Proposal for a symposium on “Matching Strategies for Causal Inference” at the SREE 2011 spring conference

Symposium title: Matching Strategies for Causal Inference

Symposium Organizer & Contact: Peter M. Steiner (psteiner@wisc.edu)

Chair: Thomas D. Cook (t-cook@northwestern.edu)

Conference Section: Research Methods

Symposium Justification:

Matching approaches like propensity score matching are frequently used in observational studies for estimating unbiased treatment effects. Though a huge body of literature emerged during the last decades on the theoretical foundations and practical implementation of individual case matching there is still a considerable lack of research on matching units in the context of multilevel data structures. Further, strategies for matching control groups can also successfully be applied to other quasi-experimental designs like regression discontinuity or interrupted time series designs. The suggested symposium tries to initiate more research on matching strategies for multilevel data which are common in educational research. The first paper by Peter M. Steiner demonstrates that the choice of a general matching strategy (i.e., at which level one should match) depends on the level at which treatment assignment/selection took place. For instance, if treatment selection is at the school level schools should be matched. If treatment selection is at the student level one should match students within schools. For the latter case he shows that biased effect estimates might result if students are matched between instead of within schools. Using ECLS-K data and student-level matching, Kelly Hallberg then investigates the importance of (i) multiple waves and (ii) different groups of pretest measures for removing selection bias. In particular, she focuses on the role of pretest and proxy-pretest measures on the outcome. Nathan Jones then presents a paper on school-level matching. In following Wilde & Hollister (2007), he uses Project STAR data in order to evaluate the role of local matching for estimating class size effects. However, in contrast to Wilde & Hollister he uses geographical coordinates for matching treatment and comparison schools. The treatment effect of the Project STAR randomized experiment is used for evaluating the success of local matching. Finally, Coady Wing explores how matching of an external control group to a regression discontinuity (RD) design might strengthen the analysis of the RD design. For instance, if the effect of an intervention is investigated by implementing a RD design in a sample of schools one could match control schools (where no intervention took place) to the RD schools and thereby increase the strength of the RD design.

Papers (in presentation order):

(i) Peter M. Steiner. Matching Strategies for Observational Data with Multilevel Structure. (contact: psteiner@wisc.edu)
(ii) Kelly Hallberg, Peter M. Steiner, & Thomas D. Cook. The Role of Pretest and Proxy-Pretest Measures of the Outcome for Removing Selection Bias in Observational Studies. (contact: kellyhallberg2013@u.northwestern.edu)

(iii) Nathan Jones, Peter M. Steiner, & Thomas D. Cook. Using Local Matching to Improve Estimates of Program Impact: Evidence from Project STAR. (contact: nathan-jones@northwestern.edu)

(iv) Coady Wing & Thomas D. Cook. How Can Comparison Groups Strengthen Regression Discontinuity Designs? (contact: coady.wing@gmail.com)

Affiliations of Authors:

Thomas D. Cook (t-cook@northwestern.edu), Northwestern University, Institute for Policy Research

Kelly Hallberg (kellyhallberg2013@u.northwestern.edu), Northwestern University, Institute for Policy Research

Nathan Jones (nathan-jones@northwestern.edu), Northwestern University, Institute for Policy Research

Peter M. Steiner (psteiner@wisc.edu), University of Wisconsin – Madison, Department of Educational Psychology

Coady Wing (coady.wing@gmail.com), Northwestern University, Institute for Policy Research

Discussant: Mark W. Lipsey (mark.lipsey@vanderbilt.edu) is willing to serve as a discussant.
Title: Matching Strategies for Observational Data with Multilevel Structure

Author(s): Peter M. Steiner
In educational research, causal inference from observational studies has gained in importance during the last decade—particularly propensity score (PS) techniques like PS matching or PS stratification are now more frequently used in estimating the effects of educational interventions. In comparison to other fields of research, observational studies in educational research typically face an additional challenge: the data usually show a multilevel structure: students are nested within classrooms and schools, or schools are nested within districts or states, for instance. Complications arise for several reasons: (i) Units within clusters are typically not independent; (ii) Interventions may be implemented at different levels (e.g., student-, classroom-, or school-level); (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. Therefore, the implementation of matching techniques for removing selection bias is more challenging than for data structures with a single level only.

