Matching Strategies for Observational Multilevel Data

Peter M. Steiner[1], Jee-Seon Kim[1], Felix Thoemmes[2]
[1]University of Wisconsin–Madison, Department of Educational Psychology, 1025 W Johnson Street, Madison, WI 53706
[2] Cornell University, Department of Human Development, MVR, Ithaca, NY 14853

Abstract

When randomized experiments cannot be conducted in practice, propensity score (PS) techniques for matching treated and control units are frequently used for estimating causal treatment effects. Despite the popularity of PS techniques, they are not yet well studied for matching multilevel data where selection into treatment takes place at the lowest level. Two main strategies for matching level-one units can be distinguished: (i) *within-cluster matching* where level-one units are matched *within* clusters and (ii) *across-cluster matching* where treatment and control units may be matched *across* clusters. Using a simulation study, we show that both matching strategies are able to produce consistent estimates of the average treatment effect. We also demonstrate that a lack in overlap between treated and control units within clusters cannot directly be compensated by switching to an across-cluster matching strategy.

Key Words: Causal inference, matching, propensity score, multilevel analysis

1. Introduction

In many fields of research, randomized experiments are considered as the canonical model for estimating causal effects of treatments or interventions. However, due to ethical or organizational reasons, for example, they cannot always be conducted. Instead, quasi-experiments like regression discontinuity designs, interrupted time series designs, instrumental variables, or non-equivalent control group designs are frequently used as second-best methods (Shadish, Cook, & Campbell, 2002; Wong et al., in press). In particular, the popularity of propensity score (PS) techniques for matching non-equivalent groups has increased during the last two decades (see Thoemmes & E.S. Kim, 2011). PS techniques like PS matching, inverse-propensity weighting or PS stratification are regularly used for removing selection bias from observational data. While a huge body of literature exists with regard to standard PS designs and techniques (e.g., Imbens, 2004; Rosenbaum, 2002; Rubin, 2006, Schafer & Kang, 2008; Steiner & Cook, in press), corresponding strategies for matching non-equivalent control groups in the context of multilevel data are still rather underdeveloped. This is surprising, given that multilevel structures pose several additional challenges in matching treatment and comparison groups. Though a few methodological publications on PS designs with multilevel data exist (Arpino & Mealli, 2011; Hong & Raudenbush, 2006; Kelcey, 2009; J. Kim & Seltzer, 2007; Stuart, 2007; Thoemmes & West, 2011), they do not address the full complexity of issues associated with multilevel data. Moreover, not all the challenges involved in matching multilevel data in practice are completely understood and explored.

This is not surprising because both PS techniques and multilevel modeling are complex topics on their own. Combining them adds an additional layer of complexity.

In comparison to observational single-level data, the main challenge with multilevel data is that they frequently exhibit a nested structure where (i) units within clusters are typically not independent, (ii) interventions may be implemented at different levels (e.g., an educational intervention might be implemented at the student, classroom, school, or district level), and (iii) selection processes may simultaneously take place at different levels and involve many stakeholders, differ from cluster to cluster, and might introduce selection biases of different directions at different levels. For these reasons, standard matching techniques that ignore the cluster- or multisite structure are, in general, not directly applicable. If the multilevel structure is ignored or not correctly reflected in matching treatment and comparison units biased impact estimates result.

The aim of this article is to briefly outline matching strategies for multilevel data. We sketch main concepts and ideas without going into formal details, and discuss practical challenges and issues using a simple simulated dataset as an illustrative example. The remainder of this article is structured as follows. Section 2 introduces the general matching strategies for hierarchically clustered data structures. Section 3 briefly introduces the Rubin Causal Model and its potential outcomes framework for multilevel settings and then discusses the main causal estimands and the assumptions required for estimating them. Using the simple simulation study, we exemplify issues involved in matching multilevel data in Section 4. Section 5 concludes with a brief summary.

2. General Strategies for Matching Multilevel Data

With a hierarchical two-level structure, treatment might be implemented either at the unit-level (level one) or cluster-level (level two). Treatment implementation at the cluster-level implies that the treatment status only varies across clusters and that all units within a cluster are assigned to either the treatment or control condition. In contrast, if a treatment is implemented at the unit-level, units self-select or are assigned into the treatment or control condition within each cluster. Thus, both treatment and control units are observed within each cluster. Depending on the level of treatment implementation and selection, the general matching strategy differs (Steiner, 2011). If treatment is implemented at the cluster-level one should match comparable treatment and control clusters because selection takes place at the cluster-level (Stuart, 2007). A cluster-level matching strategy mimics a cluster-randomized controlled trial where clusters are randomly assigned to treatment. Mahalanobis-distance matching on observed school-level covariates or standard PS techniques might be directly used since only schools need to be matched. However, with a small number of treatment and control clusters, balance on level-one covariates might not be satisfactory even after matching clusters on level-two and aggregated level-one covariates. Thus, one might consider an additional matching of level-one units within matched pairs or groups of clusters.

