THE USE OF PROPENSITY SCORES FOR NON-RANDOMIZED DESIGNS WITH CLUSTERED DATA

Felix J. Thoemmes

Texas A & M University

Stephen G. West

Arizona State University

December 22, 2010

Abstract

In this article we propose several modeling choices to extend propensity score analysis to clustered data. We describe different possible model specifications for estimation of the propensity score: single level model, fixed effects model, and two random effects models. We also consider both conditioning within clusters and conditioning across clusters. We examine the underlying assumptions of these modeling choices and the type of randomized experiment approximated by each approach. Using a simulation study, we compare the relative performance of these modeling and conditioning choices in reducing bias due to confounding variables at both the person- and cluster-levels. An applied example is provided in which the effect of retention in grade 1 on passing an achievement test in grade 3 is evaluated. We find that models that consider the clustered nature of the data both in estimation of the propensity score and conditioning on the propensity score performed best in our simulation study, however other modeling choices also performed well. An applied example illustrates practical limitations of these models when cluster sizes are small.

The use of propensity scores in multisite non-randomized trials

As evidenced by this special issue and recent work by Thoemmes and Kim (in press), propensity score (PS) methods have become a widely used tool in estimating causal treatment effects in non-randomized studies and broken randomized experiments, i.e., randomized studies that suffer from attrition, treatment non-compliance, or both (Barnard, Frangakis, Hill, & Rubin, 2003). Multilevel models (MLM; also called hierarchical linear models) are a widely used tool for the analysis of data with a nested structure. In the present context, nested structures are those that include participants within larger units such as students within schools or patients within treatment sites. However, the use of PSs within the nested data structures addressed by MLM has received little attention. The primary exception is the work by Hong and colleagues (e.g., Hong & Raudenbush, 2006; Hong & Yu, 2007, 2008) and recent publications by Arpino and Mealli (in press), Gadd, Hanson, and Manson (2008), Grieswold, Localio, and Mulrow (2010), and Kim and Seltzer (2007). Researchers who wish to draw causal inferences using PS matching from non-randomized studies with nested structure currently have little guidance on how to incorporate methods from the causal inference literature.

The use of PS in the MLM context requires several special considerations. Among these considerations are decisions about how to model the influence of variables at each of the hierarchical levels of the MLM and choices of appropriate conditioning schemes for the estimated PSs. The primary focus of this paper is to extend the use of PSs to data from nonrandomized nested designs that are analyzed using MLM. In this article we focus on designs in which individual units at the lower level of the analysis are clustered and are non-randomly assigned to either treatment or control condition. We consider two diverse perspectives on clustering that may be taken. First, the cluster level may be a central feature of the design as

when schools having different academic resources and different policies about retention decide to retain or not retain poorly performing students (Hong & Raudenbush, 2006). There may be variations in treatment implementation and interference between students within the school. Treatment effects within cluster and their generalization across clusters becomes the focus of the study. In this first case the propensity score analysis attempts to approximate a multi-site randomized trial, in which units are randomized within individual clusters. Second, clustering may be an incidental feature of the design as when randomly selected members of groups of unacquainted individuals waiting to complete forms in state unemployment office are offered the opportunity to participate in a job seeking skills program (see Caplan, Vinokur, Price, & van Ryn, 1989 for a randomized experiment in this context). The treatment is assumed to be implemented without variation. The focus of the study is to estimate the average treatment effect for the population of individuals, controlling for the potential nuisance effect of incidental clustering. In this second case the propensity score analysis attempts to approximate a single level randomized experiment on individuals who happened to be clustered. This status of clustering as a central versus incidental feature of the design and the type of randomized experiment that is approximated will have implications for the PS approaches that should be considered. In this article we will not discuss PS designs in which treatment is given to the entire cluster so the unit of analysis is the cluster itself (for an example, see Stuart, 2007).

We will consider both theoretical and practical aspects of causal inference in MLM and present guidelines on how to use PS matching in the context of non-randomized MLM designs. We limit our presentation to designs in which treatment effects are estimated for a binary treatment condition. In the single level context, Rubin (1997) proposed extending PS approaches to designs with more than two treatment conditions by comparing sets of pairs of treatment

conditions (e.g., treatment 1 vs. control; treatment 2 vs. control), constructing a PS model separately for each focused comparison. All models that we present could be extended likewise. More complex multilevel multinomial propensity score models that simultaneously compare all treatment conditions are still in development.

Causal Effects and Confounding in Multisite Non-randomized Trials

In order to better understand the issues, we begin by defining the causal effect of interest in a multisite non-randomized trial and discuss how this causal effect can be estimated from observed data. Hong and Raudenbush (2003) defined causal effects in these designs within the framework of the potential outcomes model (e.g., Rubin, 1974, 1978).

Given a binary treatment condition (e.g., treatment vs. control), a causal effect in a single level experiment can be defined as the expected difference between the potential outcomes that we would observe if all participants (units) could hypothetically be observed under both the treatment and control conditions (Holland, 1986).

$$\delta = \mathrm{E}(\mathrm{Y}_{\mathrm{iT}}) - \mathrm{E}(\mathrm{Y}_{\mathrm{iC}}) \qquad (1)$$

where Y_{iT} is the response of unit i under the treatment condition, Y_{iC} is the response of unit i under the control condition, and δ is the average causal effect (Rubin, 2005). These potential outcomes are not observed in their entirety in any given experiment—the fundamental problem of causal inference (Holland, 1986) —but they can be estimated from observed quantities. In an ideal randomized experiment[1] this causal effect can be estimated by the simple difference in observed means of the treatment and control groups.

An issue in extending the potential outcomes model to the multilevel case is the underlying assumption in the single level case that each unit has only one potential outcome that depends deterministically only on the particular unit and the treatment assignment. This

assumption implies that the potential outcome should ideally be invariant to cluster membership, cluster composition, and the treatment assignments of other participants, and that treatment delivery should be identical across clusters. This assumption might be unrealistic in many applied research contexts that involve clustered data (Hong & Raudenbush, 2006). Hong and Raudenbush (2006) propose allowing cluster specific effects on the potential outcomes, but assume additivity: "the observed group composition and agent allocation …are viewed as random events that are exchangeable" (Hong & Raudenbush, 2003, p. 1850). In other words, even though cluster-specific effects might exist, they are regarded as random effects that have an expected value of zero in the population. This assumption implies that an average causal effect can still be estimated when averaged across all clusters and the associated exchangeable random cluster effects. Hong and Raudenbush (2003) express the potential outcomes along with the random cluster effects as:

$$
\begin{aligned}
(Y_{ijC}) &= \gamma_{00} + u_{0j} + e_{0ij} \\
(Y_{ijT}) &= \gamma_{00} + u_{0j} + \delta + u_{ij} + e_{1ij} \\
\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]
\end{aligned}
\qquad (2)
$$

where Y_{ijC} is the potential outcome of unit i in cluster j in the control condition and Y_{ijT} is the potential outcome of unit i in cluster j in the treatment condition. The grand mean of the potential outcomes in the control condition of all units across clusters is expressed as γ_{00}. Random site-specific differences from this grand mean that do not depend on the treatment are captured in u_{0j}, which is assumed to have a mean of zero and a variance of τ_{00}. The treatment effect is expressed as δ and quantifies the average increase that a single unit is expected to make when observed in the treatment as opposed to the control condition. Note that the expected outcome for units that

are assigned to the treatment condition also includes the random component u_{0j} as well as a second random component u_{1j}. This second random component quantifies the cluster specific deviation over and above the treatment effect δ. Finally, e_{0ij} and e_{1ij} are unit-specific random effects. The expected value of equation (2) can be expressed as

$$\begin{aligned} E(E(Y_{ijC})) &= \gamma_{00} \\ E(E(Y_{ijT})) &= \gamma_{00} + \delta \end{aligned} \tag{3}$$

where the expectations are taken across both units and clusters. This equation makes apparent that the random effects, if aggregated across clusters, do not affect the estimation of the grand mean, γ_{00}, or treatment effect, δ.