In this study we consider a two-level structure where students are nested within schools. Thus, treatment assignment or selection might take place either at the school level or the student level. Treatment selection at the school level implies that the treatment status only varies between schools but is constant for all students within schools (all students of a school are either assigned to the treatment or control condition). On the other hand, if treatment selection takes place at the student level students might be assigned to or self-select into the treatment or control condition within each single school. Depending on the level of treatment selection, two main matching strategies are possible. First, if treatment selection is at the school level comparable treatment and control schools need to be matched—matching of individual students is not necessarily required. This type of matching mimics a cluster randomized controlled trial where schools are randomly assigned to treatment. Local and focal matching approaches that match geographically neighboring treatment and control schools of the same type is a promising strategy for obtaining unbiased school level treatment effects (Cook, Shadish & Wong, 2008). If school level matching results in matched schools that considerably differ on observed student level covariates an additional matching of students within matched schools might further increase the comparability of matched schools. Second, whenever treatment selection is at the student level students need to be matched within schools, thereby mimicking a randomized block design where students are randomly assigned to the treatment condition within schools (blocks). However, if extreme selection processes take place we might be confronted with a lack of comparable students within schools and, thus, be forced to look for matches from other schools (i.e., match between schools).

Though the popularity of matching approaches has considerably increased, only a few methodological papers on matching in the context of multilevel data are available (several of them unpublished): Hong & Raudenbush (2006) extend the Rubin Causal Model (Rubin, 1974) to the multilevel case and give an example on the effect of retaining students in Kindergartens. However, in estimating the retention effect on reading and math achievement scores, they matched retained and promoted students between schools and made no attempt to match students
within schools; similarly Hong, in press). Arpino & Mealli (2008), Kim & Seltzer (2007), and Thoemmes & West (2010) focused on student level matching and conducted simulation studies using different multilevel models for estimating propensity scores. Also these studies did not consider matching strategies that match students within schools— their suggested matching strategies allow for matches of students between schools.

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

Given the different possibilities of matching in the context of multilevel data and the lack of research on corresponding matching strategies, we investigate two main research questions. The first research question investigates the advantages and disadvantages of different matching strategies that can be pursued with multilevel data structures. The goal is first to outline possible matching strategies and then to identify an optimal matching strategy for different treatment selection scenarios (here, optimal refers to design aspects rather than technical aspects of a matching algorithm). In following Hong & Raudenbush (2006), theoretical foundations are discussed within the Rubin Causal Model framework and its potential outcomes notation (Rubin, 1974). The second research question focuses on the matching of students (when treatment is implemented at the student level) in more detail. As outline above, one can either match students within schools or match students between schools. Matching within school is time-consuming and might fail due to a lack of comparable treatment and control students within schools. Matching between schools, on the other hand, can be more conveniently implemented by estimating an overall PS model for all students together. This can be done using hierarchical linear modeling. Students can then be matched within and between schools. Thus, treatment and control students might be successfully matched even if no close matches are available within schools. The question then is whether and under which conditions we can get unbiased effect from such a matching strategy.

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

This study systematically investigates and compares matching strategies in the context of hierarchical data structures. In particular, it demonstrates that some of the matching strategies suggested by methodological papers may result in biased estimates. Another novel contribution of that study is that it discusses matching strategies for different scenarios of treatment selection (i.e., selection at different levels) and that it investigates the conditions under which less optimal strategies might lead to unbiased effect estimates.