Whenever treatment is administered at the unit-level, units should be matched within clusters because selection into treatment occurs at the unit-level within clusters. Matching units within clusters mimics a randomized block design or multisite randomized trial where units are randomly assigned to the treatment condition within clusters (i.e., blocks or sites in experimental design terminology). Thus, the ideal matching strategy consists of matching units within each observed cluster (Rosenbaum, 1986; for incidental clustering

at or after treatment selection see Thoemmes & West, 2011). We refer to this strategy as *within-cluster matching*. Once units are matched within clusters, the average treatment effect across clusters can be computed by pooling cluster-specific estimates by meta-analytic approaches (Cooper, Hedges & Valentine, 2009) or by multilevel modeling (Raudenbush & Bryk, 2002; J.-S. Kim, 2009). As before, standard matching methods like Mahalanobis-distance matching or a PS technique might be used. As simple and theoretically sound such a within-cluster matching strategy is as rarely applicable it is in practice for two reasons. First, if extreme selection processes take place—like retention of poorly performing students (units) within schools (clusters)—we might lack comparable treatment and control units within each or some clusters. Second, with small sample sizes, we might only find poor rather than perfect matches for most units within a cluster (Kelcey, 2009; J. Kim & Seltzer, 2007; Thoemmes & West, 2011). Given that within-cluster matching strategies might be bound to fail in practice, *across-cluster matching* strategies that also allow for "borrowing" units from other clusters might offer a practical solution. Yet, as we will argue, this does not directly work. Note that using an across-cluster matching strategy does not imply that all matches need to allow for an across-cluster matching; only if no close matches can be found for some treatment or control units within a cluster we allow for a matching across clusters.

Given that only standard matching techniques are required if the treatment is implemented at the cluster-level, this article focuses exclusively on matching strategies for multilevel data where level-one units self-selected or got assigned into treatment and control conditions within clusters. The investigation of different matching strategies (within- and across-cluster matching strategies) and the conditions under which they can produce consistent causal estimates is of particular importance for two reasons. First, though we do not yet have a sound theoretical basis and understanding of across-cluster matching (as opposed to within-cluster matching) it is already regularly implemented in actual research practice (e.g., Griswold, Localio & Mulrow, 2010; Hong & Raudenbush, 2006; Hong & Yu, 2008; Hong & Hong, 2009; Hughes, Chen, Thoemmes & Kwok, 2011; Reardon, Cheadle & Robinson, 2009; Wu, West & Hughes, 2008a, 2008b, 2010) Second, randomized experiments frequently cannot be conducted within clusters due to interference issues. For instance, within schools, the treatment contrast might be compromised due to interferences among teachers and students and spillover effects. Thus, in order to evaluate effects of interventions within clusters, we are often forced to resort to matching strategies that rely on observational data where units self-select or get deliberately selected into treatment conditions.

3. Potential Outcomes and Causal Estimands in Multilevel Settings

3.1 Potential Outcomes in Multilevel Settings

In order to formalize the treatment effects of interest it is convenient to use the Rubin Causal Model (Holland, 1986; Rosenbaum & Rubin, 1983; Rubin, 1974, 1978) with its potential outcomes notation and its extension to multilevel settings by Hong and Raudenbush (2006). According to this model, each unit $i = 1, \ldots, N_j$ in cluster $j = 1, \ldots, J$ has a set of potential treatment and control outcomes that can be denoted as $Y_{ij}(Z_{ij}, \mathbf{Z}_{-ij}, \mathbf{S})$. The potential outcomes depend on three factors:

(i) Unit i's treatment assignment Z_{ij}, where $Z_{ij} = 0$ for the control condition and $Z_{ij} = 1$ for the treatment condition.

(ii) The other units' assignment status $\mathbf{Z}_{-ij}$, which is a vector consisting of all units' treatment assignment except for unit *i* (the subscript –*i* indicates that unit *i* is excluded). Note that the dependence of a unit's potential outcomes on the other unit's assignment represents a violation of the stable-unit-treatment-value assumption (SUTVA; Rubin, 1986). However, with clustered data, this is often the case in practice. For instance, students within classes or schools cannot be considered as being independent of each other.