Given randomization, the unbiased average treatment effect across all clusters can be estimated with the following model:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{10} Z_{ij} + u_{0j} + u_{1j} Z_{ij} + r_{ij} \\ \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{pmatrix} \right] \end{aligned} \tag{4}$$

where γ_{10} is the estimated average causal treatment effect over all clusters. The treatment effect γ_{10} can be tested for statistical significance. Furthermore, variance in estimated treatment effects across clusters can be tested by assessing the statistical significance of the random component τ_{11}. Small and non-significant variance components can be interpreted as evidence that the treatment effect is relatively homogenous across clusters.

In trials that employ randomization, equation (4) provides a direct estimate of the causal effect. In non-randomized observational studies, equation (4) estimates the prima facie or apparent treatment effect that will nearly always require adjustment. Threats to internal validity, notably selection at both the individual and cluster levels, must be addressed and additional

assumptions made for a valid causal inference. The selection of level 1 units into either the treatment or control condition can potentially depend on covariates at the person level, the cluster level, or both. The treatment assignment variable Z could be related to a vector of variables X at level 1 and the average level of Z could be related to a vector of variables W at level 2. Consider a program offered to several schools in which students are invited to participate on a voluntary basis. The treatment selection in this case could be related to student level variables such as motivation and interest as well as school level characteristics such as the ethnic composition of the student body and public vs. private school status. The relationship between selection into treatment and level 1 covariates may differ across clusters or interact with variables at level 2. Motivation of students might be an important predictor of selection in one school, but unimportant in another school. In order to obtain an unbiased estimate of the causal effect, these potential confounding influences need to be controlled through a selection model that incorporates the confounding structure on both levels. The use of PSs in MLM is one such possible solution to model selection.

Hong and Raudenbush (2006) considered the first case in which clustering is a central feature of the design and the goal is to approximate a multi-site randomized trial. To obtain a valid causal interpretation of the treatment effect, they noted that several assumptions are needed. In their example, schools serve as clusters, and variations in treatment implementation across schools and interference between students within schools are permitted.

> "(a) Generalization of causal inferences is restricted to current school assignments, (b) there is no interference between schools, and (c) treatment assignment is strongly ignorable; that is one's own and one's peers treatment assignments are independent of the ensemble of potential outcomes given observed covariates." (p. 902).

In the second case in which clustering is an incidental feature of the design and the goal is to approximate a randomized experiment on individuals, some changes in the assumptions are necessary. In this case, it is assumed that there are no variations in treatment implementation across clusters and that selection is invariant across clusters. The multilevel analysis addresses incidental effects of the clustering of participants into groups. As in the approximation of the multisite randomized trial, the assumptions of no between-cluster interference and strongly ignorable treatment assignment given person-level covariates are needed.

Practical Issues in Estimating the Propensity Score in Trials with Clustered Data

A PS analysis involves several steps that can be briefly summarized as (a) assessment of critical covariates, (b) estimation of the PS, (c) conditioning on the PS, (d) model adequacy checks, and (e) treatment effect estimation (for a more complete overview see Caliendo and Koeping, 2007; Stuart, 2010; Thoemmes & West, 2010). In this section we consider several possible ways to conduct a PS analysis for multisite designs. Some aspects of PS analysis are easily translated to the more complex case of MLM; others require more study. As a preface to these particular considerations, we return to the issue of the role of clustering and the type of randomized experiment that is approximated by using propensity scores.

The first case in which the clusters are a central feature of the design approximates a multi-site randomized trial in which participants at level 1 (e.g., students) are randomly assigned to either treatment or control and this procedure is repeated across different clusters (e.g., schools). When approximating this design, it is of special importance that balance is achieved within each cluster. In this case the selection model (the relationship between covariates and treatment selection) may differ across clusters, and any one covariate can have a different regression weight attached to it in the prediction of treatment selection. If selection (and the

regression weight of the propensity score model) differs across clusters, matching within clusters is most suited to achieve balance on covariates within clusters. Also, matching within single clusters makes the use of level 2 variables (e.g., school characteristics) irrelevant. Units within the same cluster will by definition always have the same values on all observed and unobserved level 2 covariates.

The second case in which clusters are not a central feature of the design approximates a randomized experiment in which individuals are randomly assigned to either treatment or control conditions regardless of cluster membership. In this type of randomized experiment it would be possible to match individuals with identical propensity scores both within and across clusters and as a result achieve balance in the overall sample, but not necessarily within single clusters. Both level 1 and level 2 variables would need to be considered in the estimation of the PSs, as individual units would now be equated on a combination of individual-level and cluster-level characteristics. This case assumes that the selection model for the estimation of the PS from level 1 covariates will be identical across clusters. Under this assumption, if balance between the treatment and control groups on the PS is achieved, it is expected that balance on the covariates that contribute to the PS will also be achieved. In contrast, if the selection model could differ across clusters, units from different clusters could be matched on PSs based on different regression weights. In this case, matching of units from different clusters on PSs would not necessarily be expected to lead to balance on the covariates that contribute to the PS if the selection models differ across clusters.

Identification of Covariates

The first step of a PS analysis is to identify covariates that could potentially bias the treatment effect. In a design with clustered data the key additional consideration is that both

person-level covariates and cluster-level covariates can be considered to be potential confounders. The researcher should aim to have an exhaustive list of covariates from both level 1 and level 2 sources with a particular emphasis on those of known theoretical or empirical importance. If researchers form matches only within single clusters, level 2 covariates can be ignored.

Estimation of the Propensity Score

The PS can be estimated with diverse statistical methods including logistic regression, discriminant analysis, recursive partitioning (e.g., McCaffrey, Ridgway, & Morall, 2004), and generalized additive models. We limit our presentation here to the commonly used logistic regression model and its MLM extension, the generalized linear mixed model. This decision permits us to focus on conceptual differences concerning the treatment of clustering and estimation of effects at different levels in different modeling strategies. Following Grieswold, Localio, and Mulrow (2010) and Kim and Seltzer (2007), we examine (a) single level models in which all participants are pooled prior to estimation of parameters and clustering is ignored, (b) fixed effects regression models (see Allison, 2005) in which single level logistic regression is used for each individual cluster, and (c) multi-level model specifications that include random intercepts, slopes, or both. We consider how the PS is estimated in each of these models and which model parameters are important in the equating of units on the estimated PS.

Single level models.

The estimated propensity scores in single level models in which the vectors of person-level variables X and cluster-level variables W are included are specified as follows:

$$\text{logit}(e(x,\ w)) = \beta_0 + \sum_{p=1}^{P}\beta_p X_i + \sum_{q=1}^{Q}\beta_q W_j + \sum_{i=1}^{I}\beta_i W_j X_i\ , \qquad (5)$$

where logit(e(x,w)) is the estimated logit of the propensity score, β_0 is an intercept, $\sum_{p=1}^{P}\beta_p X_i$ is a vector of regression coefficients and predictor variables on the person level (potentially including person-level interactions and polynomial terms), $\sum_{q=1}^{Q}\beta_w W_j$ is a vector of regression coefficients and predictor variables on the cluster level (potentially including cluster-level interactions and polynomial terms), and $\sum_{i=1}^{I}\beta_i W_j X_i$ includes all possible interactions between person- and cluster-level variables. Note that a PS models applied to a real dataset will typically not include all possible predictor variables or interactions thereof.

The use of single level models with clustered data yields unbiased estimates of fixed effects, but does not include an estimate of the random effect for that cluster. This approach will only yield proper estimates of the PS for a unique cluster (approximating the multi-site randomized trial) when both (a) the intra-class correlation coefficient (ICC, a measure of the proportion of total variance that is between clusters) and (b) the random variability of regression coefficients across clusters is zero—highly unlikely in practice. Consider a data structure in which clusters differ substantially in their average level of an explanatory variable of treatment selection, e.g., motivation to participate in the intervention. Within each single cluster, motivation might be a strong predictor for treatment selection, but if a single level regression model is estimated the effect of motivation might be biased due to clustering effects (Kreft & de Leeuw, 1998). In addition, any variability in the strength of relationship between motivation and treatment selection between clusters will not be properly modeled by the basic single level model. The use of single level models to estimate PS with clustered data is expected to yield

biased estimates of the PS within single clusters. On the other hand, if selection does not differ appreciably across clusters and the goal is to approximate a randomized experiment in which individuals are assigned to treatment conditions irrespective of cluster membership, the single level model can be a viable choice for estimation of the PS.