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

We investigate the second research question (i.e., strategies for student-level matching) by conduction simulation studies and analyzing a real multilevel dataset (with a two level structure: students nested within schools). First, using a small simulated example we demonstrate that
matching between schools (instead of within schools) may result in considerably biased estimates of the treatment effect. Then, a more extensive simulation study that varies sample sizes, intraclass correlations, the complexity of both the selection process and data generating outcome model (which was not done in any of the simulation studies mentioned above), degree of group overlap, and the extent of initial covariate imbalance at each level is used to simulate more realistic scenarios for educational research. In the case of matching within schools a logistic regression model is estimated for each school in order to get the estimated propensity score. When we allow for matches between schools we estimate an overall PS model using hierarchical linear models.

The simulation study is complemented by a re-analysis of Hong & Raudenbush’s study on the effect of kindergarten retention on student achievement scores. While Hong and Raudenbush (2006) analyzed the data using a multilevel PS and allowed for matches between schools we test whether we get a different retention effect if we only allow for matches within schools.

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

The findings of the study help researches in identifying an optimal matching strategy for their data at hand. The results of the study also show that choosing a less than optimal strategy—a strategy that does not properly reflect the selection process—may not be able to reduce all the selection bias from the treatment effect of interest.

**Findings / Conclusions:**
Description of conclusions, recommendations, and limitations based on findings.

For hierarchical data structures, theoretical considerations and preliminary simulation results clearly indicate that matching approaches for causal inference need to reflect the (multilevel) selection process. If matching does not reflect the selection process that actually took place biased treatment effect may result. If selection takes place at the student level (within schools) one should match students within school. Matching students between schools may result in biased treatment effects. However, if treatment and control students cannot be matched within schools, matching students between schools might still be considered but stronger assumptions are required. If treatment selection takes place at the school level an individual-student matching is not necessarily required.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Title: The role of pretest and proxy-pretest measures of the outcome for removing selection bias in observational studies.

Author(s): Kelly Hallberg, Peter M. Steiner, & Thomas D. Cook
Abstract Body
Limit 5 pages single spaced.

Background / Context:
Description of prior research and its intellectual context.

Much of educational research is concerned with validly estimating causal effects. Experiments which employ random assignment are the preferred means for estimating causal effects, because, when implemented correctly, they assure that the treatment and control groups are equivalent in expectation on both observed and unobserved characteristics (Rubin, 1974). However, random assignment is not always ethical or feasible. In these cases, researchers must rely on quasi-experimental methods to identify causal effects (Shadish, Cook, & Campbell, 2002). Much debate has centered on whether and under what conditions these methods can produce valid estimates of causal effects (Glazerman, Levy, & Myers, 2003; Bloom, Michalopoulos, & Hill, 2005; Smith & Todd, 2005; Cook, Shadish, & Wong, 2008).

The primary challenge that all quasi-experimental studies face is the fact that without random assignment, the treatment and comparison groups can vary in both observable and unobservable from one another in ways that are independent of receipt of treatment (Rubin, 1974). Many quasi-experimental designs equate the treatment and comparison groups on observable covariates using matching or regression approaches. However, with the exception of regression-discontinuity designs, researchers can never be completely confident that they have adequately accounted for selection because they are not able to rule out the possibility that the groups vary on unobservable characteristics that are related to the outcome of interest (Rosenbaum & Rubin, 1983). The best quasi-experimental researchers can do is select observable covariates to minimize possible bias in effect estimates. For this reason, the selection of covariates in observational studies is of critical importance (Steiner, Cook, Shadish, & Clark, in press).

Two competing approaches to guiding covariate selection appear in the literature on quasi-experimental design. One, most closely associated with Donald B. Rubin (2007) and James Heckman, argues that covariate selection should be primarily concerned with modeling selection. The other, commonly associated with the work of Judea Pearl (2009), is more concerned with covariates that are associated with the outcome.