(iii) The matrix **S** which indicates the units allocation to clusters. **S** is an incidence matrix with units representing the rows and clusters the columns. Thus, **S** indicates each unit's cluster membership.

Since a unit's potential outcomes depend both on the other units' treatment assignment and the overall assignment to clusters, the resulting set of potential treatment and control outcomes is too large to be usefully estimated. Thus, it is common to restrict the set of potential outcomes by assuming SUTVA and by restricting the generalizability of estimated treatment effects to the observed units' allocation to clusters (cf. Hong & Raudenbush, 2006). In the most restrictive case we get only two potential outcomes for each unit: the potential control outcome $Y_{ij}(0) = Y_{ij}(Z_{ij} = 0, \mathbf{S} = \mathbf{s}^*)$ and the potential treatment outcome $Y_{ij}(1) = Y_{ij}(Z_{ij} = 1, \mathbf{S} = \mathbf{s}^*)$, where **s*** indicates the observed allocation of units to clusters. In assuming SUTVA (i.e., no interference between units), the assignment status of all other units, $\mathbf{Z}_{-ij}$, no longer needs to be considered. Though this very restrictive formulation may be relaxed (e.g., Hong & Raudenbush, 2006), using only two potential outcomes simplifies the following discussion of issues involved in multilevel matching strategies.

3.2 Causal Estimands

Given the two potential outcomes $Y_{ij}(0)$ and $Y_{ij}(1)$, we can define the average treatment effect (ATE) for the entire population of units across all clusters as the expected difference in units' potential outcomes:

$$\tau = E[Y_{ij}(1) - Y_{ij}(0)]. \quad (1)$$

Frequently, not only the average across all clusters is of interest but also the average treatment effect for each cluster might be of interest:

$$\tau_j = E[Y_{ij}(1) - Y_{ij}(0) \mid J = j], \text{ for all } j \in J. \quad (2)$$

In addition to the average treatment effects for all units (i.e., treated and untreated together) the average treatment effects for the treated (ATT) is another causal quantity of interest. The overall and cluster-specific average treatment effects for the treated are defined as

$$\tau_T = E[Y_{ij}(1) - Y_{ij}(0) \mid Z_{ij} = 1] \text{ and}$$
$$\tau_{Tj} = E[Y_{ij}(1) - Y_{ij}(0) \mid Z_{ij} = 1, J = j]. \quad (3)$$

3.3 Conditional Independence Assumption (Strong Ignorability)

Since both potential outcomes are never observed simultaneously, the treatment effects cannot directly be estimated without further assumptions. In general, we can estimate unbiased treatment effects only if the pair of potential outcomes ($Y_{ij}(0), Y_{ij}(1)$) is independent of treatment assignment Z_{ij}. Block randomized experiments (or multisite randomized trials) achieve this independence by randomly assigning units to treatment conditions within clusters (blocks). For observational multilevel data, we require

conditional independence—also called strong ignorability (Rosenbaum & Rubin, 1983; Rubin, 1978): The potential outcomes need to be independent of treatment assignment, given the observed vector of unit-level (level-one) covariates **X** and the observed vector of cluster-level (level-two) covariates **W**:

$$(Y(0), Y(1)) \perp Z \mid \mathbf{X}, \mathbf{W} . \quad (4)$$

The formulation of the strong ignorability assumption directly suggests an exact matching of units on unit-level and cluster-level covariates. However, an (approximately) exact matching on a large set of covariates is frequently not feasible; Techniques based on the propensity score (PS) may be used instead.

Rosenbaum and Rubin (1983) proved that matching on the PS alone also yields unbiased estimates of the overall treatment effect, given that selection is strongly ignorable for observed covariates **X** and **W** (cf. Hong & Raudenbush, 2006). Let **X** and **W** be a set of unit- and cluster-level covariates that establish strong ignorability as defined in (4) and $e_{ij}(\mathbf{X}_{ij}, \mathbf{W}_j)$ be the corresponding PS, then potential outcomes are independent given the PS:

$$(Y(0), Y(1)) \perp Z \mid e(\mathbf{X}, \mathbf{W}) . \quad (5)$$

The PS $e_{ij}(\mathbf{X}_{ij}, \mathbf{W}_j) = P(Z_{ij} = 1 \mid \mathbf{X}_{ij}, \mathbf{W}_j)$ is defined as a unit's conditional probability of receiving the treatment, given the observed covariates $(\mathbf{X}_{ij}, \mathbf{W}_j)$. Since the true PSs are rarely known in practice they need to be estimated from observed pretreatment covariates using a parametric binomial regression model (e.g., a logit or probit model) or more flexible semi- or non-parametric approaches like generalized additive models (Wood, 2006) or statistical learning algorithms (McCaffrey, Ridgeway & Morral, 2004; Berk, 2008).

