Abstract Title Page
Not included in page count.

Title:

Within-Cluster and Across-Cluster Matching with Observational Multilevel Data

Authors and Affiliations:

Jee-Seon Kim, University of Wisconsin-Madison

Peter M. Steiner, University of Wisconsin-Madison

Courtney Hall, University of Wisconsin-Madison

Felix Thoemmes, Cornell University

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

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 from observational data. Despite the popularity of PS techniques, they are not yet well studied for matching multilevel data where selection into treatment takes place among level-one units within clusters. For instance, students self-select into treatment conditions within schools (or teachers or classrooms). We investigate two different strategies for matching level-one units (students): (i) within-cluster matching where matches are only formed within clusters (schools) and (ii) across-cluster matching where treatment and control units may be matched also across clusters. Using a simulation study, we show that both matching strategies are able to produce consistent estimates of the average treatment effect. However, across-cluster matching requires stronger assumptions than within-cluster matching. We also demonstrate that a lack of overlap between treated and control units within clusters cannot directly be compensated by switching to a between-cluster matching strategy.

Matching treatment and control units in the context of multilevel data is typically more challenging than for data structures with a single level only because (i) units within clusters are not independent, (ii) interventions may be implemented at different levels (e.g., student-, classroom-, or school-level), and (iii) selection processes may simultaneously take place at different levels and involve many stakeholders (students, peers, parents, teachers, school management, parent teacher association), differ from school to school or district to district, and might introduce selection biases of different directions at different levels. Similarly, also the data-generating outcome model might differ across clusters. Because selection processes and outcome model might differ considerably across clusters, matching strategies need to take this heterogeneity across clusters into account (Authors, in press).

Given that treatment selection takes place among units within clusters, level-one units are ideally matched within clusters, mimicking a randomized block or multisite design. However, a within-cluster matching strategy might not always be feasible. First, with small sample sizes within clusters (as often the case in educational research) we might obtain only poor within-cluster matches. Second, extreme selection processes within clusters typically results in quite heterogeneous treatment and control groups lacking overlap. In both cases, within-cluster matching estimators might be considerably biased due to poor matches. Thus, the idea is to allow for matches across clusters—as a general strategy or only for units for which no close match within a cluster can be found. However, an across-cluster matching strategy relies on much stronger assumptions than a within-cluster matching approach: the level-one units' propensity score need to be correctly estimated across clusters, which requires the reliable measurement and correct modeling of all level-two covariates that explain selection and outcome differences across clusters (this is not necessary for within-cluster matching since the PS is estimated for each cluster separately). Using different simulated data scenarios, this paper compares the performance of within-cluster and across-cluster matching estimators.

Purpose / Objective / Research Question / Focus of Study:
Description of the focus of the research.

Using a simulation study, the purpose of the study is (a) to demonstrate under which conditions within-cluster matching breaks down but across-cluster matching strategies work, (b) to compare fixed-effects and random-effects models for estimating the unknown propensity score, and (c) to investigate the relative effectiveness of different PS techniques (optimal full matching on the PS, PS stratification and inverse-propensity weighting). The results of the simulation study are evaluated against the true population effect which is known for the simulation.

Setting:
Description of the research location.
(May not be applicable for Methods submissions)
NA

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.
(May not be applicable for Methods submissions)
NA

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.
(May not be applicable for Methods submissions)
NA

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

Though several recent studies already investigated across-cluster matching strategies (Arpino & Mealli, 2008; Hong & Raudenbush, 2006; Kelcey, 2009; Kim & Seltzer, 2007; Thoemmes & West, 2011) they did not address the full complexity involved in matching units across clusters. For instance, none of these studies varied both the data-generating selection model and the outcome model, or explicitly investigated situations of small samples or lack of overlap. Moreover, the published studies are not explicit about the assumptions that are required for identifying average causal effects with an across-cluster matching strategy. Since we systematically investigate and compare across-cluster matching to within-cluster matching, this study will guide researchers in selecting an appropriate matching strategy and matching technique.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

