 and standard deviation of 0.011. The nonrandom treatment assignment mechanism did not affect the standard errors of ATT obtained with the multilevel model for the outcome without propensity score weights, given that the relative bias of standard errors of ATT ranged from –0.048 to 0.010, with mean of –0.015 and standard deviation of 0.013.

The largest effect of a manipulated condition on relative bias of ATT estimates was of PS method (G $\eta^2 = 0.020$). Matching produced the lowest level of relative bias ($B(\gamma_1) = 0.023$), followed by weighting ($B(\gamma_1) = 0.029$), OFM ($B(\gamma_1) = 0.042$) and stratification ($B(\gamma_1) = 0.046$). The effect of cluster size (G $\eta^2 = 0.019$) resulted in decreases in relative bias as the cluster size increased (i.e, $B(\gamma_1) = 0.047$, 0.034, and 0.024 for cluster sizes of 5, 10, and 20, respectively). We summarized the relative bias across PS methods and cluster sizes in Table 1.

TABLE 1
Descriptive Statistics of Relative Bias of ATT Estimates Across Propensity Score Methods and Cluster Sizes

PS Method	Cluster Size	Mean	*SD*	Min	Max
Matching	5	0.020	0.022	—0.040	0.063
	10	0.026	0.014	—0.004	0.062
	20	0.022	0.013	—0.003	0.058
OFM	5	0.059	0.018	0.018	0.107
	10	0.040	0.014	0.010	0.070
	20	0.026	0.013	0.002	0.059
Stratification	5	0.065	0.015	0.036	0.103
	10	0.044	0.013	0.017	0.076
	20	0.031	0.013	0.006	0.061
Weighting	5	0.043	0.010	0.023	0.063
	10	0.028	0.010	0.005	0.060
	20	0.018	0.013	−0.003	0.056

Note. SD = Standard deviation; Min = Minimum; Max = Maximum; OFM = Optimal full matching.

We found that the estimates of the ATT obtained with MPML estimation were not substantially affected by the weighting scaling method used (G $\eta^2 = 0.009$). Also, there were no interactions between the other manipulated factors and the scaling method with G η^2 greater than 0.002. Collapsing across the other manipulated conditions, the relative bias of the ATT with normalized, cluster-normalized, and effective weights was 0.026, 0.038, and 0.040, respectively.

There were no substantial differences between PS models with respect to bias of ATT estimates (G $\eta^2 = 0.009$). The relative bias of ATT estimates were similar between conditions with propensity scores estimated with the multilevel logistic regression model ($B(\gamma_1) = 0.028$), the fixed-effects logistic regression model ($B(\gamma_1) = 0.035$), and the logistic regression model ignoring cluster effects($B(\gamma_1) = 0.042$). Although differences were small, we observed that the multilevel logistic regression model consistently provided the lowest level of bias, followed by the fixed-effects logistic regression model and the logistic regression model ignoring clustering effects. The manipulation of the multilevel pseudo-R^2 to create different levels of cluster effect on treatment assignment and the different levels of ICC of the outcome had little impact (i.e., G $\eta^2 < 0.005$) on the relative bias of ATT.

The mixed-design ANOVA showed that none of the manipulated factors had a substantial effect on the standard errors of the ATT, because no G η^2 was above 0.001. Inspection of the relative bias of standard errors indicated that model estimation with multilevel pseudo-maximum likelihood estimation with weights from PS methods produced acceptable standard error in all conditions, except in six cells of the simulation design. The range of relative bias of standard errors was from –0.090 to 0.613, with mean of 0.004 and standard deviation of 0.032. Further investigation indicated that there were some outlier standard error estimates that were responsible for the increased bias of six conditions. In the distribution of the deviations of estimated standard errors from empirical standard errors, the minimum was –0.580 and the 99.9% quantile was 0.366, but the maximum was 291.590. We found that conditions with the smallest cluster size (i.e., 5), the largest size of cluster effect (i.e., 0.2), and that used one-to-many matching sometimes produced these outlier standard errors because matching was the only PS method used that reduced sample sizes. After trimming 0.1% of the right side of the distribution of deviations of estimated standard errors from empirical standard errors, the relative bias of standard errors ranged from –0.117 to 0.061, with mean of 0.001 and standard deviation of 0.024.

The relative biases of the between-cluster variance of the ATT were unacceptably large in most conditions. The effects of the conditions manipulated on the relative bias of the between-cluster variance of the ATT were complex, as indicated by the presence of two-, three-, and four-way interactions with sizable G η^2 (see Table 2). It is readily noticeable that the weight-scaling method had the strongest effect on the relative bias of the between-cluster variance of the ATT (G $\eta^2 = 0.397$), and most of the interaction effects involved the weighting scale method. Tables 3, 4, and 5 allow inspection of these effects. The relative biases of the between-cluster variance of the ATT were smaller when effective weights were used. With cluster-normalized and normalized weights, the relative bias was always positive, indicating an overestimation of the between-cluster variance of the ATT. Also, these two scaling methods resulted in bias that reduced as cluster size and ICC increased. However, with effective weights, the

TABLE 2
Effects of Manipulated Conditions on the Relative Bias of the Between-Cluster Variance of the ATT

Effect	G η^2
Cluster size (A)	0.103
ICC (B)	0.244
Cluster effect (C)	—
PS model (D)	0.047
PS method (E)	0.024
Scaling method (F)	0.397
A × B	0.049
B × D	0.014
C × D	0.015
B × E	0.01
A × F	0.069
B × F	0.135
D × F	0.026
E × F	0.119
A × B × F	0.012
A × E × F	0.048
B × E × F	0.035
A × B × E × F	0.013

Note. ICC = Intra-class correlation; PS = Propensity score; G η^2 = Generalized eta squared. Only G η^2 greater than 0.01 are displayed.