It is important to note that conditioning on covariates **X** and **W** (instead of the PS) in equation (4) implies a within-cluster matching if cluster-level covariates **W** allow for a unique identification of clusters (either via variations in cluster characteristics or fixed effect dummies). Such a unique identification of clusters is no longer possible if we condition on the PS as in equation (5), because units with identical PSs, $e_{ij} = e_{i'j'}$ $(i \neq i', j \neq j')$, might actually come from different clusters which implies $\mathbf{W}_j \neq \mathbf{W}_{j'}$. As discussed in Thoemmes and West (2011), a pair of PS-matched treatment and control units might be very different with regard to unit- and cluster-level covariates (despite having the same PS). This led Thoemmes and West to the conclusion that across-cluster matching should only be used if we can reasonably assume that the selection mechanism is identical across clusters. Similarly, Kim and Seltzer (2007) argue that across-cluster matching makes an unbiased estimation of cluster-specific treatment effects difficult or even impossible. Though the authors' reservations against across-cluster matching seem plausible, we think that they are too restrictive and that it is in fact possible to estimate unbiased causal effects even when units are matched across clusters with different selection mechanisms. We argue below that across-cluster matching produces consistent estimates of the overall treatment effects (ATE and ATT), given a correctly specified joint PS model and sufficient overlap of treatment and comparison cases within each cluster.

4. Issues and Strategies in Multilevel Matching: An Illustration

4.1 Simulation Setup

In order to see the challenges involved in matching multilevel data at the unit-level, consider the following simple simulation study with only four schools and with 200 to 500 students per school. In this example students represent the level-one units and schools the clusters. Assume that we are interested in estimating the average effect of retaining (instead of promoting) a student where the retention decision is exclusively based on students reading achievement scores. For both the data-generating selection and outcome models we use simple models involving only the reading pretest as single student-level covariate and the school-average of the pretest as single school-level covariate. The inclusion of a school-level covariate allows for different selection mechanisms and outcome models across schools. Different selection models across schools imply that students with high probabilities of retention in one school might have comparatively low retention probabilities in another school.

More formally, let the logit of the retention probabilities be a linear function of the pretest,

$$\text{logit}(\pi_{ij}) = \alpha_j + \beta_j X_{ij}, \tag{6}$$

where π_{ij} is the latent retention probability of student i (i = 1, …, N_j) in school j (j = 1, …, 4), X_{ij} the corresponding pretest of the reading achievement score, and α_j and β_j the school-specific intercepts and slopes, for j = 1, …, 4. We use a rather general notation for formulating multilevel models since in analyzing actually observed data we might either use fixed effects models (i.e., dummies and corresponding interactions) or random effects models. The actually observed treatment status $Z_{ij} \in \{0,1\}$, with 0 representing promotion and 1 indicating retention, is modeled as a Bernoulli-distributed random variable with retention probability π_{ij}: $Z_{ij} \sim \text{Bernoulli}(\pi_{ij})$. In generating the potential control and treatment outcomes, we used outcome models with varying slopes across schools but constant intercepts:

$$\begin{aligned} Y_{ij}(0) &= \upsilon + \omega_j X_{ij} + \varepsilon_{ij} \quad \text{for } Z_{ij} = 0, \\ Y_{ij}(1) &= \upsilon + \tau + \omega_j X_{ij} + \varepsilon_{ij} \quad \text{for } Z_{ij} = 1, \end{aligned} \tag{7}$$

where $Y_{ij}(0)$ and $Y_{ij}(1)$ are the potential outcomes of student i in school j, τ is the treatment effect, ω_j the school-specific pretest slopes, and υ is the intercept that is held constant across schools. We modeled the school-specific slopes as a linear function of the school-averages of the achievement score: $\omega_j = f(\overline{X}_j)$. Though we generated both potential outcomes for each student, we determined the actually observed outcome according to each student's treatment status: $Y_{ij} = (1 - Z_{ij})Y_{ij}(0) + Z_{ij}Y_{ij}(1)$.

Figures 1 to 3 describe the simulated data and simulation setup. Figure 1 shows the pretest distribution of the reading score (X) for the four schools. It can be seen that the schools differ with regard to their student composition. School 1 has on average the lowest performing students, while School 4 has the highest performing students.