In our simulations we used a model with two level-one ($p = 2$) and two level-two covariates ($q = 2$). In using different coefficient matrices, we created target populations that differ in the degree of overlap, the heterogeneity of the selection and outcome models across clusters, and the complexity of the models (i.e., with or without interaction terms and cross-level interactions). Details about how we generated the target populations are given in Appendix B. In simulating repeated sampling from an underlying target population of clusters and units, we sampled 30

clusters (out of 500) and 40% of units within each sampled cluster. In each iteration of the simulation, we first estimated different PS models and then the treatment effect using different PS techniques.
Estimation of PS. In estimating the unknown PS we used several fixed and random effects models. We estimated three fixed effects model, the first only including the main effects of level-one and level-two covariates, the second adds cluster fixed-effect (i.e., dummy variable for clusters), and the third also includes all interaction terms of level-two covariates (among each other and with level-one covariates). In addition to the three fixed effects models we also estimated four random effects models: two random intercept models, one with level-one and level-two main effects only, the other with all interaction terms of level-two covariates; two random slopes models where the coefficients of the two level-one covariates were modeled as (i) simple random coefficient and (ii) as a linear function of level-two covariates plus a stochastic error term. The models are shown in the notes to Table 1 in Appendix B.
Matching Strategies. Using the different estimated PSs, we investigated two main matching strategies: within-cluster matching and across-cluster matching. For the within-cluster matching strategy we used both the jointly estimated PSs and the PS that was estimated for each cluster separately. In estimating the average treatment effect (ATE), we applied three PS techniques: optimal full matching on the PS-logit, PS stratification, and inverse propensity weighting (Rubin, 2006; Schafer & Kang, 2008). Further, combining PS adjustments with an outcome regression allows for an estimation of the average treatment effect via a fixed-effects or random effects outcome regression. We used two fixed effects models and three random effects models for estimating the treatment effect (see the notes to Table 1 in Appendix B).

Usefulness / Applicability of Method:
Demonstration of the usefulness of the proposed methods using hypothetical or real data.

The findings of this study guide researches in choosing an optimal matching strategy for their multilevel data at hand. Depending on their data (e.g., availability of level-two covariates, cluster sizes, and overlap of treatment and control cases within clusters) researchers might either consider within-cluster matching or across-cluster matching as their primary matching strategy. Moreover, the simulation results help to select the best performing combination of a matching strategy and analytic method. But our results also demonstrate under which conditions the different matching strategies and analytic methods break down, that is, do not successfully remove most of the selection bias.

Research Design:
Description of the research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).
(May not be applicable for Methods submissions)
NA

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)
NA

Findings / Results:
Description of the main findings with specific details.

(May not be applicable for Methods submissions)

Table 1 in Appendix B shows the estimates for the average treatment effect for a population of units and clusters that was generated using main effects, interactions effects, cross-level interaction effects, random slopes, and random intercepts in both models (the selection and outcome model). Moreover, a slightly heterogeneous treatment effect was modeled in generating the outcomes (i.e., a treatment by level-two covariate interaction was included in the outcome-generating model). Note that Table 1 shows the results only for a single simulated population, but our final study will cover results on other populations with different outcome and selection models or different cluster sizes as well.

The first three lines of Table 1 indicate that approximately unbiased estimates result for all three PS methods (weighting, stratification and matching) if the true PS is used. PS-based estimates are close to the true population effect of 10 points. Results are only "approximately" unbiased since weighting and matching estimators are consistent but not unbiased, and PS stratification with 5 strata removes only about 90% of the initial selection bias. Within-cluster matching (shown in the last column of Table 1) also results in nearly unbiased estimates.

While fixed-effects PS models [1] and [2] produce biased estimates, PS adjustments based on fixed-effects model [3] (which includes cross-level interactions) result in nearly unbiased effect estimates although the model does not account for variations in slopes across clusters. We obtain similar results for PS models estimated with random effects. If the PS model includes cross-level interactions or random slopes nearly unbiased effect estimates can be obtained.