TABLE 3
Relative Bias of Between-Cluster Variances From Analyses Using Effective Weights

PS Method	PS Model	ICC = 0.1			ICC = 0.3		
		$n = 5$	$n = 10$	$n = 20$	$n = 5$	$n = 10$	$n = 20$
Matching	Multilevel	0.055	−0.109	−0.074	−0.226	−0.134	**–0.032**
	Fixed	**0.006**	−0.090	−0.061	−0.243	−0.132	**–0.025**
	Logistic	0.098	−0.160	−0.151	−0.162	−0.134	−0.071
OFM	Multilevel	−0.099	−0.252	−0.259	−0.272	−0.160	−0.112
	Fixed	−0.148	−0.234	−0.229	−0.265	−0.145	−0.092
	Logistic	−0.116	−0.259	−0.242	−0.235	−0.157	−0.121
Stratification	Multilevel	−0.182	−0.224	**–0.048**	−0.313	−0.117	**–0.020**
	Fixed	−0.315	−0.242	**0.003**	−0.404	−0.127	**0.006**
	Logistic	−0.066	−0.151	−0.086	−0.162	−0.074	**–0.040**
Weighting	Multilevel	0.127	0.070	0.118	−0.113	**0.030**	**0.047**
	Fixed	0.116	**0.034**	0.051	−0.085	**0.010**	**0.021**
	Logistic	0.082	**0.005**	0.065	−0.152	**–0.021**	**0.022**

Note. OFM = Optimal full matching; PS = Propensity score; ICC = Intra-class correlation. *n* indicates cluster size. Bold numbers are acceptable levels of relative bias.

bias was positive for some conditions but negative for others, and did not consistently decrease as cluster size increased.

In sharp contrast to the results obtained with multilevel pseudo-maximum likelihood estimation and three PS weight-scaling methods, the relative bias of the between-cluster variance of the ATT with the baseline model (i.e., a multilevel model for the outcome without weights and estimated with robust maximum likelihood estimation) was acceptable for all conditions except one condition with cluster size equal to five, ICC = 0.1, size of cluster effect (reduction of the pseudo-R^2) = 0.2, and with cv = 1.0. With the baseline model, the relative biases ranged from –0.050 to 0.120, with mean of –0.002 and standard deviation of 0.038.

Summary of Results

We presented and evaluated an approach for the implementation of four PS methods with multilevel data that involved the creation of weights, scaling of weights, and estimation of weighted multilevel models with MPML estimation (Asparouhov, 2006). We found that this approach works well for obtaining unbiased estimates of the ATT and standard errors with all PS methods evaluated, and therefore we recommend it to researchers interested in applying PS methods to multilevel data. Also, the choice of weight-scaling method did not matter for estimation of the ATT and its standard error, but had a large effect on estimating the between-cluster variance of the ATT. These estimates were positively biased in the majority of conditions examined, which agrees with Asparouhov's (2006) findings, but were substantially less biased when effective weights were used, which agrees with Stapleton's (2002) findings. However, the estimates of between-cluster variances obtained using effective weights only had acceptable levels of bias in a few conditions, particularly when cluster size was 20 and ICC was 0.3. Although the multilevel model with no weights provided acceptable

TABLE 4
Relative Bias of Between-Cluster Variances From Analyses Using Cluster-Normalized Weights

PS Method	PS Model	ICC = 0.1			ICC = 0.3		
		$n = 5$	$n = 10$	$n = 20$	$n = 5$	$n = 10$	$n = 20$
Matching	Multilevel	4.465	3.233	1.850	1.247	0.927	0.549
	Fixed	4.077	3.023	1.795	1.146	0.886	0.543
	Logistic	4.287	2.454	1.109	1.257	0.727	0.345
OFM	Multilevel	2.375	1.914	1.358	0.664	0.565	0.411
	Fixed	2.035	1.829	1.360	0.600	0.549	0.422
	Logistic	1.591	1.084	0.656	0.440	0.307	0.202
Stratification	Multilevel	1.866	1.499	1.022	0.534	0.461	0.329
	Fixed	2.108	1.645	1.074	0.631	0.502	0.351
	Logistic	1.004	0.611	0.350	0.277	0.177	0.116
Weighting	Multilevel	1.945	1.636	1.199	0.568	0.502	0.372
	Fixed	2.537	1.889	1.299	0.752	0.570	0.407
	Logistic	1.452	1.063	0.692	0.393	0.313	0.218

Note. OFM = Optimal full matching; PS = Propensity score; ICC = Intra-class correlation. *n* indicates cluster size.