Within this larger debate sits the more pragmatic question of whether a pretest of the outcome measure plays a special role in reducing bias in observational studies. Some researchers, most notably those associated with the Campbell tradition of causal inference, have placed special emphasis on the pretest, primarily because of its correlation with the outcome (Campbell, 1957; Campbell & Stanley, 1963; Shadish, Cook and Campbell, 2002). While others, including Rubin (2007) and Cronbach (1982), give the pretest no special emphasis, stressing instead compiling a composite of variables that explain selection.

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

The purpose of this paper is threefold. The first is to test whether the pretest plays a greater role in bias reduction than any other single covariate, which we predict it will. The second is to
examine the marginal improvement in bias reduction offered by having two pretest measurement waves. We predict that there will be some marginal gain in bias reduction as a result of including an additional pretest wave. The third purpose is to examine the extent to which a proxy pretest measure can substitute for a real pretest whose form is invariant between pretest and posttest.

**Significance / Novelty of study:**

*Description of what is missing in previous work and the contribution the study makes.*

Within study comparisons of experimental and quasi-experimental results suggest that the quality of covariates in observational studies matters and that there may reason to think that the pretest plays a unique role in these studies. However, the conclusions are drawn by comparing approaches that are employed across within-study comparisons rather than by examining the importance of certain kinds of covariates within the same study. This paper contributes to the literature by undertaking this task in a case study using a single dataset to examine how true pretest measures of the outcome perform in reducing bias in an observational study compared to other predictors.

**Statistical, Measurement, or Econometric Model:**

*Description of the proposed new methods or novel applications of existing methods.*

We conduct a secondary data analysis on the data set used by Hong and Raudenbush (2005) to examine the effect of kindergarten retention on children’s cognitive development in reading and mathematics. In this case, there is no randomly assigned experimental benchmark. This is a substantial limitation which means that we remain unable to determine whether any of the methods employed would replicate an experimental benchmark. Hong and Raudenbush (2005) conducted a thorough analysis of the data using a rich set of covariates making the strong ignorability assumption plausible. However, we cannot rule out the possibility that residual statistical bias remains due to the fact that the matched non-retention students come from the lower end of the distribution of non-retained students.

Instead of using an experimental benchmark, this study draws on our understanding of the selection process of kindergarten retention and a comparison of various effect estimates. We know that students are primarily retained “to remedy inadequate academic progress and to aid in the development of students who are judged to be emotionally immature” (Jackson, 1975, p. 614). Alexander, Entwisle, and Dauber (2003) show that prior to retention, future retainees lag behind their non-retained peers on demographic, academic and social predictors of academic success. These predictors of selection into the treatment are likely to be negatively correlated with future academic performance and thus the naïve, unadjusted comparison of retained and non-retained students is likely to be negatively biased. This is supported by the past research on retention which consistently finds a negative bias in the unadjusted mean differences between the two groups (Alexander, R. Entwisle, & Dauber, 2003; Karweit, 1999). For this reason, we view an estimate as more realistic based on how much it reduces this negative bias.

Hong and Raudenbush provided the subset of the ECLS-K used in their initial analysis as well as information about the covariates included in their propensity score models. In their original papers, Hong and Raudenbush examine several different effect estimates to fully understand the various effects of retaining students in kindergarten. For the purpose of this paper, we focus
solely on estimating the effect of retention on retained students. Following Hong and Raudenbush’s lead, the ECLS-K data for this analysis were limited to schools in which at least some students are retained in kindergarten. The resulting data set included 10,726 students in 1,080 schools.

**Best fitting model using all of the covariates.** In any evaluation of a policy where participants are not randomly assigned to treatment or control, the problem of selection is paramount. We know that students who are retained are different from those that are not retained. These differences can bias estimates of program effectiveness (Shadish, Cook, & Campbell, 2002).