Fixed effects models.

An alternative conceptual approach to estimating PSs in multisite trials is to use a separate logistic regression for each cluster. In practice, the selection model includes indicator variables that represent clusters and capture variation in cluster means. The selection model can also theoretically include interactions between the indicator variables and person-level variables; inclusion of such interactions is limited in practice because they utilize a large number of degrees of freedom. No cluster-level covariates can be included in the regression equation because they are constant within each cluster (perfectly collinear). The fixed effect model can be expressed as:

$$\text{logit}(e(x, w)) = \sum_{p=1}^{P} \beta_p X_i + \sum_{c=1}^{C} \beta_c C_i + \sum_{i=1}^{I} \beta_i C_i X_i \qquad (6)$$

where C is an indicator variable for cluster membership and remaining regression coefficients are defined as in Equation (5). Note that the model specification is very similar to the single level model from Equation (5), with the difference being that the indicator variable C in the fixed effects model allows for estimation of intercepts and possibly regression slopes of the X variables for every single cluster. As a result, the fixed effects model theoretically allows for estimation of unbiased within cluster regression slopes, regardless of presence or absence of any cluster-level covariates. Due to the presence of the cluster membership indicator variable, observed or unobserved cluster-level covariates cannot change the estimated PS. Two important limitations of this model are that very large sample sizes within each cluster are needed for the estimation of

this model (if cluster level interactions are included), and that the estimated PSs across different clusters are not comparable, because they are based on different regression equations within each cluster. This estimation model is appropriate when a multisite randomized trial is approximated.

Multi-level models.

Finally, we specify a multi-level model (MLM) with potential random intercept and slope components. In practice, not all random components might be modeled: Random effects should be included in the model when they are significantly different from zero and estimation does not yield a non-positive definite matrix of variance and covariance components.

MLM comes with additional complexities relative to the previously considered models. In models with random intercepts, slopes, or both, that are allowed to vary across clusters, predicted propensity scores can be either based on fixed effects only or both fixed and random effects (see e.g., McLean, Sanders, & Stroup, 1991; Robinson, 1991; Zeger, Liang, & Albert, 1998). This distinction between the inclusion vs. exclusion of random effects in estimating PSs is also referred to as the distinction between narrow (subject-level) and broad (population-wide) inference spaces. If specific PSs for the sampled units are desired, the narrow inference space with inclusion of random effects is used. These predicted values are usually referred to as best linear unbiased predictors (BLUPs). The narrow inference space captures PS models that mimic a randomized multisite trial, whereas the broad inference space captures PS models that mimic a randomized individual trial with incidental clustering.

The narrow inference space uses the full MLM model with fixed and random effects:

$$\text{logit}(e(x,w)) = \gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{ij} + \sum_{q=1}^{Q}\gamma_{0q}W_j + \sum_{i=1}^{I}\gamma_{1i}W_j X_{ij} + u_{0j} + \sum_{p=1}^{P}X_{ij}u_{1j}, \quad (7)$$

where $\sum_{p=1}^{P}\gamma_{p0}X_{ij}$ is a vector of regression coefficients and subject-level covariates, $\sum_{q=1}^{Q}\gamma_{0q}W_j$ is a vector of regression coefficients and cluster-level covariates, $\sum_{i=1}^{I}\gamma_{1i}W_j X_{ij}$ is a vector of all possible interaction terms between subject- and cluster-level covariates, u_{0j} is a random effects component influencing the intercept of each cluster j, and $\sum_{p=1}^{P}X_{ij}u_{1j}$ is a vector of random effects components that influence each of the regression slopes of subject-level predictors. All random effects u are assumed to be normally distributed with mean 0 and estimated variance τ. Random effects may be allowed to correlate with each other. The model used to estimate the PS based on the broad inference space is identical to Equation 7 with the omission of all random effects:

$$\text{logit}(e(x,w)) = \gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{ij} + \sum_{q=1}^{Q}\gamma_{0q}W_j + \sum_{i=1}^{I}\gamma_{1i}W_j X_{ij} \quad (8)$$

Several important observations can be made about the MLM models. First, two different types of predicted probabilities can be extracted from the MLM, depending on whether the broad or narrow inference space is used. Predicted values based on the two inference spaces can differ dramatically if the estimates of the random effects components are large. Of key importance is to differentiate between random components influencing the intercept vs. slope estimates. The random effect component of the intercept will change the average level of the estimated PS within a single cluster, and should therefore not have any profound effect on later conditioning, because the rank ordering of estimated PS will not be changed within the cluster. Random effect

components of random slopes on the other hand can have strong effects on both the estimation of the PS and potentially the conditioning and resulting treatment effects, because the rank ordering of estimated PS can change dramatically based on inclusion or exclusion of the slope-specific random effects. A noteworthy difference between the model parameters from MLM as compared to the single level or fixed effects model is that estimates of within cluster relationships in MLM are based on the shrinkage estimator (see e.g. Little, Milliken, Stroup, and Wolfinger, 1996).

The narrow inference space MLM allows for the selection model to differ across clusters, and is most appropriate for matches within single clusters, approximating a multisite randomized trial. The broad inference space MLM does not allow for the selection model (the relationships between covariates and treatment selection) to differ across clusters, except as a function of cross-level interactions and is more suited to the approximation of randomized experiments of individuals in which cluster membership is not a central feature.

Summary: Similarities and Differences between Estimation Methods

What are the key conceptual differences and similarities between these estimation methods and how do they relate to different approximations of randomized experiments? The single level model ignoring clustering (SL) will yield increasingly biased estimates of PSs within single clusters to the extent that the ICC and random slope variance both diverge from 0. However, all estimated PSs are based on the same model, i.e., the same regression coefficients relating covariates and treatment selection are used in estimation. One would therefore expect that covariate balance could be achieved in the whole sample if conditioning on the estimated PSs from the SL model were performed. However, covariate balance would unlikely be achieved within clusters and residual bias of treatment effects could remain.

The MLM with broad inference space (MLM-B) is conceptually similar to the SL model so that we would expect to see predicted PSs that are closely related to each other across the SL and MLM-B methods. Both models apply a single set of regression predictors to the estimation of each PS score. The difference between the models is that the estimates of the MLM-B model are shrunken towards the overall mean. Both estimation techniques therefore map on to situations in which randomized experiments are approximated in which individuals are assigned to treatment or control irrespective of cluster membership. These approaches are suited for matching across the full sample, but not necessarily within single clusters.

Conceptual similarity can also be observed between the fixed effects (FE) model and the MLM with narrow inference space (MLM-N). Both of these models use regression slope estimates relating covariates to treatment selection that can vary substantially across clusters and might not even be comparable across different clusters. As a result these estimation techniques can be used to approximate multisite randomized trials in which matching is constrained to units that share the same cluster membership, approximating a randomized multisite trial.

Conditioning

Conditioning on the PS can be achieved using diverse methods (e.g., matching, stratification). In the context of multisite trials, conditioning can be performed within single clusters or across clusters. With matching or stratification this implies that matches or strata would be formed only within each cluster (i.e., matching would be restricted to units from the same cluster) or across clusters (i.e., matching could be performed across the whole sample, ignoring cluster membership). Conditioning within clusters (CWC) is conceptually an option that maps on to the approximation of the multisite randomized trial, because treatment effects are ultimately computed for each single cluster and then averaged across clusters. CWC ensures that

covariate balance is achieved within clusters and that individual treatment effects within cluster can be estimated without bias. The practical drawback of CWC is that for small sample sizes the number of appropriate matches (and ultimately the overall size of the matched dataset) can diminish sharply therefore making conditioning within single clusters nearly impossible, a point also raised by Arpino and Mealli (in press).