TABLE 5
Relative Bias of Between-Cluster Variances From Analyses Using Normalized Weights

		ICC = 0.1			ICC = 0.3		
PS Method	PS Model	$n = 5$	$n = 10$	$n = 20$	$n = 5$	$n = 10$	$n = 20$
Matching	Multilevel	1.265	1.156	0.861	0.238	0.297	0.229
	Fixed	1.141	1.130	0.837	0.181	0.264	0.219
	Logistic	1.021	0.729	0.473	0.214	0.194	0.126
OFM	Multilevel	2.889	1.707	0.989	0.848	0.517	0.292
	Fixed	3.388	1.793	0.987	1.002	0.534	0.292
	Logistic	1.616	0.908	0.480	0.464	0.252	0.141
Stratification	Multilevel	2.444	1.365	0.737	0.749	0.425	0.220
	Fixed	2.932	1.440	0.736	0.920	0.441	0.223
	Logistic	1.076	0.497	0.224	0.305	0.133	0.064
Weighting	Multilevel	2.495	1.523	0.879	0.745	0.466	0.266
	Fixed	3.420	1.781	0.939	1.003	0.525	0.281
	Logistic	1.697	0.913	0.477	0.482	0.266	0.151

Note. OFM = Optimal full matching; PS = Propensity score; ICC = Intra-class correlation. *n* indicates cluster size.

bias of the between-cluster variance of the ATT across all conditions and previous research using the MPML estimator obtained biased variance estimates (Asparouhov, 2006), we cannot attribute the bias of the between-cluster variance solely to the use of MPML estimation. Therefore, it is not possible to ascertain from the current results the extent that the bias in between-cluster variance is due to MPML estimation with weights from PS methods, the selection bias, or an interaction between them.

It is important to note that the choice of PS model had little effect on the estimates, which provides support to Thoemmes and West's (2011) finding that the fixed-effects model and multilevel model work equally well. However, we also found that the logistic regression model ignoring cluster effects provided adequate results. The reason is that when a multilevel model is used to estimate a treatment effect, such as the model presented in Equations (6) and (7), some degree of selection bias is removed both with the PS model and the outcome model. This agrees with Li et al. (2013), who found that in studies where both the treatment assignment mechanism and outcome of interest are multilevel in nature, selection bias due to clusters will be reduced if the multilevel structure of the design is considered in either the PS model or the outcome model. Therefore, if neither fixed- nor random-cluster effects had been included in the outcome model, or a proxy to cluster effects such as the proportion treated, we would expect that the estimates obtained from conditions where the logistic regression model ignoring cluster effects was used would have higher levels of bias.

With the PS method implementation used, we did not find any consequence of ignoring between-cluster variability of individual-level covariate effects on treatment assignment on the bias of ATT estimates and standard errors, even though we simulated a large amount of variability by setting the between-cluster standard deviations equal to the means of the covariate effects. This finding is of practical importance to applied researchers who want to combine PS methods with multilevel modeling, because accounting for between-cluster variability of covariate effects in the PS model (e.g., by adding random slopes of covariates to a multilevel logistic regression model), greatly complicates model specification and estimation when the number of covariates is large.

To provide additional insight on this finding, we undertook a secondary study and simulated data according to Equations (8–10), but varying the size of the standard deviations of the random slopes of the five individual-level predictors of treatment assignment [(see Equation (8)]. We defined the 10 sizes of standard deviations of random slopes as proportions of their means from 0.1 to 1.0 in increments of 0.1. We simulated 100 datasets with 100 clusters and cluster sizes of 5, 10, or 20 with each of the 10 sizes of standard deviations, resulting in 30 conditions. For each dataset, we estimated propensity scores using the multilevel logistic regression model in Equation (13), which does not include random slopes of individual-level covariates, and calculated the correlation between the estimated propensity scores and the true probabilities of treatment. These correlations are presented in Figure 1. It can be observed that the correlation between estimated propensity scores and true probabilities decreases as the between-cluster standard deviations of the slopes of individual-level covariates increase, but the reduction is not substantial. Figure 1 also shows that these correlations decrease as cluster sizes decrease. Therefore, we conclude that there is evidence that the PS method implementation using a PS model without random slopes is robust to the simulated levels of random variability in covariate effects on treatment assignment across clusters, but the evidence we found is limited to a small number of covariates (i.e., five) This finding is supported by Kelcey (2011b), who simulated treatment assignment with a multilevel model with random slopes and cross-level interactions, and found that a PS model ignoring the random slopes performed similarly to a model including