To address this issue, Hong and Raudenbush employed a multi-level propensity score approach. Propensity scores allow for modeling the probability that a given student participates in the program based on available observed characteristics. Because characteristics of the students themselves and their schools are likely to influence whether a student participates in the program, a hierarchical modeling approach was employed to calculate an individual-level propensity score for each student in the data set, denoted as \( q \):

\[
\hat{q} = \Pr(z_i = 1 | D_j = 1, X_{ij}, W_j, u_j)
\]

Where \( q \) is the conditional probability that student \( I \) in school \( j \) is retained as a function of his or her individual and school characteristics, \( X_{ij} \) and \( W_j \) respectively, and the residual random effect of school \( j \).

We employed a similar approach, starting at first with all 144 pre-treatment covariates as candidates for inclusion in the propensity score model. To select an initial propensity score model, we began by regressing each of the covariates on kindergarten retention. All covariates with a p value of greater than .2 were then included in a forward stepwise regression function to produce an initial propensity score model. Propensity scores and propensity score logits were then estimated using this model. We examined overlap in the treatment and comparison groups and deleted non-overlapping cases. We then looked at balance across the two groups on all 144 covariates. Balance statistics (standardized mean differences and variance ratios) were used to guide model selection.

We balanced pretreatment group differences in observed covariates using a propensity score stratification and marginal mean weighting approach. The propensity score logit was also included in the outcome model to control for within strata differences. Student outcomes were modeled in using two-level hierarchical linear models to account for the nested nature of the data (students within schools).

\[
y_{ij} = \gamma_0 + \delta_z Z_{ij} + \gamma_1 (\text{Logit}_q) + \sum_{s=2}^{15} \alpha_s L_{sij} + \gamma_2 (\text{Dur}_F)_{ij} + u_{0j} + u_{1j} Z_{ij} + e_{ij}
\]

Where \( y_{ij} \) is the reading or math score for child \( I \) in school \( j \) at the end of first grade, \( \delta_z \) is the average retention effect on retained students, \( L_{sij}, s = 2, \ldots, 15 \) are the dummy indicators for the propensity score strata, \( \text{Logit}_q \) is an additional adjustment for the student’s propensity of being retained, and \( \text{Dur}_F \) indicates the length of time since the beginning of the treatment year that
passed before the student was assessed. We chose to estimate the average effect of treatment on retained students rather than the average effect of retention on at risk students, as Hong and Raudenbush did, because this is generally the substantive estimate of interest in program evaluations.

**Sub Setting the Available Covariates.** Some 144 covariates were available in the ECLS-K for possible inclusion in the propensity score. We initially subdivided these covariates into three primary group: pretests (pre-intervention data on the same measure as the outcome), proxy pretests (teachers ratings of students’ academic performance), and all other covariates. The “all other covariates” group was later subdivided into eight additional groups: child demographics, child social skills, classroom demographic composition, classroom learning environment, home environment, school structures and supports, school demographic composition, and teacher demographics.

Each group of covariates was then included separately in the propensity score model. The initial propensity score model for each set of covariates was created using stepwise regression as described above. The model was then finalized by examining balance statistics to determine whether additional covariates (from that group of covariates), higher order terms, and interactions should be included. Once the propensity score model was finalized, the predicted values were used to model the effect of kindergarten retention on mathematics and reading achievement using the procedure described above. Effect estimates were then compared with one another. Bootstrap standard errors were calculated to allow us to determine whether the estimates varied significantly.

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

This paper will provide valuable insight to applied researchers grappling with covariate selection in observational studies or trying to determine whether propensity score matching is a valid strategy given the available covariates. It will also provide guidance to consumers of research in their assessment of whether the appropriate variables were included in observational studies.

**Findings / Results:**
*Description of the main findings with specific details.*
(May not be applicable for Methods submissions)

Table 1 shows the effect estimates for each set of covariates. The results support the notion that the true pretest of the outcome plays a special role in bias reduction compared to other covariates as well as the hypothesis that two pretest waves is preferable to one. However, the propensity score model with two proxy pretest waves performed almost as well.