In practice, conditioning across clusters (CAC) may often be an attractive option, if CWC is infeasible due to smaller sample sizes within clusters. A single unit in a cluster for which no appropriate within cluster match was found could be matched with a similar unit from another similar cluster, thereby increasing the overall sample size. The key consideration for this approach is that under certain estimation methods (FE or MLM-N) the estimated PSs will be based on different regression equations. The PSs might not (even in expectation) yield balance on the background covariates, either in the entire sample, or within single clusters. This would prohibit estimation of an unbiased treatment effect. Instead, when CAC is used (and an experiment is approximated in which individuals are randomized without consideration of cluster membership), the estimation techniques SL and MLM-B should be used, as they both do not model selection within single clusters.

It is illustrative to consider how two randomly picked units (one treated; one control) could differ from each other in their estimated PSs. The estimated PS for each unit will determine whether they will be potentially equated, whether the basis for equating is matching, propensity score weighting, or stratification. Following Kim and Seltzer (2007) we consider how two units (X_T and X_C) and their associated PSs could be different in each of the estimation methods. Differences in estimated PSs could potentially be obtained when conditioning occurs

within or between clusters. Here we only consider combinations of estimation and conditioning that map on to approximations of the two different types of randomized experiments.

Single level models (SL) under conditioning across cluster (CAC)

The difference between any two estimated PSs from the SL model with CAC would be:

$$\text{logit}(e(x, w)_T) - \text{logit}(e(x, w)_C) =$$

$$= \left(\beta_0 + \sum_{p=1}^{P} \beta_p X_{xT} + \sum_{q=1}^{Q} \beta_q W_{qjT} + \sum_{i=1}^{I} \beta_i W_{qT} X_{pT} \right) - \left(\beta_0 + \sum_{p=1}^{P} \beta_p X_{pC} + \sum_{q=1}^{Q} \beta_q W_{qjC} + \sum_{i=1}^{I} \beta_i W_{qC} X_{pC} \right)$$

$$= \left(\sum_{p=1}^{P} \beta_p X_{pT} + \sum_{q=1}^{Q} \beta_q W_{qjT} + \sum_{i=1}^{I} \beta_i W_{qT} X_{pT} \right) - \left(\sum_{p=1}^{P} \beta_p X_{pC} + \sum_{q=1}^{Q} \beta_q W_{qjC} + \sum_{i=1}^{I} \beta_i W_{qC} X_{pC} \right)$$

$$= \sum_{p=1}^{P} \beta_p \left(X_{pT} - X_{pC} \right) + \sum_{q=1}^{Q} \beta_q \left(W_{qjT} - W_{qjC} \right) + \sum_{i=1}^{I} \beta_i \left(W_{qT} X_{pT} - W_{qC} X_{pC} \right). \quad (9)$$

Equation (9) implies that the difference in the PSs is based on a weighted combination of differences on the person-level X variables, cluster-level W variables, and the interaction of the two sets of variables. Two units will have an identical estimated PS if all values on all X and W variables are identical, or if the weighted combinations of X and W variables are identical. This implies the not widely known fact that two units with very similar estimated PSs do not have to be identical in their composition of background variables. Balance on the background variables is achieved on average in the distribution in the matched sample, but not necessarily for any single matched pair. A further implication is that the CAC type conditioning is expected to create balanced covariates on both levels across the whole sample. This implication does not necessarily imply that the person-level X variables within each cluster will be perfectly balanced.

However, when one is approximating the type of randomized experiment with incidental clustering, the primary goal is to achieve balance in the full sample.

Single level models (SL) under conditioning within cluster (CWC)

In the case that conditioning is restricted to units within the same cluster, all W variables become constants and the difference between two estimated PSs in the SL model simplifies to:

$$\text{logit}(e(x)_T) - \text{logit}(e(x)_C) = \sum_{i=1}^{I}\sum_{p=1}^{P}(\beta_i + \beta_p)(X_{pT} - X_{pC}) \quad (10)$$

This result implies that under CWC the difference between estimated PSs only depends on person-level variables, the regression coefficients of the X variables, and the cross-level interaction regression coefficients. The interaction regression coefficients have a potential effect on each single regression coefficient of the person-level covariates. This can lead to situations in which different clusters (with accompanying different interaction weights) yield estimated PSs that are based on different regression coefficients of the person-level covariates. This combination of estimation and conditioning approximates a multisite randomized trial, however also assumes that the SL model is a proper approximation of the selection within single clusters.

Fixed effects model (FE) under conditioning within cluster (CWC)

Under CWC the difference in estimated PSs can be expressed as:

$$\text{logit}(e(x)_T) - \text{logit}(e(x)_C) = \sum_{i=1}^{I}\sum_{p=1}^{P}(\beta_i + \beta_p)(X_{pT} - X_{pC}) \quad (11)$$

which is identical to the SL model. However, individual regression coefficients might differ substantially between the SL model and the FE model, depending on the amount of intra-class variability and random slope variability. The FE model specification should be restricted to CWC, because the estimated PSs from different clusters are based on incomparable models. In

the FE model the estimation of the PS under CWC only depends on person-level variables and person-level regression coefficients. This result is important because it implies that as long as conditioning is performed within cluster, cluster-level variables do not need to be considered in the estimation of the PS. Observed and unobserved cluster variables cannot bias the estimation of the PS and the treatment effect. Performing CWC with the FE model shields the estimate of the causal effect from any bias that could arise from omitted cluster-level variables.

MLM-B model under conditioning across clusters (CAC)

Estimated PSs in the MLM-B model are unaffected by the random effects of intercepts and slopes across clusters. Estimation of PSs is only based on fixed effects in the model. As a result, the estimated PSs are conceptually similar to those of the SL model with the difference that fixed effects in the MLM-B model are shrunken towards the average slope across clusters. The difference between two estimated PSs can be expressed as:

$$\text{logit}(e(x, w)_T) - \text{logit}(e(x, w)_C) =$$

$$\left(\gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{pTj} + \sum_{q=1}^{Q}\gamma_{0q}W_{qj} + \sum_{i=1}^{I}\gamma_{1i}W_{qj}X_{pTj}\right) - \left(\gamma_{00} + \sum_{p=1}^{P}\gamma_{p0}X_{pCj} + \sum_{q=1}^{Q}\gamma_{0q}W_{qj} + \sum_{i=1}^{I}\gamma_{1i}W_{qj}X_{pCj}\right) =$$

$$\sum_{p=1}^{P}\gamma_{p0}\left(X_{pTj} - X_{pCj}\right) + \sum_{q=1}^{Q}\gamma_{0q}\left(W_{qjT} - W_{qjC}\right) + \sum_{i=1}^{I}\gamma_{1i}\left(W_{qT}X_{pT} - W_{qC}X_{pC}\right) \quad (12)$$

where γ is a regression weight associated with either person-level, cluster-level variables, or interactions thereof, depending on the subscript. As in the SL model, this result illustrates that conditioning on the estimated PS will depend on weighted combinations of differences on the

subject and cluster level. Conditioning on this estimated PS will usually not balance covariates within any cluster, but may provide balance across clusters in the whole sample.

MLM-B model under conditioning within clusters (CWC)

In the case of restriction to CWC, the difference in PSs simplifies to:

$$\text{logit}(e(x)_T) - \text{logit}(e(x)_C) = \sum_{p=1}^{P}\sum_{i=1}^{I}\left(\gamma_{p0} + \gamma_{1i}\right)\left(X_{pTj} - X_{pCj}\right). \quad (13)$$

Note again that the interaction term can alter the value of regression coefficients systematically across clusters depending on the values of the cluster-level covariates and that it is assumed that the model with the broad inference space is still a valid approximation of the selection within single clusters.

MLM-N model under conditioning within clusters (CWC)

Finally, we consider the case of the MLM-N model under CWC. This model is different from the MLM-B model because the random components of intercepts and slopes across clusters are included in the estimation of the PS. In real data examples it is doubtful that all possible random effects can or should be estimated. As a practical solution, individual random effects could be probed for significance and only significant random effects would then be included in the model estimation. In the case of CWC the difference between PSs is simply:

$$\text{logit}(e(x)_T) - \text{logit}(e(x)_C) = \sum_{i=1}^{I}\sum_{p=1}^{P}\left(\gamma_i + \gamma_p\right)\left(X_{pT} - X_{pC}\right) \quad (14)$$

All random effects and cluster-level effects do not influence the difference of units within a single cluster and can therefore be ignored in the difference between two units that reside in the same cluster. Again, this approach approximates a multisite randomized trial.