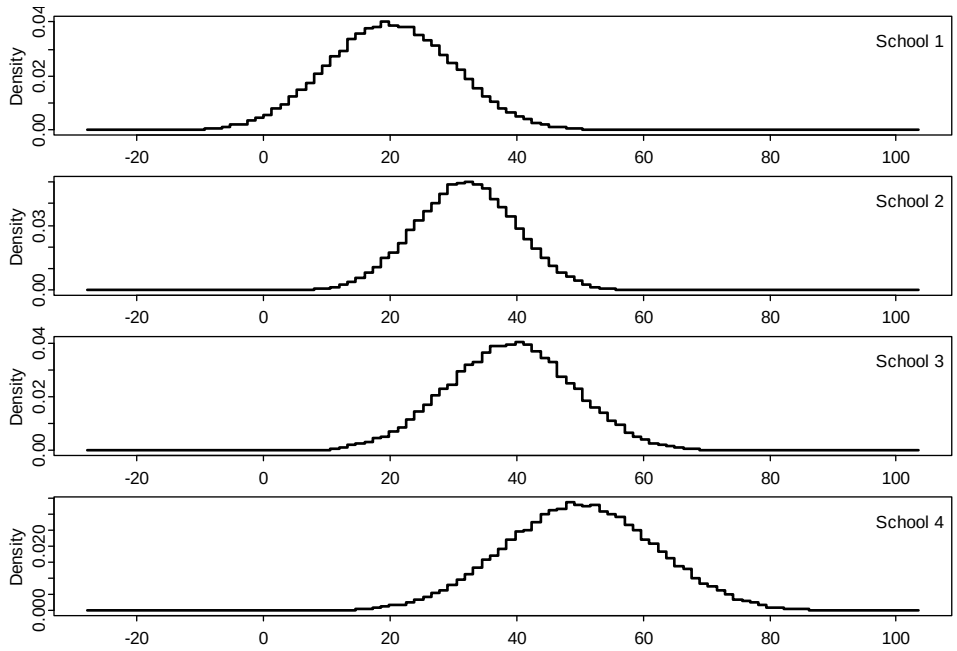

Figure 1: Distribution of pretest achievement scores. The four distributions show that Schools 1 to 4 differ with respect to their students' pretest scores.

Figure 2 shows the same pretest distribution but separated by retention status Z_{ij}. Though the poorest 20% to 30% of all students get retained in each school, the schools differ with regard to the strength by which they discriminate students to be retained or promoted (as indicated by the skewness and overlap of distributions). School 2 and school 4 exhibit a stronger discrimination of retained and promoted students than schools 1 and 3 do. Figure 2 makes it very clear that, due to the lack of overlap in school-specific distributions, we cannot find comparable promoted students for all retained students within each school. As a consequence, the school-specific average retention effects on retained students cannot be estimated without assuming constant treatment effects or severely restricting the generalizability of results to the overlapping population only. Allowing for matches across schools might solve this problem. For example, for all retained students in School 4 we could find students with comparable pretest scores in School 1. As we discuss below, however, such a strategy does not work in general.

Figure 3 shows the school-specific expectations of potential control outcomes $Y(0)$ (i.e., the potential control regression lines). Note that the higher a school's average pretest score, the higher its average posttest score but also the steeper its slope. For example, School 4 has the steepest slope, which indicates that this school has the students with the steepest average growth rate in achievement scores from one year to the next. But School 4 also has the best performing students to begin with.