The propensity score that included all covariates except the pretests and proxy pretests did not do as well as the pretest and proxy pretest models. However, this model was able to reduce a fair amount of the bias. The researchers hypothesized that this is a result of the particularly rich set of other covariates available in this data set which spanned multiple domains of child development. When the group of all non-pretest and proxy pretest covariates was further subdivided into
domain specific subgroups, no single sub groups performed as well as all of the other covariates combined (see Table 2).

Table 3 shows the correlation of the propensity scores created with each sub set of covariates with both selection into treatment and the outcome. By examining these relationships we are better able to understand why some groups of covariates perform better or worse than others.

**Conclusions:**

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

This case study provides a unique opportunity to examine the relationship between covariate selection and bias reduction in observational research. Past research has suggested the importance of covariate selection and the importance of certain types of covariates, such as the true pretest of the outcome. However, this study is somewhat unique in its comparison of the performance of different sets of covariates in the same study. From this study, we can glean initial insights into how much bias certain types of covariates can reduce in observational studies. However, as a case study, the issue of generalizability remains at the forefront. It is unclear the extent to which the findings from this study generalize more broadly.
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


### Table 1. Effect Estimates from Covariate Subsets

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th>All Covariates</th>
<th>One Pretest</th>
<th>Two Pretests</th>
<th>First Pretest and Slopes</th>
<th>One Proxy Pretest</th>
<th>Two Proxy Pretests</th>
<th>All Other Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.38)</td>
<td>(0.43)</td>
<td>(0.43)</td>
<td>(0.38)</td>
<td>(0.40)</td>
<td>(0.48)</td>
<td>(0.39)</td>
<td>(0.46)</td>
</tr>
<tr>
<td><strong>Math</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-11.86</td>
<td>-5.21</td>
<td>-6.23</td>
<td>-4.92</td>
<td>-5.03</td>
<td>-8.37</td>
<td>-5.02</td>
<td>-6.91</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.32)</td>
<td>(0.30)</td>
<td>(0.29)</td>
<td>(0.28)</td>
<td>(0.34)</td>
<td>(0.29)</td>
<td>(0.32)</td>
</tr>
</tbody>
</table>

### Table 2. Effect Estimates with Domain-Specific Sets of Covariates Included

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Math</th>
</tr>
</thead>
<tbody>
<tr>
<td>All covariates</td>
<td>-9.06</td>
<td>-5.21</td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(0.32)</td>
</tr>
<tr>
<td>All other covariates</td>
<td>-12.45</td>
<td>-6.91</td>
</tr>
<tr>
<td></td>
<td>(0.46)</td>
<td>(0.32)</td>
</tr>
<tr>
<td>Child demographics</td>
<td>-17.19</td>
<td>-10.39</td>
</tr>
<tr>
<td></td>
<td>(0.51)</td>
<td>(0.35)</td>
</tr>
<tr>
<td>Child social skills</td>
<td>-13.87</td>
<td>-7.69</td>
</tr>
<tr>
<td></td>
<td>(0.49)</td>
<td>(0.34)</td>
</tr>
<tr>
<td>Classroom demographic composition</td>
<td>-19.50</td>
<td>-11.87</td>
</tr>
<tr>
<td></td>
<td>(0.55)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>Classroom Learning</td>
<td>-18.18</td>
<td>-10.93</td>
</tr>
<tr>
<td></td>
<td>(0.56)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>Environment</td>
<td>-17.87</td>
<td>-10.88</td>
</tr>
<tr>
<td></td>
<td>(0.52)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>Home environment</td>
<td>-20.38</td>
<td>-12.16</td>
</tr>
<tr>
<td></td>
<td>(0.55)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>School structures and supports</td>
<td>-20.14</td>
<td>-12.03</td>
</tr>
<tr>
<td></td>
<td>(0.55)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>School demographic composition</td>
<td>-20.14</td>
<td>-12.17</td>
</tr>
<tr>
<td></td>
<td>(0.55)</td>
<td>(0.36)</td>
</tr>
<tr>
<td>Teacher demographics</td>
<td>-20.14</td>
<td>-12.17</td>
</tr>
<tr>
<td></td>
<td>(0.55)</td>
<td>(0.36)</td>
</tr>
</tbody>
</table>