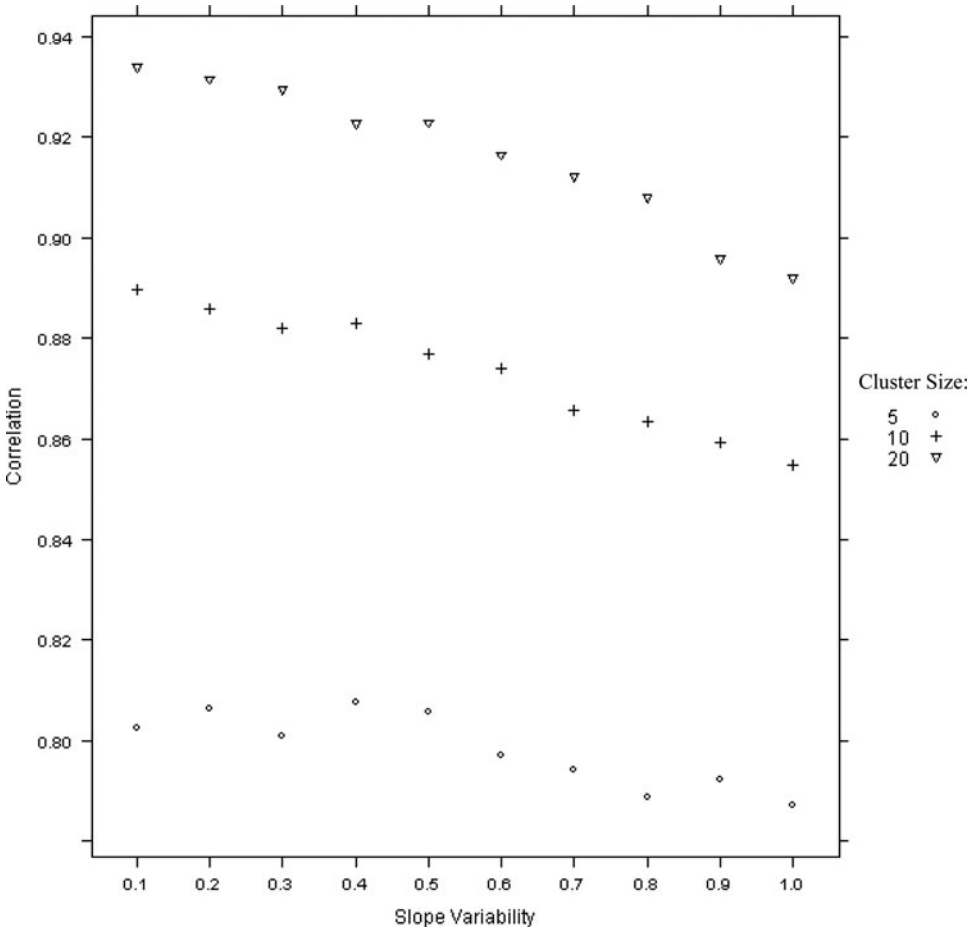

FIGURE 1 Correlation of estimated propensity score and true probability of treatment assignment (*Y* axis) as a function of the between-cluster variability of covariate effects on treatment assignment (*X* axis) and cluster size (lines).

random slopes with respect of bias, as long as there were no omitted covariates. With one omitted school-level and/or student-level covariate, he found that the PS model with random slopes resulted in lower bias than the PS model ignoring random slopes. Furthermore, Kelcey (2009) found that when treatment assignment has a complex multilevel structure depending on individual- and cluster-level covariates with cross-level interactions, a multilevel PS model, which accounts for this complex structure only performs slightly better than simpler multilevel models, and selection of true confounders for the PS model is more important than the choice of PS model.

DEMONSTRATION OF ATT ESTIMATION WITH MULTILEVEL MODELING AND PROPENSITY SCORE WEIGHTS

In this section, we provide a demonstration of how to combine PS weighting with multilevel modeling to estimate the ATT of children receiving special education services in third grade on reading achievement, using multilevel survey data from the ECLS-K. We compare estimates obtained with normalized, cluster-normalized, and effective weights. This example is similar to part of the study published by Morgan, Frisco, Farkas, and Hibel (2010), because we controlled for the same set of covariates, but with a PS analysis strategy that takes into account the nested structure of the ECLS-K. We abstain from discussing the implications of special education services, but details can be found in Morgan, Frisco, Farkas, and Hibel (2010).

The implementation of PS weighting with multilevel models to estimate the ATT from multilevel survey data involves the following major steps: (1) Examine the sample available, the size of the treated/untreated groups, the number of clusters, the cluster sizes, and the proportion treated per cluster. This information is helpful for deciding whether to pool cluster-specific ATTs or estimate a marginal ATT from the entire sample. It is also important to identify whether the SUTVA assumption is tenable and how the sample was selected (e.g., stratified cluster sampling); (2) Identify and prepare data on individual-level and cluster-level covariates related to both treatment assignment and the outcome, which

are true confounders, as well as outcome proxies (Kelcey, 2011a). Variables strongly related to only the outcome should also be included because they increase the power to test the ATT (Brookhart et al., 2006; Cuong, 2013). Data preparation in this step includes deciding on how to handle missing data, such as by obtaining multiple imputations of the data and averaging the propensity scores obtained from the imputed datasets (Mitra & Reiter, 2012); (3) Estimate propensity scores accounting for the multilevel structure of the data and the sampling design (if applicable), and examine common support (i.e., the overlap assumption) both within and across clusters. In this step, the researcher may decide on whether to pool cluster-specific ATTs or estimate a marginal ATT; (4) Obtain weights for the PS method(s) of choice and evaluate covariate balance (i.e., the strong ignorability of treatment assignment assumption). If sampling weights are informative (Pfeffermann, et al., 1998), they should be multiplied by the weights from PS methods (Hahs-Vaughn & Onwuegbuzie, 2006). If adequate covariate balance is not achieved, the researcher should revisit the estimation of PS scores (e.g., add interactions in the PS model); (5) Select and fit a multilevel model to estimate the ATT including weights, and potentially other covariates of interest; and (6) Perform a sensitivity analysis to evaluate how strong the effect of an omitted covariate would have to be for the significance of the treatment effect to change (Rosenbaum, 2010).

For the first step, we obtained a sample from the ECLS-K with 7,783 third grade students nested in 1,454 schools. The minimum, maximum, mean, and standard deviation of the number of students per school were 1, 27, 5.353, and 4.939, respectively. In this sample, 6.20% of students received special education services. From the schools in the sample, 1,139 had a sample that did not contain any treated students and 35 only contained treated students. The sample of the ECLS-K was obtained with stratified multistage sampling, but it is only representative of kindergarten and first grade students and not third grade students (National Center for Education Statistics, 2010). For this reason, and also because of missing strata indicators for third grade that would reduce substantially the number of treated observations, we decided not to use sampling weights and strata indicators in this demonstration.