The results in the first five columns of Table 1 represent the estimates of different outcomes models applied to the across-cluster matching strategy. The outcome models differ with respect to the fixed- or random-effects modeling of the intercept and the treatment effect (for details see the notes to Table 1). At least in this simulation, the choice of a specific outcome model—with random or fixed effects—had no significant effect on the point estimates of the average treatment effect (but it certainly has an effect on the standard errors which are not shown here).

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

The results indicate that matching approaches for causal inference need to reflect the multilevel structure of the selection process. If matching does not reflect the cluster-specific differences in the selection process biased effect estimates result. Whenever possible, a within-cluster matching strategy should be used since it rests on weaker assumptions than across-cluster matching. However, the sparseness of observations within clusters or the lack of overlap might force researchers to match across clusters. The simulation shows that an across-cluster matching strategy successfully removes approximately all the bias if the PS model is (approximately) correctly specified across clusters. However, these preliminary results will be complemented by further simulations that explore variations in the target populations (like degree of overlap or cluster sizes) but also different across-cluster matching strategies (e.g., strategies that only allow for an across-cluster matching if no close matches within a cluster can be found).

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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

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.

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.

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

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

Authors, (in press).

Appendix B. Tables and Figures

Not included in page count.

Generation of Simulated Target Populations

Generating potential control and treatment outcomes. In order to investigate different matching strategies, we simulated a population of approximately 150,000 units nested within 500 clusters with cluster-sizes between 250 and 350 units. For each unit we generated a set of covariates, a propensity score, a treatment status, and a set of potential outcomes. Similarly, for each cluster we computed a set of level-two covariates. Given a set of p level-one covariates $\mathbf{X}$ and q level-two covariates $\mathbf{W}$ (the generation of covariates is described in more detail below), we compute for all units $i = 1, \ldots, n_j$ of each cluster $j = 1, \ldots, J$ the potential control and treatment outcomes according to

$$\mathbf{Y}_j^0 = \mathbf{X}_j^0\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \quad \text{for the potential control outcomes, and}$$

$$\mathbf{Y}_j^1 = \mathbf{X}_j^1\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \quad \text{for the potential treatment outcomes,}$$

where $\mathbf{Y}_j^0$ and $\mathbf{Y}_j^1$ are the $(n_j \times 1)$ dimensional vectors of potential control and treatment outcomes, respectively $\mathbf{X}_j^0$ and $\mathbf{X}_j^1$ represent the corresponding $(n_j \times p_d)$-dimensional design matrices of predictors including the constant but also interaction terms of the treatment indicator with covariates and interaction terms among level-one covariates (thus, $p_d > p$). Design matrices $\mathbf{X}_j^0$ and $\mathbf{X}_j^1$ only differ with regard to the treatment indicator Z ($Z = 1$ indicates the treatment condition, $Z = 0$ the control condition); All predictors involving Z are set to zero in $\mathbf{X}_j^0$, thus $\mathbf{Y}_j^0$ is unaffected by treatment. $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a $(n_j \times 1)$-dimensional vector of normally distributed level-one errors with expectation zero and a diagonal matrix of homogeneous variances. The $(p_d \times 1)$-dimensional vector of level-one outcome coefficients $\boldsymbol{\beta}_j$ is computed as a linear combination of level-two predictors $\mathbf{W}_j^d$ plus an additive error term $\boldsymbol{\omega}_j$:

$$\boldsymbol{\beta}_j = \mathbf{W}_j^d\boldsymbol{\gamma}_j + \boldsymbol{\omega}_j,$$

where $\mathbf{W}_j^d$ is the $(p_d \times q_d)$-dimensional matrix of level-two predictors (including the constant, interaction effects and higher order terms), $\boldsymbol{\gamma}_j$ the $(q_d \times 1)$-dimensional vector of level-two coefficients, and $\boldsymbol{\omega}_j \sim N(\mathbf{0}, \boldsymbol{\Omega})$ is the $(p_d \times 1)$-dimensional vector of normally distributed level-two errors with expectation zero and a variance-covariance matrix $\boldsymbol{\Omega}$.