### Table 3. Propensity Score Correlation with Selection and Outcomes

<table>
<thead>
<tr>
<th>Propensity score with...</th>
<th>Correlation with Selection</th>
<th>Correlation with the Reading Outcome (Control Group Only)</th>
<th>Correlation with Math Outcome (Control Group Only)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All pretests</td>
<td>0.59</td>
<td>-0.30</td>
<td>-0.29</td>
</tr>
<tr>
<td>All proxy pretests</td>
<td>0.61</td>
<td>-0.27</td>
<td>-0.26</td>
</tr>
<tr>
<td>All covariates except pretests and proxy pretests</td>
<td>0.59</td>
<td>-0.20</td>
<td>-0.21</td>
</tr>
</tbody>
</table>
Title: Using Local Matching to Improve Estimates of Program Impact: Evidence from Project STAR.

Author(s):

Nathan Jones  
Northwestern University

Peter Steiner  
University of Wisconsin--Madison

Tom Cook  
Northwestern University
Background / Context:
Description of prior research and its intellectual context.

Experimental designs are widely considered the preferable method of validly determining a program’s causal impacts, primarily because, through the use of random assignment, experiments produce a reliable counterfactual. There will likely be cases where experiments are either not feasible or not desirable, and in such situations, alternatives to experiments will always be needed.

To determine the kinds of observational studies that are most likely to lead to unbiased results, researchers have increasingly conducted within-study comparisons, in which non-experimental designs are evaluated by comparing their causal estimates of program impacts to the estimates drawn from an experiment (e.g., LaLonde, 1986; Heckman, Ichimura, & Todd, 1997; 1998; Heckman et al., 1998; Smith & Todd, 2005). As reviews of the within-study literature in job training have shown, although observational studies frequently produce causal estimates that are different than those produced by experiments, there are characteristics of observational studies than can improve estimates; these include having a pretest measure of the outcome variable, having a local comparison group, and similar measures across the experimental and counterfactual groups (Bloom, Michalapoulos, & Hill, 2005; Glazerman, Levy, & Myers, 2003).

More recently, Cook, Shadish, and Wong (2008) compared three conditions under which observational studies may produce comparable causal estimates – regression discontinuity (RD) designs, intact matching from maximally similar groups, and using statistical procedures (e.g., propensity scores, OLS regression, instrumental variables) to attempt to make groups equivalent. Of the three kinds of studies, RD designs and intact groups appeared to reduce the most bias; and, in cases where the selection process was completely known, methods relying on statistical procedures also performed well. As Cook et al. suggest, it is in cases where researchers rely on “off-the-shelf” covariates that these studies typically fail to reduce bias.

In the educational context, a recent prominent example of a within-study comparison of experimental and quasi-experimental designs is Wilde and Hollister's (2007) analysis of data from the Tennessee class size experiment (Project STAR), in which kindergarten classrooms were randomly assigned to either regular or small class sizes to determine whether small classes affects student achievement. The authors evaluated the impact of estimates derived from propensity score matching relative to the study's experimental results (i.e., students from treatment classrooms were matched to students in control classrooms in the other 79 schools in the sample), concluding that the propensity scores do not reliably produce estimates of the “true” impact of small classes.

Given what we know from previous within-study comparisons, these results are not entirely surprising. In the absence of pre-test information on students in the control and treatment classes, Wilde and Hollister were limited to making comparison groups based on extant data on student, teacher, and school characteristics, leaving open the possibility that there were one or several unmeasured covariates that were correlated to selection and to the outcome. Further, rather than
matching based on intact local comparison groups, the authors matched at the student-level. As Cook et al. show, intact group matching will almost always reduce initial selection differences more than individual matching based on covariates.