For the second step, we first obtained from Morgan et al. (2010) a list of all 27 individual-level covariates and five cluster-level covariates that they used to model the probability of receipt of special education. To select these covariates, Morgan et al. used theory and prior empirical research to identify both background characteristics and academic factors that increase a child's possibility of being identified as disabled and predict his or her placement into special education services. Similarly to Morgan et al. (2010), we handled missing data with listwise deletion.

For the third step, we estimated propensity scores using a multilevel logistic regression model with 32 covariates (i.e., 27 individual level and 5 cluster level) and a random intercept, but without random slopes of individual-level covariates. We evaluated common support for the entire data by using histograms to compare the distribution of the treated and the untreated, and found it to be adequate. Within-cluster common support is not achievable with these data because of the small cluster sizes and large proportion of schools with only untreated students, and therefore we chose to estimate a marginal ATT.

For the fourth step, we implemented one-to-many matching within a 0.25 caliper, stratification using 10 strata, OFM, and PS weighting. One-to-many matching, stratification and OFM were performed with the *MatchIt* package of R (Ho, et al., 2011). Weights for each PS method were obtained according to Equations (3–5). To evaluate balance, we used the *survey* package of R (Lumley, 2010) to obtain weighted means M_{tc} and M_{uc} for each covariate c for treated and untreated groups, respectively, and the weighted standard deviations SD_c for the entire sample. We estimated the weighted standardized difference (d_c) between the groups as $d_c = (M_{uc} - M_{tc})/SD_c$. We considered covariate balance adequate if $|d_c| \leq 0.1$ for all covariates. We achieved adequate balance with PS weighting, but d_c was larger than 0.1 for 5, 4, and 11 covariates with one-to-many matching, stratification, and OFM, respectively. The maximum absolute values of d_c were .295, .204, and .375 for one-to-k matching, stratification, and OFM, respectively. These results demonstrate how different PS methods using the same PS vector can produce different levels of covariate balance. We proceeded to the next step with PS weighting because it was the only PS method that produced adequate covariate balance, and therefore for this method we have evidence that strong ignorability holds.

For the fifth step, we estimated the ATT of special education services on reading achievement in third grade using a multilevel model estimated with the MPML estimation method in Mplus 7.0 (Muthén & Muthén, 2013) using PS weights as sampling weights. We used normalized, cluster-normalized, and effective scaling of weights and compared results across weight-scaling methods. We also fit the model without weights using robust maximum likelihood estimation. We fit the following multilevel model:

$$\begin{aligned} y_{ij} &= \gamma_0 + \gamma_1 T_{ij} + \gamma_2 \left(X_{ij} - \bar{X}_t\right) + \gamma_3 \left(X_{ij} - \bar{X}_t\right) T_{ij} \\ &\quad + u_{0j} + u_{1j} T_{ij} + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim N(0, \sigma^2), u_{0j}, u_{1j} \sim N(0, \Phi) \end{aligned} \tag{14}$$

where y_{ij} is the IRT-scaled measure of reading achievement in third grade for child i in school j, T_{ij} is a dummy-indicator of receiving special education services, u_{0j} is the random intercept of school j, and u_{1j} is the random effect of receiving special education services in school j. In Equation (11), X_{ij} is the average of the child's IRT-scaled reading scores measured in the fall and spring of kindergarten and spring of first grade, $\bar{X}_t$ is the overall mean of X_{ij} for the treated. The inclusion of an important covariate (i.e., the average of

TABLE 6
Multilevel Modeling Results for the Effect of Children Receiving Special Education Services in Third Grade on Reading Achievement

Fixed Effects	No Adjustment	Normalized Weights	Cluster-Normalized Weights	Effective Weights
Intercept	118.304**(0.481)	113.458**(0.836)	112.845**(0.515)	113.314**(0.516)
Special education services (treatment)	−14.532**(1.029)	−8.429**(1.161)	−8.448**(1.084)	−8.740**(1.068)
K/first average IRT reading (covariate)	1.074**(0.024)	1.752**(0.107)	1.434**(0.042)	1.393**(0.047)
Treatment × covariate interaction	0.650* (0.148)	−0.027 (0.183)	0.258 (0.149)	0.329* (0.156)
Random effects				
Variance of intercept	71.652	19.578	188.701	83.888
Variance of effect of special education services	0.512	21.376	302.503	15.705
Covariance between intercept and effect of special education services	−2.659	7.663	−100.249	11.549
Within-school residual variance	326.418	370.413	195.866	320.503

Note. IRT = Item response theory. Standard errors in parentheses. $^{**}p < .001$. $^{*}p < .05$. $df = 1{,}452$.

previous reading scores) in the model makes it doubly robust (Li et al., 2013), because the results become less vulnerable to misspecifications of both the propensity score and outcome models. Also, the inclusion in the model of a covariate that is strongly related to the outcome will increase power to test the treatment effect. We included the interaction term because Schafer and Kang (2008) have shown that γ_{1ij} will only be equal to the ATT in models including covariates if the covariates are centered around the mean of the covariates for the treated and the treatment by covariate interaction is included in the model.

For the sixth step, we did not perform a sensitivity analysis because its implementation for multilevel models is outside of the scope of this paper, but a related method can be found in Brumbach, Hernán, Haneuse, and Robins (2004).