Summary: Model components and conditioning

Our examination of the influence of model components on conditioning permits us to reach several conclusions. First, only under the FE model are the cluster-level effects of covariate vector W completely ignored. All other models consider the influence or at least the potential interaction effects of cluster-level covariates. All models that use CWC are expected to yield balance within clusters.

In the case of CAC, where conditioning is performed across the whole sample, balance is achieved in the total sample but imbalances can remain within clusters under any estimation model. The CAC approach also requires that selection models do not differ across cluster. Imagine that a single person is matched on his or her estimated PS to a person from another cluster. Under a model that allows for different selection models across cluster, e.g., the MLM-N model, the estimated PSs from individuals from different clusters could be based on substantially different regression equations. Therefore, matching two units with the same PS from different clusters may not actually achieve balance on the person-level covariates because the estimated PSs were differently weighted across different clusters.

The FE model with CWC and the MLM-N model with CWC approximate multi-site randomized trials and allow the selection model to differ across clusters. The SL model with CWC, and the MLM-B model with CWC also approximate this type of randomized experiment but make the assumption that the selection model can only differ across clusters as a function of cross-level interactions. In all cases in which a multi-site randomized trial is approximated and CWC is used, covariate balance within clusters is of main importance. These models allow estimation of average treatment effects and treatment effect variability across clusters mimicking a randomized multisite trial. In contrast, the SL model with CAC, and the MLM-B model with

CAC approximate randomized experiments in which randomization occurred at the individual level without regard to clustering. In these models covariate balance is achieved in the total sample and not within single clusters.

Estimation of treatment effect

The final step of a PS analysis in the context of a trial with clustered data is the estimation of the treatment effect. Given prior conditioning on the estimated PS, the treatment effect could be estimated using either a fixed or random effects hierarchical linear model with the outcome variable being regressed on treatment assignment or models that use OLS, but correct the standard errors due to the clustered nature of the data. For theoretical or empirical reasons, the effect of the treatment could be allowed to vary across clusters. The significance of the treatment effect can be tested by examining the significance of the fixed effect for treatment, whereas treatment heterogeneity across clusters in the case of a multisite randomized trial can be assessed by examining the significance of the random slope component associated with the estimate of the average treatment effects in an MLM.

Simulation Study

Design

To explore differences between the estimation and conditioning choices and the effects of certain data characteristics, we conducted a simulation study. We varied the type of estimation strategy, the type of conditioning, overall sample size (simultaneously varying cluster size and number of clusters), and the degree of intra-class correlation in a full 4 x 2 x 2 x2 factorial design. The levels of the manipulated factors and key values of other features that were held constant are summarized in Table 1. Our main interest was to observe how well different

propensity score methods (both with regard to estimation and conditioning) could reproduce the treatment effect in the population.

We varied several factors that we believed might be influential in the performance of the PS models. In particular, we varied the amount of the intra-class correlation coefficient for all X variables from a very low .05 to a very high .50. We expected that single level models would be most biased under presence of a strong intra-class correlation coefficient. The ICCs for the outcome variable and treatment assignment were held constant. The ICC of the continuous outcome variable was held constant at a value of .10. The ICC of the binary treatment indicator was manipulated by introducing variability in the intercept estimate of the treatment assignment variable. The average treatment assignment probability was set to .5, indicating an even split between treated and untreated units. This value of .5 was allowed to vary randomly across clusters. We chose the random intercept variance to achieve a specified distribution of expected splits between treatment and control assignments across all clusters. This distribution was referenced to a normal distribution with specified mean and variance. In other words, based on our specification we were able to determine the frequency of particular treatment assignment probabilities across clusters. Based on this information we computed the ICC of treatment assignment using formulas provided by Ridout, Demetrio, and Firth (1999) and Zou and Donner (2004). The ICC of the treatment assignment variable was approximately .20. The treatment effect was set to account for 13% of the explained variance (moderate effect size per Cohen's, 1988, definition). The strength of confounding, defined as the relationship between covariates and treatment selection, and covariates and the outcome variable, was set to a constant value of 26% of the total explained variance (large effect size, Cohen, 1998) separately for both subject-level X and cluster-level W covariates. The amount of random slope variance for all subject-level

X variables was set to a relatively small value of .0026. This value was chosen based on consideration that the spread of slope coefficients across clusters could reasonably form a 95% confidence interval of +/- 0.1 on a standardized regression coefficient metric. In other words, the value for the slope variance implied that the more extreme values would be about +/- 0.1 from the mean slope coefficient. Finally, we set the amount of explained variance in the slope coefficients due to cluster-level W variables to 13% (moderate effect size, Cohen, 1988). We considered two levels of sample size: 20 clusters with 50 units each to represent a realistic large sample size and 200 clusters with 500 people each to represent an asymptotic sample size. These conditions were fully crossed with the estimation and conditioning schemes to provide a design with a total of 32 conditions. Each condition was replicated 1,000 times.

Data Generation and Analysis

Data were simulated using Mplus 5 (Muthén, & Muthén, 2009; Input file available from first author) using the following model specifications.

Selection model:

$$\log it(p) = \gamma'_{00} + \gamma'_{p0} X + \gamma'_{0q} W + \gamma'_{pq} XW + u'_{0j} + u'_{pj} X \tag{15}$$

$$\begin{bmatrix} u_{0j} \\ \dots \\ u_{pj} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & \\ 0 & \tau_{..} & \\ 0 & 0 & \tau_{pp} \end{pmatrix} \right]$$

where X is a vector of subject-level covariates with dimension p, W is a vector of cluster-level variables with dimension q, and u are random effects components of the intercepts or slopes depending on the subscript. The matrix of variance and covariance terms of random effects is

abbreviated with ellipses (…) that represent elements between our indexing variables. The residual term in this logistic regression is held constant at $\frac{\pi^2}{3}$.

Outcome model:

$$Y_{ij} = \gamma_{00} + \gamma_{z0}Z + \gamma_{p0}X + \gamma_{0w}W + \gamma_{pw}XW + u_{0j} + u_{zj}Z + u_{pj}X + r_{ij} \tag{16}$$

$$\begin{bmatrix} u_{0j} \\ u_{zj} \\ \dots \\ u_{pj} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & & & \\ 0 & \tau_{zz} & & \\ 0 & 0 & \tau_{\cdot\cdot} & \\ 0 & 0 & 0 & \tau_{pp} \end{pmatrix}\right]$$

where all definitions are identical to the selection model with the difference that the outcome variable is now the continuous variable Y and additional effects of the treatment Z (both fixed and random parts) are also included in the model. In both models, the value of p (the dimension of the person-level variable vector), and the value of q (the dimension of the cluster-level variable vector) were set to three, meaning that we had three variables on the person level and three variables on the cluster level. Specific sizes of parameters are given in the form of explained variance in Table 1 and a path model representation of the selection model is displayed in Figure 1. We made the selection mechanism complex, allowing for all possible interaction effects between cluster-level and person-level variables, and random slope variances for the person-level variables. However, we also made several simplifying assumptions. First, all person-level and cluster-level covariates were standardized (mean of 0 and variance of 1) and were within their respective level uncorrelated with each other. This lack of correlation of the covariates simplifies calculations, but is unrealistic in real datasets. However, this data structure does not appreciably change the behavior of the model in terms of the either the selection

mechanism or the relationship to the outcome. It is the unique effect of a covariate in selection and its relationship to the outcome variable over and above the other covariates that determines its status as a confounder. The second simplifying assumption was that relationships between person-level covariates and treatment assignment were specified in the selection model as additive and linear (with the exception of cross-level interactions). This might not be a realistic representation of conditions in substantive research; however the focus of the investigation was not on the performance of PS analyses under non-linear selection models, but rather on the comparison of estimation and conditioning strategies under controlled circumstances. The third simplifying assumption was that random intercept components and random slope components were not correlated with each other. Again, this was done to simplify calculations and estimation; more realistic conditions that allow for correlations between random slopes and intercepts could be specified. Lastly, we only included three variables as covariates on each level; propensity score models often include many more variables. However, the performance of the different estimation approaches is not expected to change dramatically with differing numbers of covariates given proper specification.