Generating propensity scores and treatment statuses. In order to create the units' true PS we use a logistic model that creates the cluster-specific logits of the PS as a linear combination level-one predictor ($j = 1, \ldots, J$):

$$\boldsymbol{\Lambda}_j = \mathbf{X}_j^s\boldsymbol{\delta}_j,$$

where $\boldsymbol{\Lambda}_j$ is the $(n_j \times 1)$ dimensional vector of PS logits and $\mathbf{X}_j^s$ the $(n_j \times p_s)$-dimensional design matrix of predictors in the selection model (including the constant but also interaction terms of the treatment indicator with covariates and interaction terms among level-one covariates). The

($p_s \times 1$)-dimensional vector of level-one selection coefficients $\boldsymbol{\delta}_j$ is determined as a linear combination of level-two predictors $\mathbf{W}_j^s$ and an additive normally error term $\boldsymbol{\psi}_j$,

$$\boldsymbol{\delta}_j = \mathbf{W}_j^s \boldsymbol{\eta}_j + \boldsymbol{\psi}_j ,$$

where $\mathbf{W}_j^s$ is the ($p_s \times q_s$)-dimensional design matrix of level-two predictors (including the constant, interaction effects and higher order terms), $\boldsymbol{\eta}_j$ the ($q_s \times 1$)-dimensional vector of level-two coefficients and $\boldsymbol{\psi}_j$ the ($p_s \times 1$)-dimensional vector of normally distributed level-two errors with expectation zero and a variance-covariance matrix $\boldsymbol{\Psi}$, i.e., $\boldsymbol{\psi}_j \sim N(\mathbf{0}, \boldsymbol{\Psi})$.

For each level-one unit the PS logit λ_{ij} is transformed into the PS π_{ij} according to $\pi_{ij} = 1/(1+\exp(-\lambda_{ij}))$. Finally, a random draw from a Bernoulli distribution with selection probability π_{ij} determines each unit's treatment status ($Z_{ij} \sim \text{Bernoulli}(\pi_{ij})$).

Generation of level-one and level-two covariates. Since we simulated level-one covariates that depend on level-two covariates we generated the level-two covariates first. For each cluster $j = 1, \ldots, J$, we randomly drew a ($1 \times q$)-dimensional vector of level-two covariates $\mathbf{W}_j$ from a multivariate normal distribution with expectation $\boldsymbol{\mu}_{\mathbf{W}}$ and a covariance matrix $\boldsymbol{\Sigma}_{\mathbf{W}}$ ($\mathbf{W}_j \sim N(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Sigma}_{\mathbf{W}})$). Then, for each unit ij, with $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, we randomly sampled a ($1 \times p$)-dimensional level-one covariate vector $\mathbf{X}_{ij}$ from a normal distribution, $\mathbf{X}_{ij} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, with cluster-specific expectation $\boldsymbol{\mu}_j = \mathbf{W}_j^x \boldsymbol{\varphi} + \kappa_j$ and a covariance matrix $\boldsymbol{\Sigma}_j$ with randomly determined variances and covariances. $\mathbf{W}_j^x$ is the ($p \times q_x$)-dimensional design matrix that includes the constant term, interaction and higher-order terms of $\mathbf{W}$, and $\boldsymbol{\varphi}$ represents the corresponding ($q_x \times 1$)-dimensional coefficient vector. $\kappa_j \sim N(\mathbf{0}, \mathbf{K})$ is a multivariate normally distributed error term with expectation zero and covariance matrix $\mathbf{K}$.

Table 1. Estimates of the Average Treatment Effect.