Demonstration Results

In Table 6, we present the estimates obtained with the multilevel model both without weights and with normalized, cluster-normalized, and effective PS weights. As expected from the simulation study results, the estimates of the fixed effects were similar across weight-scaling methods, but the random effect estimates differed dramatically. More specifically, the between-cluster variance estimate from the analysis with cluster-normalized weights was much larger than with normalized and effective weights, while the estimate from the unweighted analysis was the smallest. On the other hand, the ATT estimates with all three PS weights as well as without weights indicate that children who received special education services in third grade showed lower reading skills compared to those children who did not receive such services but had similar distributions of covariates. The ATT estimates were lower with PS weighting than without weighting, but were still statistically significant. These results agree with Morgan et al.'s (2010) findings using stratification and kernel matching, and indicate a shortcoming of the services provided for these students. The discussion of these negative results is presented in detail by Morgan et al.'s (2010). There were only small differences between the PS weight-scaling methods in the estimate of the main effect of the covariate, but there were larger differences in the estimate of the interaction between treatment and the covariate. In particular, no weighting and effective weighting resulted in statistically significant positive interactions, while estimates obtained with normalized and cluster-normalized weighting were nonsignificant. For these data, a positive main effect of the covariate combined with a significant positive interaction indicates that the slope of the regression of third grade reading scores on the average of previous reading scores is steeper for individuals who received special education services than for those who did not. Because the simulation conducted in this study did not include an interaction effect, the mechanisms by which the three weight-scaling methods may affect the estimate of the interaction effect differently deserve consideration in future research.

The main limitation of this analysis is the possibility that important confounders have been omitted, and therefore results should be considered with care. An improvement over PS methods to evaluate the effect of special education services would be to obtain a sample where some individuals were given services based on meeting a specific cutoff on a quantitative measure. If the dataset contains scores on this measure used for treatment assignment and outcomes for both treated and untreated, a regression discontinuity analysis (Cook & Steiner, 2009; Imbens & Lemieux, 2008) can be performed, which estimates a treatment effect around the cutoff with as strong internal validity as an experimental design, but with a loss of efficiency (Schochet, 2009).

DISCUSSION AND CONCLUSION

Given the results of this study, we argue that it is possible to eliminate most selection bias in the estimation

of the ATT and its standard error by using multilevel models with weights from different PS methods and MPML estimation. This strategy was evaluated for obtaining the marginal ATT rather than the pooled ATT, because cluster sizes found in social and educational research (e.g., 5, 10, 20) are frequently too small to obtain adequate common support and covariate balance within clusters. However, this strategy should not be used to obtain between-cluster variances of the treatment effect. Because PS methods to obtain the marginal ATT do not guarantee unbiased cluster-specific ATT estimates, we expected that there would be some bias in the between-cluster variance of the ATT across all conditions. Furthermore, with small cluster sizes such as the ones we simulated, previous research has found that the MPML estimator produces positively biased estimates of between-cluster variances (Asparouhov, 2006). However, we found unbiased estimates of between-cluster variances in the baseline condition, which used robust maximum likelihood estimation without weights. This result may be due to the levels of selection bias we simulated (i.e., multilevel pseudo-R^2 of 0.10, 0.20, 0.25, or 0.30), which may have not been sufficient to create substantial bias of the between-cluster variance of the ATT estimates, even though the ATT estimates were biased. Therefore, generalization of our results to stronger levels of selection bias is limited, and additional research is needed to examine the extent that selection bias affects variance estimates in multilevel models.

When the internal validity of a research design is threatened by selection bias (Shadish, et al., 2002), unbiased treatment effect estimates can be obtained by specifying either a correct treatment assignment model or a correct outcome model (Schafer & Kang, 2008). Because it is challenging to determine a correct specification of either the PS model or the outcome model, a "doubly robust" (Bang & Robins, 2005) procedure is recommended. With multilevel data, we found that combining a PS model that is incorrect because it omits cluster effects, with an outcome model that is incorrect because it omits important covariates, can produce adequate treatment effect estimates because the PS model and the outcome model compensate for the other model's limitations. This finding supports the idea of double-robustness discussed in the context of multilevel modeling by Li et al. (2013).

In applied studies, we recommend that researchers investigate whether the effects of covariates vary across clusters, which can be accomplished with multilevel models by comparing models with or without random slopes of covariates, or with fixed effects models by comparing models with or without covariate-by-cluster interactions. In practice, PS models tend to have a very large number of covariates and fitting a multilevel model with many random slopes or a fixed-effects model with many covariate-by-cluster interactions could become unfeasible as the number of covariates increases. Both this study and Kelcey's (2011b) study found that there is some degree of robustness of the PS model to the omission of random slopes of covariates. However, the extent of the robustness of PS models to omission of random slopes is not known, given that in this study the selection bias in the simulated data was due to only five individual-level and five cluster-level covariates with no cross-level interactions, and in Kelcey's (2011b) study there were three individual-level and three cluster-level covariates. With a large number of covariates whose effects on treatment assignment differ across clusters, this robustness is expected to degrade as the number of covariates increases, but additional research is needed on this issue.