After data were simulated, we estimated PSs using SAS PROC LOGISTIC for single level and fixed effects models and SAS PROC GLIMMIX for random effects models. The model used to estimate the PSs included all person-level covariates X, all cluster-level covariates W, and all interaction effects. The random effects matrix was a diagonal matrix, with diagonal elements freely estimated and off-diagonal elements constrained to zero. This mirrored the data generating model. Conditioning was accomplished by stratification on the estimated PS using 10 strata, and then computing and averaging treatment effects across strata. For cross-cluster conditioning, the strata were formed across the whole sample, ignoring the cluster structure. Then within each

stratum, a multi-level model was used to estimate the treatment effect based on treatment assignment, allowing for both random intercepts and slopes. For within-cluster conditioning, we formed 10 strata based on the estimated PS within each cluster and treatment effects were estimated as in the cross-cluster conditioning models.

We collected several statistics as measures of performance: the amount of average raw bias, defined as the difference between the treatment effect in the population and the estimated treatment effect averaged across 1,000 replications, the Mean Square Error (MSE) defined as the squared deviation of the estimated treatment effect from the true population treatment effect averaged across 1000 replications, the percent coverage, defined as the percent of replications in which a 95% confidence interval covered the true treatment effect, and balance statistics within and across clusters for person-level and cluster-level variables.

Results

The results are presented in Table 2. In the low ICC conditions, differences on the performance measures across the different PS estimation and conditioning choices were subtle. For small sample sizes and a small ICC value, virtually all estimation and conditioning combinations performed very well, with both low values for bias and coverage rates for the true treatment effect close to 95%. Within this sample size and ICC combination, the CWC MLM-N model performed slightly better than other models; however, it did not appreciably outperform the other models. This slight advantage of the MLM-N model became more apparent in the large sample size, small ICC condition. The MSE did not differ appreciably across models in this combination of sample size and ICC, but was again lowest for the MLM-N model.

In the large ICC, small sample size condition, most models performed relatively well. As expected, the SL model seemed to perform worse than other models. The SL model was the only

model that ignored clustering effects, which were prominent in the large ICC conditions. In the large ICC condition, the MLM-N model slightly outperformed all other models in terms of raw bias, MSE, and coverage regardless of sample size. The CWC SL and CAC SL models seemed to perform worse than other models. The remaining models performed similarly, but all models that utilized CAC seemed to have slightly larger biases. In this study the CWC MLM-N model performed best.

In addition to bias and coverage, we also examined balance statistics for each condition (see Table 3). We assessed the average balance (defined as the median of all standardized differences on the means of the covariates) within clusters and across clusters for each data generating mechanism and PS model condition. For the balance across clusters, we further differentiated between balance on level 1 and level 2 variables. Consistent with our results on bias, balance was overall very good. This was expected as we included all covariates from the data generating model in our estimation and used the correct functional form. Some notable differences were that models that used CAC had markedly worse balance within clusters when sample sizes were small. With large sample sizes imbalance even within clusters disappeared. Larger average imbalances for level 2 variables were also observed for models that used CAC and large sample sizes. Finally, the imbalances within cluster were almost always largest for the SL model when comparing models that used CWC. Especially for conditions with large ICC, these imbalances within cluster were noticeable.

Our results indicate that the SL model becomes problematic as ICC values become larger. When the ICC was .05, bias was small, indicating that ignoring cluster membership can yield relatively unbiased results. When the ICC was .50, the SL model can yield badly biased results with coverage values as low as 11% rather than the expected 95%. As a caveat it should be noted

that in many applied circumstances, the expected ICC will usually be much smaller than .5. The MLM-N model performed best in almost all circumstances, especially when coupled with the CWC conditioning scheme. The combination of the MLM-N model and the CWC scheme takes clustering fully into account, both estimating PSs and conditioning within clusters. As expected, the MLM-N model coupled with CWC yielded good balance within clusters, which is crucial for the estimation of treatment effects when a multi-site randomized trial is approximated. The CWC FE and CWC MLM-B models also showed good performance in most circumstances, even though decreased coverage rates were observed under conditions of large sample size, and ICC = .50. Based on the results of our simulation study, the use of the SL model is discouraged and the use of the MLM-N model or FE model with CWC is encouraged, presuming that the sample size in the clusters allows for this approach. The MLM-N model worked very well in our simulation because the cluster size and distribution of treated and untreated subjects were similar across clusters. If this had not been the case (i.e., highly imbalanced clusters in terms of treatment assignment), any sort of CWC scheme could potentially fail, because too many units would not have suitable matches within each cluster. We explored this limitation in our applied example.

Applied Example: Retention in grade 1 and its effects on later academic performance

To illustrate the use of PS in MLM, we present a partial re-analysis of a study by Hughes, Chen, Thoemmes, and Kwok (2010), in which the impact of repeating grade 1 on passing the grade 3 Texas Assessment of Knowledge and Skills (TAKS) math achievement test was assessed. In this study 769 students across three districts in Texas were recruited when in grade 1. Of these 769 students, 165 were retained in first grade and 604 were promoted. We excluded all schools that had fewer than 2 retained children, because of perceived problems with matching and estimation of effects in the MLM. The reported sample size reflects this exclusion criterion.

At the end of grade 3, students took the TAKS math test. The research question was to determine whether retention in grade 1 had any impact on passing the grade 3 test. To bypass issues of missing data in our illustrative example we used a single imputation using SAS PROC MI to create a complete data set.

We used a total of 67 comprehensive covariates that were measured on the person level, and two variables that were measured on the cluster (school) level. These variables were intended to be a comprehensive set of covariates that have been shown in prior research to be related to grade retention, to achievement in the elementary grades, or ideally both. In this illustration, we estimated treatment effects for grade 1 retention on grade 3 TAKS math achievement using all of the estimation and conditioning strategies discussed above. To ensure comparability across different estimation and conditioning models, we first fit a saturated model (including all subject-level and cluster-level covariates and interactions). We used the same combinations of estimation and conditioning schemes that we used for the simulation study. We used 1:1 nearest neighbor matching without replacement and a caliper width of 0.1 standard deviation. The treatment effects were estimated based on the matched samples using a random effects MLM. To ensure convergence of the models with random effects, each regression slope of level 1 predictors was first tested individually for significant variation across clusters using SAS PROC MIXED. Only 5 slopes that showed significant variation were entered as random effects in the model. Slopes of other level 1 predictors were assumed to be fixed. The covariance matrix of random effects was constrained to zero on off-diagonal elements. For each combination of PS estimation and conditioning (CWC, CAC), we estimated the treatment effect (with 95% confidence interval), the balance of all covariates between the retained and the promoted groups in the matched sample as a whole, and the balance of all covariates within each

single cluster. Balance was defined as the standardized difference in means (Cohen's d) on all covariates. Balance statistics were averaged across all variables and clusters.

The results are presented in Table 4. This applied example made several limitations of the FE model approach apparent. Because of small sample sizes within clusters, virtually none of the single FE PS models could be properly estimated, at least not if all interaction effects between dummies and person-level covariates were included. Recall that the FE model posits a single logistic regression model for each cluster. In many clusters the number of predictors was larger than the number of students in this particular cluster, and even in larger clusters, problems of complete separation of groups occurred frequently. This problem was not observed with any of the other models that estimated the PS using the complete sample (either completely pooled as in the case of the SL model or partially pooled as in the case of the MLM-N and MLM-B models). A second complication that became apparent was that all combinations that used the CWC scheme, ended up with very small sample sizes. This was due to the fact that in many schools no appropriate matches could be found for many of the retained students. In fact, on average models that used CWC were able to include approximately 5% of children in the original sample, and only 11% of all retained children. As a result, estimated treatment effects were extremely variable and confidence intervals were exceedingly large. The theoretical advantages of the CWC scheme in approximating a multisite randomized trial could not be realized in this applied dataset in which extreme imbalances between treated and untreated units existed. The small remaining sample size led to unacceptably high variance in estimates of the treatment effect. In contrast, the CAC scheme retained more subjects because matches could be formed across different clusters. Of note is that the SL model retained almost 90% of all retained children and achieved an almost perfect balance on covariates when computed across the whole sample.