		across-cluster matching					within-cluster matching
		Outcome model					
PS model	Matching method	FE-Y[1]	FE-Y[2]	FE-Y[3]	RE-Y[1]	RE-Y[2]	
true PS	weighting	10.01	10.10	10.16	10.10	10.14	10.14
	stratification	10.32	10.31	10.35	10.31	10.35	10.13
	matching	10.15	10.17	10.22	10.17	10.20	10.13
FE-PS[1]	weighting	13.07	11.79	11.81	11.80	11.82	10.88
	stratification	12.52	11.62	11.67	11.63	11.67	10.87
	matching	12.27	11.47	11.51	11.48	11.51	10.88
FE-PS[2]	weighting	11.92	11.90	11.88	11.90	11.89	10.88
	stratification	11.98	11.85	11.86	11.85	11.87	10.88
	matching	11.75	11.72	11.73	11.72	11.74	10.87
FE-PS[3]	weighting	9.64	9.79	9.95	9.79	9.94	10.15
	stratification	10.25	10.27	10.40	10.27	10.39	10.16
	matching	10.07	10.11	10.23	10.11	10.22	10.14
RE-PS[1]	weighting	12.05	11.87	11.86	11.87	11.87	10.85
	stratification	12.09	11.84	11.84	11.84	11.85	10.84
	matching	11.83	11.70	11.69	11.70	11.70	10.84
RE-PS[2]	weighting	9.60	9.80	9.97	9.80	9.96	10.17
	stratification	10.16	10.21	10.35	10.21	10.35	10.17
	matching	9.94	10.03	10.18	10.03	10.16	10.15
RE-PS[3]	weighting	10.16	10.15	10.20	10.15	10.19	10.06
	stratification	10.44	10.33	10.36	10.33	10.35	10.06
	matching	10.19	10.12	10.16	10.12	10.14	10.05
RE-PS[4]	weighting	9.83	9.91	9.97	9.91	9.96	10.01
	stratification	10.22	10.16	10.20	10.16	10.19	10.01
	matching	9.95	9.95	10.00	9.95	9.98	10.00

Notes to Table 1.
(i) The true effect in the population is 10.

(ii) The "within-cluster matching"-column shows the results one obtains when the cluster-specific treatment effects are estimated using the PS from the joint multilevel model. Results where the PS is estimated for each cluster separately are not shown (but they are close to the within-cluster matching estimates using the true propensity score as shown in the first three lines).

(iii) PS models (the PS logit l_{ij} is modeled)
Fixed effects models (FE) with/without cluster dummies D_g:

FE-PS[1] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj}$

FE-PS[2] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj} + \sum_{g=1}^{J-1} \delta_g D_{gj}$

FE-PS[3] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj} + \sum_{h=1}^{2} \sum_{k=1}^{2} \eta_{hk} w_{kj} x_{hij} + \sum_{g=1}^{J-1} \delta_g D_{gj}$

Random effects models (RE):

RE-PS[1] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj} + \psi_{0j}$

RE-PS[2] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj} + \sum_{h=1}^{2} \sum_{k=1}^{2} \eta_{hk} w_{kj} x_{hij} + \psi_{0j}$

RE-PS[3] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj} + \sum_{h=1}^{2} \psi_{hj} x_{hij} + \psi_{0j}$

RE-PS[4] $l_{ij} = \eta_{00} + \sum_{h=1}^{2} \eta_{h0} x_{hij} + \sum_{k=1}^{2} \eta_{0k} w_{kj} + \sum_{h=1}^{2} \sum_{k=1}^{2} (\eta_{hk} w_{kj} + \psi_{hj}) x_{hij} + \psi_{0j}$

(iv) Outcome models

Fixed effects models (FE) with/without cluster dummies D_g:

FE-Y[1] $y_{ij} = \gamma_{00} + \gamma_{10} Z_{ij} + \varepsilon_{ij}$

FE-Y[2] $y_{ij} = \gamma_{00} + \gamma_{10} Z_{ij} + \sum_{g=1}^{J-1} \delta_g D_{gj} + \varepsilon_{ij}$

FE-Y[3] $y_{ij} = \gamma_{00} + \gamma_{10} Z_{ij} + \sum_{g=1}^{J-1} \delta_g D_{gj} + \sum_{g=1}^{J-1} \delta_g Z_{ij} D_{gj} + \varepsilon_{ij}$

Random effects models (FE):

RE-Y[1] $y_{ij} = \gamma_{00} + \gamma_{10} Z_{ij} + \omega_{0j} + \varepsilon_{ij}$

RE-Y[2] $y_{ij} = \gamma_{00} + \gamma_{10} Z_{ij} + \omega_{0j} Z_{ij} + \omega_{0j} + \varepsilon_{ij}$