A sensible strategy for PS model specification with multilevel data when the number of covariates is large is to apply the principle of parsimony with respect to random slopes: start with a PS method implemented with a simpler PS model (e.g., Equation (1) without random slopes), then evaluate covariate balance, and only proceed to examine PS models with random slopes and cross-level interactions if the first attempt is not successful in obtaining covariate balance. A potentially laborious but adequate follow-up would be to add one or a small set of random slopes and cross-level interactions in Equation (1) at a time, implement a PS method and evaluate covariate balance, then repeat this process until adequate covariate balance is achieved. Data mining methods (Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008) are promising alternatives to logistic regression models for estimating propensity scores because they automatically detect interaction terms, but their performance with multilevel data has not been evaluated. A challenge of using data mining methods for PS estimation is the determination of a stopping rule to prevent overfitting. Ridgeway, McCaffrey, Morral, Burgette, and Griffin (2013) have developed the *twang* package for the R statistical software that implements four stopping rules to generalized boosted regression when used to estimate propensity scores, but research comparing different stopping rules for PS estimation both within and across data mining methods is still needed.

When evaluating propensity scores estimated with multilevel data, if the distribution of propensity scores of some treated individuals does not have common support in the distribution of untreated individuals, inference about treatment effects is limited to the area of common support. Therefore, untreated individuals outside the area of common support of the treated could in theory be discarded, but while applications of PS matching usually discard individuals, applications of weighting and stratification do not (Stuart, 2010), which may impact estimates and standard errors if lack of common support results in extreme weights (Freedman & Berk, 2008; Strumer, 2010). The benefits of discarding untreated units outside of common support of the treated for estimation of the ATT have not been examined across PS methods or with multilevel research designs. For the estimation of ATE with single-level designs, Crump et al. (2009) argued that discarding untreated individuals with no common support in the treated increases the efficiency of estimates. However, with multilevel data, discarding treated individuals with no

common support has the disadvantage of reducing cluster sizes.

We only found small differences between different PS methods implemented with multilevel models, which was expected because the PS methods used have been found to work well with data that is not multilevel. Based not only on our results but also on the existing literature (Arpino & Mealli, 2011; Li, et al., 2013; Thoemmes & West, 2011), we would not expect substantial differences in the comparison of PS methods with stronger levels of selection bias due to individual and/or cluster-level variables. With stronger selection bias we would expect worse common support, which could bring complications such as extreme weights (Freedman & Berk, 2008) and discarded observations, and therefore favoring PS stratification because it is less vulnerable to extreme weights (Hong, 2010, 2012) or matching within caliper because it implements a precise enforcement of common support.

The analysis strategy presented can be used with other variations of weights based on propensity scores, such as inverse-probability of treatment weights (Cole & Hernan, 2008) to estimate the ATE. The results obtained are expected to generalize to more complex outcome models, such as multilevel structural equation models (Asparouhov, 2006; Lüdtke et al., 2008) and multilevel latent growth models (Duncan, Duncan, Okut, Strycker, & Li, 2002), because the same MPML estimation method can be used. However, the use of PS weighting strategies with models for data that are cross-classified rather than nested such as the cross-classified random effects model (Beretvas, 2008; Goldstein, 1994), the multiple-membership model (Browne, Goldstein, & Rasbash, 2001), and the cross-classified structural equation model (Asparouhov & Muthén, 2012) is an open area of research. In particular, it is not clear how weights based on propensity scores would affect estimates and standard errors with cross-classified data, and whether and how weights should be scaled.

ARTICLE INFORMATION

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: No external funding support.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like thank Paul L. Morgan for providing information for the applied example. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

REFERENCES

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*, 235–267.

Allison, P. D. (2009). *Fixed effects regression models*. Los Angeles, CA: Sage.

Anand, P., Mizala, A., & Repetto, A. (2009). Using school scholarships to estimate the effect of private education on the academic achievement of low-income students in Chile. *Economics of Education Review*, *28*, 370–381.

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

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics: Theory and Methods*, *35*, 439–460.

Asparouhov, T., & Muthén, B. O. (2012). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters.* Retrieved from http://www.statmodel.com/papers.shtml

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*, 399–424.

Bandalos, D. L., & Leite, W. L. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 625–666). Greenwich, CT: Information Age.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–973.

Bates, D., Maechler, M., & Bolker, B. (2011). Lme4: Linear mixed-effects models using s4 classes. Retrieved from http://cran.cnr.berkeley.edu/web/packages/lme4/index.html

Berends, M., Goldring, E., Stein, M., & Cravens, X. (2010). Instructional conditions in charter schools and students' mathematics achievement gains. *American Journal of Education*, *116*, 303–335.

Beretvas, S. N. (2008). Cross-classified random effects models. In Ann A. O'Connell & D. Betsy McCoach (Eds.), *Multilevel modeling of educational data* (pp. 161–198). Charlotte, NC: Information Age.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*, 1149–1156.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, *1*, 103–124.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. P. A., & Robins, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, *23*, 749–767.

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Optimal matching with a variable number of controls vs. a fixed number of controls for a cohort study: Trade-offs. *Journal of Clinical Epidemiology*, *56*, 230–237.

Cochran, W. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*, 295–313.

Cole, S. R., & Hernan, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*, 656–664.

Cook, T. D., & Steiner, P. M. (2009). Some empirically viable alternatives to random assignment. *Journal of Policy Analysis & Management*, *28*, 165–166.

Crump, R., Hotz, V. J., Imbens, G. W., & Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*, 187–199.

Cuong, N. V. (2013). Which covariates should be controlled in propensity score matching? Evidence from a simulation study. *Statistica Neerlandica*, *67*, 169–180.