Treatment effects of the conditioned model were all positive, but had 95% confidence intervals that included 0.

The present applied example paints a more complicated picture of the prospect of using PSs in MLM. In cases of CWC too few units could be matched and resulted in highly variable treatment effects. As a result CAC models had to be preferred, approximating a randomized experiment in which individuals are assigned disregarding cluster membership.

Conclusion and Discussion

PS analysis with clustered data has been only rarely applied and previously has not received thorough examination in the literature. Our article considered several important modeling choices that need to be made in the context of clustered data, both at the stage of estimation and at the stage of conditioning. For the data generated in our simulation studies, modeling the clustered nature of the data using random effects multilevel models with a narrow inference space (inclusion of random effects in prediction of the PS) and conditioning within clusters yielded the best results; however, other modeling choices such as the fixed effects model and conditioning within clusters also performed well. We also provided theoretical arguments relating different estimation and conditioning choices to two different types of randomized experiments. In particular, researchers can choose to approximate a multi-site randomized trial and rely on conditioning within clusters. Our simulations show that in this case the multi-level model with narrow inference space is a viable choice, if researchers have large enough samples to support it. We also note that in a number of applied circumstances some approaches might not always be feasible. In pursuit of its goal of equating the treatment and control groups at baseline, propensity score methods often have the effect of reducing the number of cases available for analysis. Our applied example illustrated that severe imbalances in the proportion of retained and

promoted children yielded dramatic reductions in sample size, particularly when conditioning was restricted to occur within single clusters. This reduction in sample size can yield highly variable estimates of treatment effects that are unlikely to be informative for applied researchers. An alternative approach that can circumvent this problem of reduced sample sizes is to approximate a randomized experiment with individuals in which clusters are an incidental feature of the design. Here, conditioning on the estimated propensity score across clusters can be used, and according to our simulation studies, the multi-level model with broad inference space provides a viable estimation strategy.

References

Allison, P. (2005). Fixed Effects Regression Methods for Longitudinal Data Using SAS. Cary, NC: SAS Institute Inc.

Arpino, B., & Mealli, F. (in press). The specification of the propensity score in multilevel studies. Computational Statistics and Data Analysis.

Barnard, J., Frangakis, C., Hill, J., & Rubin, D. B. (2003). A principal stratification approach to broken randomized experiments: A case study of vouchers in New York City (with discussion and rejoinder). Journal of the American Statistical Association, 98, 299–323.

Caliendo, M., & Kopeinig, S. (2007). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22, 31–72.

Caplan, R. D., Vinokur, A. D., Price, R. H., & van Ryn, M. (1989). Job seeking, reemployment, and mental health: A randomized field experiment in coping with job loss. Journal of Applied Psychology, 74, 759-769.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ: Erlbaum.

Gadd, H., Hanson, G., & Manson, J. (2009). Evaluating the impact of firm subsidy using a multilevel propensity score approach. Working Paper Series, Centre for Labour Market Policy Research, 3, 1-25.

Grieswold, M., Localio, A., & Mulrow C. (2010). Propensity score adjustments with multilevel data: Setting your sites on decreasing selection bias. Annals of Internal Medicine, 152, 393-396.

Holland, P. W. (1986). Statistics and causal inference (with discussion). Journal of the American Statistical Association, 81, 945-970.

Hong, G., & Raudenbush, S.W. (2003). Causal inference for multi-level observational data with application to kindergarten retention study. Proceedings of the American Statistical Association, Social Statistics Section, [CD-ROM, pp. 1849-1856], Alexandria, VA: American Statistical Association.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multi-level observational data. Journal of the American Statistical Association, 101(475), 901-910.

Hong, G., & Yu, B. (2007). Early grade retention and children's reading and math learning in elementary years. Educational Evaluation and Policy Analysis, 29(4), 239-261.

Hong, G., & Yu, B. (2008). Effects of kindergarten retention on children's social-emotional development: An application of propensity score method to multivariate multi-level data. Developmental Psychology, 44(2), 407-421.

Hughes, J., Chen, Q., Thoemmes, F., & Kwok, O. (2010). Effect of retention in first grade on performance on high stakes tests in 3rd grade. Educational Evaluation and Policy Analysis, 32(2), 166-182.

Kim, J., & Seltzer, M. (2007). Causal inference in multilevel settings in which selection process cary across schools. Working Paper 708, Center for the Study of Evaluation (CSE), UCLA: Los Angeles.

Kreft, I., & de Leeuw, J. (1998). Introducing multilevel modeling. Thousand Oaks, CA: Sage.

Littell, R., Milliken, G., Stroup, W., & Wolfinger, R. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute Inc.

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

McLean, R., Sanders, W., & Stroup, W. (1991). A unified approach to mixed linear model. American Statistican, 45, 54-64.

Muthén, L.K., & Muthén, B.O. (2009). *Mplus users's guide. 6*th edition. Los Angeles, CA: Muthén, & Muthén.

Ridout, M., S., & Demetrio, C.G., & Firth, D. (1999). Estimating intraclass correlation for binary data. Biometrics, 55, 137-148.

Robinson, G. (1991). That BLUP is a good thing. Statistical Science, 6, 15-51.

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

Rubin, D. B. (1978).Bayesian inference for causal effects: The role of randomization. Annals of Statistics, 6, 34-58.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. Annals of internal medicine, 127 (8), 757-763.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association, 100, 322-331.

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

Stuart, E. (2007). Estimating causal effects using school-level datasets. Educational Researcher, 36, 187-198.

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

Thoemmes, F., & Kim, E. (in press). A systematic review of propensity scores in the social sciences. Multivariate Behavioral Research.

Thoemmes, F. & West, S. G. (2010). Equating Groups: Propensity Score and Other Matching Methods. New York: Taylor & Francis, in preparation.

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. Psychological Methods, 15, 18-37.

Zeger, S., Liang, K., & Albert, P. (1998). Models for longitudinal data: a generalized estimating equation approach. Biometrics, 44, 1049-1060.

Zou, G., & Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics, 60, 807-811.

Author note

Correspondence should be directed to Felix Thoemmes, University of Tübingen, Europastrasse 6, 72072 Tübingen, Germany, e-mail: felix.thoemmes@gmail.com or Stephen G. West, Psychology Department, Arizona State University, Tempe, AZ 85287-1104, e-mail: sgwest@asu.edu. Stephen G. West was supported by a Forschungspreis (research prize) from the Alexander von Humboldt Foundation. The data collected in the empirical example was supported by a grant from the National Institute of Child Health and Development (5 R01 HD39367) to Jan N. Hughes (PI), Department of Educational Psychology, Texas A & M University.

Footnotes

[1] An ideal randomized experiment is defined as meeting certain assumptions, such as proper randomization, no attrition, full treatment compliance, identical treatment for all participants, and that assignment of one participant does not alter the response of another (Holland, 1986; West & Thoemmes, 2010). The last two assumptions are sometimes referred to as the stable unit treatment value assumption (SUTVA, Rubin, 2010).

Table 1. Factors and levels of simulation study.

Factor	Level(s)
Number of X variables (person-specific confounders)	3
Number of W variables (cluster-specific confounders)	3
Total explained variance of all X variables in treatment	.26
Total explained variance of all X variables in outcome	.26
Total explained variance of W variables on random intercepts of treatment	.26
Total explained variance of W variables on random intercepts of outcome	.26
Unique explained variance in outcome by treatment	.13
Total explained variance of W variables on slopes of X variables	.13
ICC of treatment variable	.10
ICC of outcome variable	.20
Random slope variance of X variables on treatment	.0026
Random slope variance of X variables on outcome	.0026
ICC of X variables	.05 .5
Sample size (cluster × units)	20 × 50 200 × 500
Type of estimation of propensity score	SL FE MLM-B MLM-N
Type of conditioning used	CAC CWC

Note: SL= single level model, FE= fixed effects model, MLM-B= multi-level model with broad inference space, MLM-N= multi-level model with narrow inference space, CAC= conditioning

across clusters, CWC= conditioning within clusters. Only one level is presented for factors that were held constant.