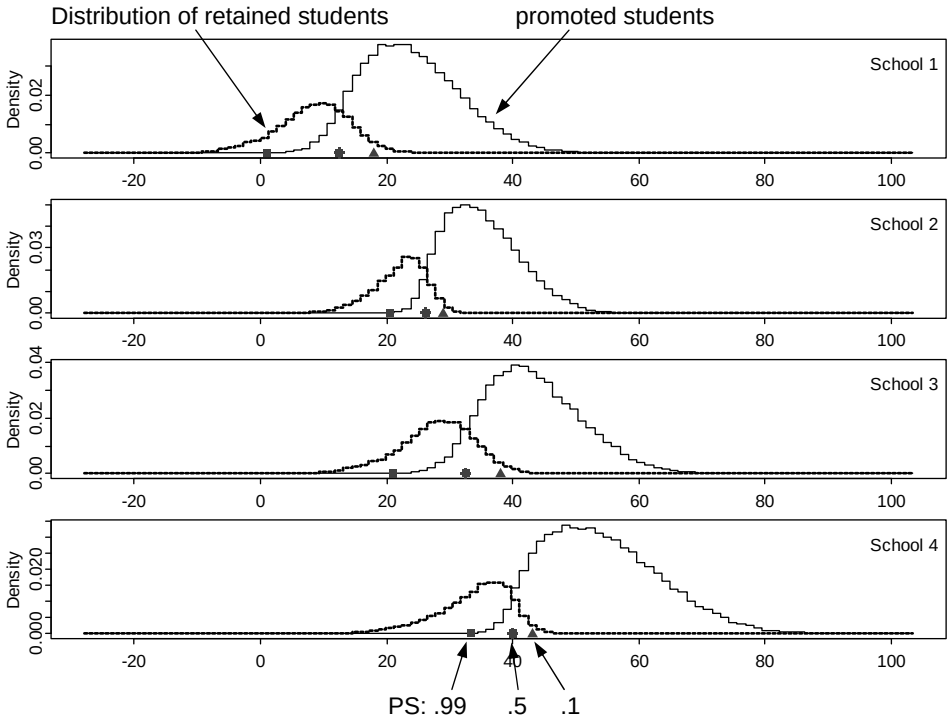

Figure 2: Distribution of pretest achievement scores for School 1 to School 4 by treatment status (retained/promoted). The darker dashed line represents retained students, the thinner line promoted students. The three different symbols represent students across schools that have the same PS: .99 (■), .5 (●), .1 (▲).

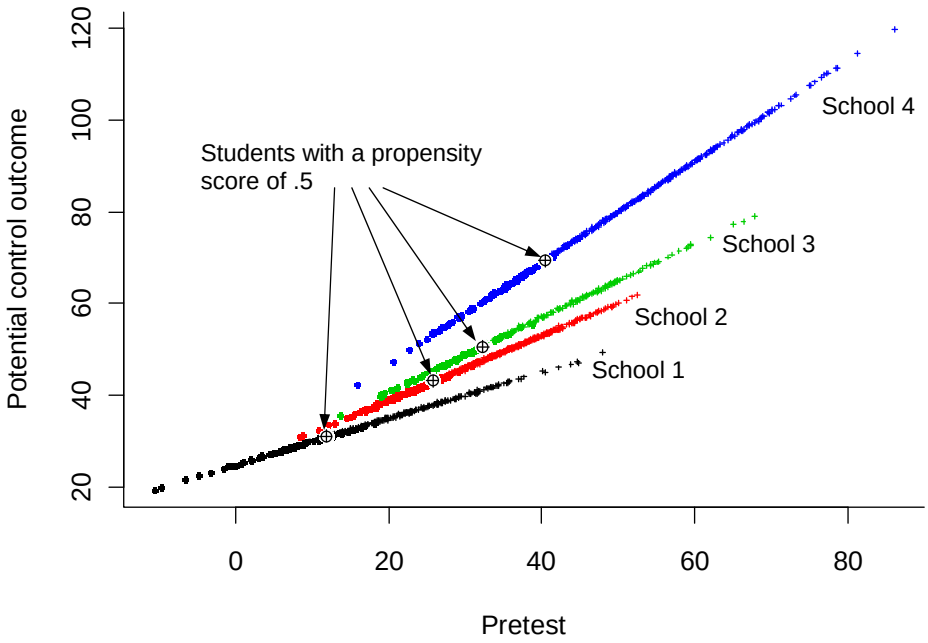

Figure 3. Potential control outcomes of School 1 to School 4. Large dots indicate students across schools having the same PS of .5. This plot shows that schools differ with respect to the distribution of pretest scores (*X*-axis) and the growth rates (slopes).

4.2 Simulation Results

Using the heterogeneous selection and outcome models across schools, we investigate which matching strategy—within-school or across-school matching—can produce unbiased estimates of the overall and school-specific treatment effects for retained students as defined in equation (3)? *Within-school matching* requires that we match students within each school separately and then pool school-specific estimates using weights that reflect the number of retained students in each school. For estimating the treatment effect, we used a PS stratification approach implemented via marginal mean weighting (Hong, 2010) and an additional covariance adjustment. The PSs were estimated using a joint logistic regression model across schools. Table 1 shows that within-school matching results in approximately unbiased estimates of the average retention effect on retained students—overall for all schools together but also for each school separately. The resulting estimates are unbiased in spite of the fact that we had to discard many students that did not overlap on the PS. This is so, because we generated the data assuming a constant treatment effect. We get nearly unbiased estimates even if the joint PS model ignores the cluster structure, that is, by not modeling school-fixed (or random) effects for intercepts and slopes. By ignoring the cluster structure, no bias is introduced since retention is a strictly monotonically decreasing function of the pretest score across all schools. Thus, school-specific PSs estimated via separate school-specific models or a joint (mis-specified) model does not change the PSs' rank order within schools (cf. J. Kim & Seltzer, 2007). However, with more complex selection models, this no longer holds.