Doyle, W. R. (2009). Impact of increased academic intensity on transfer rates: An application of matching estimators to student-unit record data. *Research in Higher Education*, *50*, 52–72.

Duncan, T. E., Duncan, S. C., Okut, H., Strycker, L. A., & Li, F. (2002). An extension of the general latent variable growth modeling framework to four levels of the hierarchy. *Structural Equation Modeling*, *9*, 303–326.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, *32*, 392–409.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, *22*, 364–375.

Goldstein, H. (2003). *Multilevel statistical models*. New York, NY: Halsted.

Griswold, M. E., Localio, A. R., & Mulrow, C. (2010). Propensity score adjustment with multilevel data: Setting your sites on decreasing selection bias. *Annals of Internal Medicine*, *152*, 393–396.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, *2*, 405–420.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education*, *75*, 31–65.

Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News*, *7*, 18–24.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis.* Boca Raton, FL: CRC.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2006). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*, 1–28.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2014). *How exactly are the weights created?* Retrieved from http://r.iq.harvard.edu/docs/matchit/2.4-20/How_Exactly_are.html

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Holmes, W. M. (2014). *Using propensity scores in quasi-experimental designs*. Thousand Oaks, CA: Sage.

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics*, *35*, 499–531.

Hong, G. (2012). Marginal mean weighting through stratification: A generalized method for evaluating multivalued and multiple treatments with nonexperimental data. *Psychological Methods*, *17*, 44–60.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel 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 & Research*, *26*, 329–367.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Kang, J. D. Y., & Schafer, J. L. (2007a). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.

Kang, J. D. Y., & Schafer, J. L. (2007b). Rejoinder: Demystifying double robustness, a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 574–580.

Kelcey, B. M. (2009). *Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (304929925.)

Kelcey, B. M. (2011a). Covariate selection in propensity scores using outcome proxies. *Multivariate Behavioral Research*, *46*, 453–476.

Kelcey, B. M. (2011b, April). Propensity score matching within versus across schools. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection processes vary across schools. Los Angeles, CA: Center for Study of Evaluation (CSE).

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Leite, W. L., & Zuo, Y. (2011). Modeling latent interactions at level two in multilevel structural equation models: An evaluation of mean-centered and residual-centered unconstrained approaches. *Structural Equation Modeling*, *18*, 449–464.

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

Lockheed, M., Harris, A., & Jayasundera, T. (2010). School improvement plans and student learning in Jamaica. *International Journal of Educational Development*, *30*, 54–66.

Lohr, S. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229.

Lumley, T. (2010). *Complex surveys: A guide to analysis using r*. New York: Wiley.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*, 103–120.

Mitra, R., & Reiter, J. P. (2012). A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*.

Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters: Which is preferable? *The American Statistician*, *59*, 173–179.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, *43*, 236–254.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods and Research*, *35*, 3–60.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York, NY: Cambridge University Press.

Muthén & Muthén. (2013). Mplus (Version 7.0). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398.

National Center for Education Statistics. (2010). *Early childhood longitudinal study (ECLS)*, *2010*. Retrieved from http://nces.ed.gov/ecls

National Center for Education Statistics. (2011). 2007–08 *SASS methods and procedures, 2010.* Retrieved from http://nces.ed.gov/surveys/sass/methods0708.asp

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447.

Ou, S. R., & Reynolds, A. J. (2010). Grade retention, postsecondary education, and public aid receipt. *Educational Evaluation and Policy Analysis*, *32*, 118–139.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society, Series B*, *60*, 23–56.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. New York, NY: Springer-Verlag.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel sem framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*, 805–827.

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

Ridgeway, G., Mccaffrey, D., Morral, A., Burgette, L., & Griffin, B. A. (2013). *Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package.* Retrieved from http://cran.r-project.org/web/packages/twang/vignettes/twang.pdf

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550–560.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, *34*, 441–456.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.

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

Rubin, D. B. (1973). The use of matching and regression adjustment to remove bias in observational studies. *Biometrics*, *29*, 185–203.

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). Comment: Which ifs have causal answers? *Journal of the American Statistical Association*, *81*, 961–962.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, *13*, 279–313.

Schochet, P. Z. (2009). Statistical power for regression discontinuity designs in education evaluations. *Journal of Educational and Behavioral Statistics*, *34*, 238–266.

Schochet, P. Z. (2012). Estimators for clustered education RCTS using the Neyman model for causal inference. *Journal of Educational and Behavioral Statistics*.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., 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.

Shadish, W. R. (2002). Revisiting field experiments: Field notes for the future. *Psychological Methods*, *7*, 3–18.

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

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

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling*, *9*, 475–503.

Steiner, P. M., Kim, J.-S., & Thoemmes, F. (2013). *Matching strategies for observational multilevel data. Paper presented at the Joint Statistical Meetings*, Alexandria, VA.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, *44*, 711–740.

Strumer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution, a simulation study. *Practice of Epidemiology*, *172*, 842–854.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*, 1–21.

Su, Y.-S., & Cortina, J. (2009, September). *What do we gain? Combining propensity score methods and multilevel modeling. Paper presented at the Annual Meeting of the American Political Science Association*, Toronto, Canada.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118.

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

Van Landeghem, G., De Fraine, B., & Van Damme, J. (2005). The consequence of ignoring a level of nesting in multilevel analysis: A comment. *Multivariate Behavioral Research*, *40*, 423–434.

Winship, C., & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology*, *25*, 659–706.

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2011). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis*, *34*, 45–68.