Table 2. Results of simulation study averaged across 1,000 replications within each condition.

	Sample Size											
	small			large			small			large		
	ICC						ICC					
	.05						.5					
PS strategy	Bias	MSE*	%Cov	Bias	MSE*	%Cov	Bias	MSE*	%Cov	Bias	MSE*	%Cov
CWC SL	-.011	.307	93.2	-.005	.006	77.5	-.046	.620	83.9	-.014	.025	31.9
CWC FE	-.014	.325	94.1	-.005	.006	76.9	-.019	.437	92.5	-.006	.007	75.3
CWC MLM-N	-.002	.299	94.1	-.002	.004	87.1	-.014	.408	91.2	-.003	.006	85.5
CWC MLM-B	-.011	.303	93.7	-.005	.006	77.4	-.021	.416	90.4	-.006	.009	75.0
CAC SL	-.005	.293	94.5	-.005	.006	76.1	-.035	.488	87.5	-.019	.043	11.1
CAC MLM-B	-.005	.293	94.4	.-005	.006	76.1	-.016	.367	91.9	-.010	.015	53.0

Note: SL= single level model, FE= fixed effects model, MLM= multilevel model, CAC= conditioning across clusters, CWC= conditioning within clusters. Bias = Raw average bias, MSE = Mean Square Error multiplied by 100, %Cov = Percentage coverage.

Table 3. Balance statistics of simulation study averaged across 1,000 replications within each condition.

	Sample Size											
	small			large			small			large		
	ICC						ICC					
	.05						.5					
PS strategy	Within clusters	Total sample L1	Total sample L2	Within clusters	Total sample L1	Total sample L2	Within clusters	Total sample L1	Total sample L2	Within clusters	Total sample L1	Total sample L2
CWC SL	-.013	-.014	-.003	-.009	-.008	-.001	-.081	-.005	-.006	-.029	-.004	-.005
CWC FE	-.024	-.011	-.001	-.012	-.007	-.001	-.024	-.008	-.004	-.016	-.004	-.003
CWC MLM-N	.011	.023	.003	-.006	.011	.002	-.013	.014	.019	-.011	.000	.001
CWC MLM-B	-.016	-.012	.000	-.009	-.007	-.002	-.029	.018	-.019	-.013	.027	-.055
CAC SL	-.148	-.023	-.035	-.009	-.008	-.034	-.195	-.055	-.152	-.024	-.018	-.140
CAC MLM-B	-.156	-.023	-.039	-.009	-.008	-.034	-.139	-.015	-.144	-.007	-.006	-.141

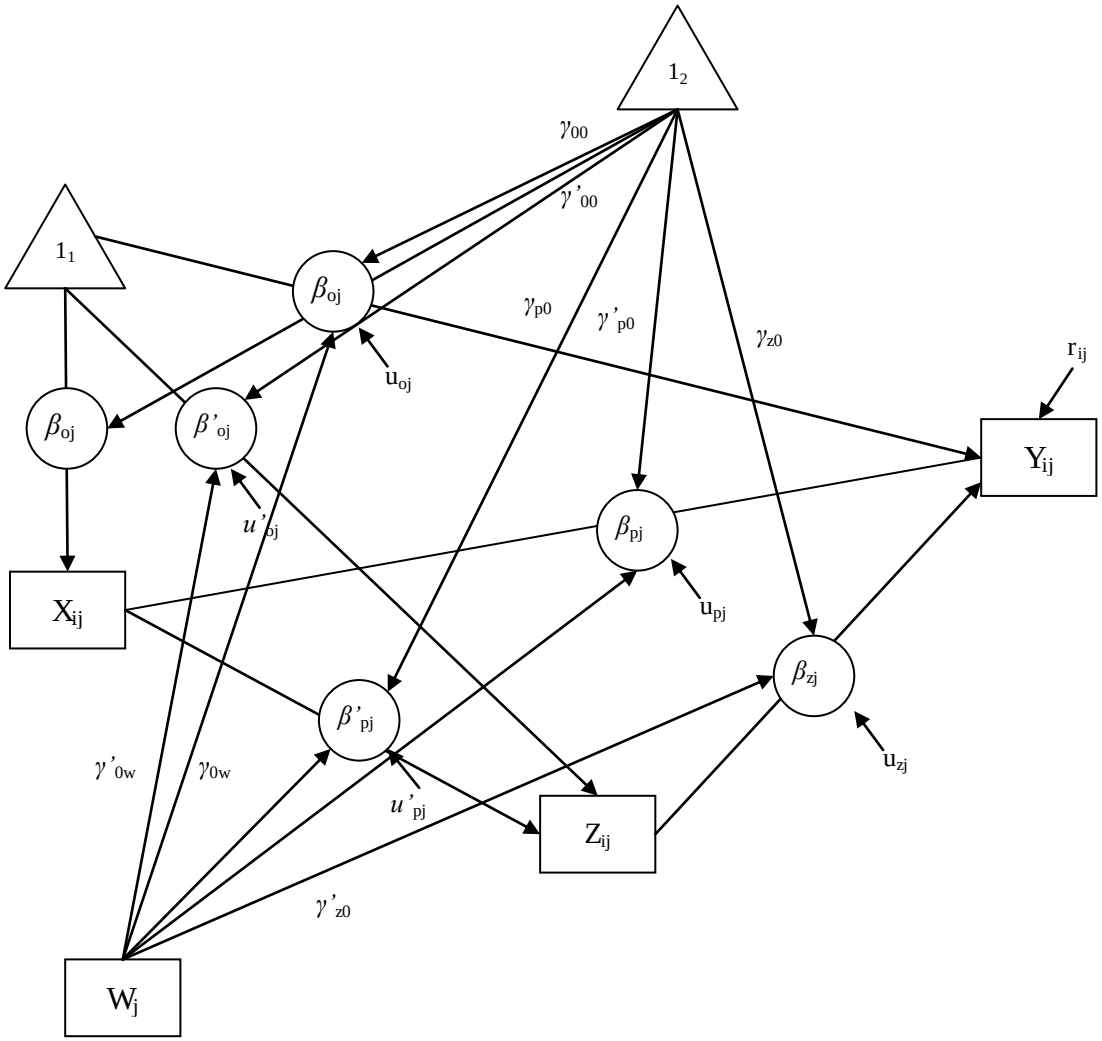
Note: SL= single level model, FE= fixed effects model, MLM= multilevel model, CAC= conditioning across clusters, CWC= conditioning within clusters. Within cluster = Median balance of all L1 covariates within single clusters averaged across all clusters and strata. Total sample L1 = Median balance of all L1 covariates in the whole sample averaged across all strata. Total sample L2 = Median balance of all L2 covariates in the whole sample averaged across all strata.

Table 4. Treatment effects and standard errors of applied example across all estimation and conditioning choices.

PS Strategy	Treatment effect	95% CI	Total matched sample	Average balance total sample	Average balance within cluster
Prima facie	-.39	(-.08, .04)	656	.04	.05
CWC SL	-1.25	(-3.67, 1.14)	32	-.03	.10
CWC FE			Not estimable		
CWC MLM-N	.51	(-1.49, 2.51)	24	.03	.11
CWC MLM-B	1.67	(.06, 3.27)	48	.07	.09
CAC SL	.42	(-.04, 1.27)	142	.02	.00
CAC MLM-B	.08	(-.96, 1.12)	79	.04	.00

Note: SL= single level model, FE= fixed effects model, MLM= multilevel model, CAC= conditioning across clusters, CWC= conditioning within clusters.

Figure 1. Data generation structure of simulated example, including a treatment effect from Z to Y, and confounding variables X and W on both levels of the analysis.



Note: Z is a treatment variable, Y an outcome of interest, X is a vector of level 1 confounders, W is a vector of level 2 confounders. Circles on regression paths, denote a random effect that varies across clusters. Triangles with "1" denote intercepts.