Table 1. Average Treatment Effects for Retained Students
(Simulation standard deviations are in parenthesis; In simulating the data, the "true" treatment effect was set to -5 points, i.e., a slightly negative retention effect. Thus, estimates considerably deviating from -5 are biased.)

Estimates for:	All four schools	School 1	School 2	School 3	School 4
True effect	-5.00	-5.00	-5.00	-5.00	-5.00
Correctly specified PS model (school-specific intercepts and slopes)					
Within-school matching	-5.02 (0.04)	-5.01 (0.04)	-4.98 (0.05)	-5.01 (0.01)	-5.10 (0.13)
Across-school matching	-5.14 (0.20)	-5.54 (0.47)	-5.39 (0.08)	-5.47 (0.45)	-5.56 (0.57)
Mis-specified PS model (constant intercept and slope across schools)					
Within-school matching	-5.02 (0.04)	-5.00 (0.04)	-4.97 (0.05)	-5.01 (0.01)	-5.10 (0.13)
Across-school matching	0.18 (0.03)	-9.02 (0.34)	-11.63 (0.44)	-12.74 (0.16)	-13.32 (1.54)

The results are quite different for the *across-school matching* strategy. Unbiased estimates of the overall and school-specific retention effect on retained students are obtained only if the PS model is correctly specified—in our case if it includes intercept- and slope-fixed effects for each school. If the PS model is mis-specified, that is, if the variability of intercepts or slopes across schools is not (correctly) modeled, biased effect estimates result. Though this finding seems plausible, it is not clear why we can get unbiased effect estimates for an across-school matching. Figure 3 shows that students

with the same PS of .5 differ heavily with regard to the pretest across schools (*X*-axis). In School 1, students with a PS of .5 have a pretest score of around 15 points, while corresponding students in School 4 have a pretest score of around 40 points. Thus, in allowing for PS matches across schools, a treatment student in School 1 might get matched to a control student with the same PS from one of the other three schools. Yet those students' potential control and treatment outcomes are different from the treatment students' outcomes. Figure 3 shows the different potential control outcomes (*Y*-axis) for students with the same PS. So, why does across-school matching work even though we match students that might considerably differ on pretest measures?

Figure 2 shows an intuitive explanation. Whenever the joint PS model is correctly specified (as it is for this simple example), students with the same PS have the same relative position with respect to the school-specific pretest distributions of retained and promoted students. For example, students with a PS of .5 have a relative position that is exactly where the retained and promoted students' pretest distributions intersect, that is, where the risk of being retained (.5) equals the chance of being promoted. Those students with a PS of .99 are positioned exactly where the respective school's retention-promotion ratio is .99/.01. Thus, the relative risk of being retained is the same for students with the same PS (as we would expect), but the same relative risk is associated with different pretest scores across schools. The important point here is that students coming from different schools but with identical PSs have the same relative retention risk despite very different pretest values. That would not be the case if the PS model were incorrectly specified! Thus, given a correctly specified joint PS model, in matching promoted to retained students across schools, we get on average the same composition of promoted students as we would get for a within-school matching (the composition of students refers to their covariate distribution—here the pretest only). In other words, the seemingly mismatched students for School 1 are perfect matches for the other schools, while a part of the mismatches in Schools 2 to 4 are perfect matches for School 1. Since the proportions of mismatches are well balanced across all four schools (given the correct specification of the PS model) unbiased treatment effects for the overall population result. In our simulation, also the school-specific estimates are approximately unbiased because cluster sizes and the size of the treatment group are not very different across clusters (and the treatment effect was modeled as a constant term).

4.3 Overlap Issues and Strategies for Dealing with It

Interestingly, the simple simulation study also suggests that if we cannot find comparable retained and promoted students within a school—due to the lack of overlap, as shown in Figure 2—then we cannot find students with a comparable PS from other schools either. This is because students with a PS of .99 always exhibit the same extremity with respect to the school-specific pretest distribution as illustrated in Figure 2. If we take a treated student with a PS of .99 from School 1 we cannot find a corresponding comparison student in the same school nor in the other three schools. Thus, an across-school matching cannot directly solve overlap issues within schools. However, across-school matching might slightly improve overlap just due to the increased sample size of students across all clusters as compared to number of students within each single cluster.

However, severe overlap issues might be addressed by adapted versions of across-cluster matching strategies. One possibility is to look for comparison schools that never apply treatment (retention) to their students, or apply it to a considerably smaller portion of students. For non-nested data, Stuart and Rubin (2008) discussed a similar strategy of using local and non-local comparison groups when matches within a well specified target

population cannot be found. Another possibility to generate overlap is to "deliberately" mis-specify the joint PS model by partially ignoring the cluster-structure. This may work as long as it results only in a rank-preserving transformation of the true PS. However, the success of such a strategy might strongly depend on the identification of homogeneous groups of schools with similar selection processes or covariate distributions.

5. Summary

The results from this illustrative simulation study suggest the following:
(i) If the joint PS model is correctly specified then both within- and across-cluster matching produce consistent effect estimates across clusters.
(ii) Given sufficient overlap within clusters, within-cluster matching is preferable to across-cluster matching because it relies on weaker modeling assumptions—in particular, no cluster-level covariates are required.
(iii) A lack of overlap within clusters cannot directly be compensated by allowing for matches across clusters. More elaborate across-cluster matching strategies are required.

These findings from the illustrative simulation study will be more thoroughly investigated in future research using formal derivations and more complex simulation settings (including multiple level-one and level-two covariates, data-generating PS and outcome models with cross-level interactions, or different degrees of intraclass-correlations). More thorough simulations will compare fixed-effects and random-effects PS models, different PS techniques (matching, stratification, weighting) and mixed methods that combine PS adjustments with an additional covariance adjustment in estimating the treatment effect.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305D120005 (all three authors) and R305D100033 (first author). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

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

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York, NY: Springer.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. 2nd edition. New York, NY: Russell Sage Foundation.

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

Hong, G., & Hong, Y. (2009). Reading instruction time and homogeneous grouping in kindergarten: An application of marginal mean weighting through stratification. *Educational Evaluation and Policy Analysis, 31*, 54–81.

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

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*, 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,* 166–182.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics, 86*, 4–29.

Kelcey, B. M. (2009). Improving and assessing propensity score based causal inferences in multilevel and nonlinear settings. Dissertation at The University of Michigan. Available from: http://deepblue.lib.umich.edu/bitstream/2027.42/63716/1/bkelcey_1.pdf

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

Kim, J.-S. (2009). Multilevel analysis: An overview and some contemporary issues. In R.E. Millsap & A. Maydeu-Olivares (Eds.), *Handbook of quantitative methods in psychology* (pp. 337-361). Sage.

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.

Raudenbush, S., & Bryk, A., (2002). *Hierarchical Linear Models*. Thousand Oaks, CA: Sage Publishing.

Reardon, S. F., Cheadle, J. E., & Robinson, J. P. (2009). The effect of Catholic schooling on math and reading development in kindergarten through fifth grade. *Journal of Research on Educational Effectiveness, 2*, 45–87.

Rosenbaum, P.R., (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics, 11*, 207–224.

Rosenbaum, P. R. (2002). *Observational studies* (2nd Ed.). New York, NY: Springer-Verlag.

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. (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. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association, 81*, 961–962.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

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

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

Steiner, P. M. (2011). Matching strategies for multilevel data. Presentation at the *Society for Research on Educational Effectiveness*, Washington, DC. Abstract available from: http://www.sree.org/conferences/2011/program/downloads/abstracts/167.pdf

Steiner, P. M., & Cook, D. L. (in press). Matching and propensity scores. In Little, T. D. (Ed.), *The Oxford handbook of quantitative methods*.

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

Stuart, E. A., & Rubin, D. B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics, 33*, 279–306.

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

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

Wood, S. N. (2006). *Generalized additive models. An introduction with R.* Boca Raton, FL: Chapman & Hall/CRC.

Wong, V. C., Wing, C., Steiner, P. M., Wong, M., & Cook T. D. (in press). Research designs for program evaluation. In W. Velicer & J. Schinka (Eds.), *Handbook of Psychology: Research Methods in Psychology*. 2nd edition. Hoboken, NJ: Wiley and Sons.

Wu, W., West, S. G., & Hughes, J. N. (2008a). Short-term effects of grade retention on the growth rate of Woodcock-Johnson III Broad Math and Reading Scores. *Journal of School Psychology, 46*, 85–105.

Wu, W., West, S. G., & Hughes, J.N. (2008b). Effect of retention in first grade on children's achievement trajectories over four years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology, 100*, 727–740.

Wu, W., West, S. G., & Hughes, J. N. (2010). Effect of grade retention in first grade on psychosocial outcomes and school relationships. *Journal of Educational Psychology, 102*,135–152.