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

In this study we test whether matching using intact local groups improves causal estimates over those produced using propensity score matching at the student level. Like the recent analysis of Wilde and Hollister (2007), we draw on data from Project STAR to estimate the effect of small class sizes on student achievement. We propose a strategy for intact group matching in which we match treatment cases to control cases in other schools.

A secondary goal of this analysis is to determine whether the use of geographic covariates (including latitude and longitude, as well as Census variables) improve the quality of matches over the use of student, teacher, and school covariates alone. We hypothesize that by incorporating physical distance into our matching, we will increase our likelihood of finding maximally similar comparison classrooms.

**Setting/Population**
*Description of the research location.*

The setting for this study is a sample of 79 kindergarten classrooms in the state of Tennessee. The Project STAR data was collected beginning during the 1985 school year, although longitudinal data collection continued on students through fourth grade. In total, over 7,000 students participated in the experiment.

For the purpose of comparison with Wilde and Hollister’s (2007) study, we restrict our treatment sample to the 11Project STAR schools with over 100 kindergartners, which ensures that there are multiple treatment classrooms per school and allows for 11 separate experimental and observational causal estimates; a pooled analysis is also conducted to determine the average treatment effect across the sample. In constructing the counterfactual groups, treatment schools were matched not just to the 11 comparison classrooms, but to all other 78 schools in the sample.

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

Students in participating schools were randomly assigned to one of three settings: the treatment condition (small classrooms with 13-17 students) or one of two control conditions (regular size classrooms or regular classrooms with an instructional aide). Past studies have shown that students in the small classes generally outperform their peers in large classes (Krueger, 1999; Word et al., 1990).

**Significance / Novelty of study:**
*Description of what is missing in previous work and the contribution the study makes.*
We provide a method of matching that is likely to reduce the bias associated with previous propensity score matching techniques. For one, our preliminary findings show that matching at the school level (while maintaining intact classrooms) is a potentially useful method for obtaining less biased results from an observational study. And, further, our results suggest that, in the absence of pretest information, the inclusion of geographic information as a covariate can improve the quality of matches obtained.

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

We match schools using Mahalanobis distance metric matching, as developed by Rubin and Cochran (Cochran & Rubin, 1973; Rubin, 1976, 1979). We restrict the covariance matrix to treated schools, given that the schools that serve as controls change depending on the covariates used in calculating Mahalanobis distances. Similarly, when the analyses are conducted at the classroom level, we use the covariance matrix of only treated classrooms within the 11 schools with over 100 students.

To find matches, schools are first randomly sorted, and the distances are calculated between each “treatment” school and all available control schools. In finding matches that minimize the Mahalanobis distance, we allowing for replacement; i.e., control schools are allowed to be used more than once.

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

As described in the section on the setting/population used in our research, we demonstrate the usefulness of the method using the Project STAR data. We have chosen this dataset for a number of reasons. For one, our results expand on the within-study comparison conducted by Wilde and Hollister (2007). Additionally, Project STAR is well-known within the educational research community as being a strong example of a large-scale experiment that has shown its intervention (class size) to be effective. Lastly, the fact that all schools are drawn from a common state is advantageous for us, given that we use geographic distance as a matching strategy.

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

As suggested above, in the application of our matching, we introduce a series of models that incorporate additional covariates at each step. We begin by using our distance measures alone, i.e., latitude and longitude, which provides a match that is geographically closest to the “treatment” school. We then conduct the same analyses using only the STAR covariates: % free and reduced lunch, % minority, % special education, % repeats, and % pullouts (all of which are aggregated to the school level). Next, we run a model with both distance and the STAR covariates. Finally, we incorporate a series of Census variables to the existing model to determine whether they improve further on our matches. These variables include: median
income, percent vacant homes, percent families, percent manufacturing/professional positions, percent high school graduates, and percent with graduate/professional degrees.

Based on these various matching strategies, we estimated the treatment effect by calculating mean differences between students in treatment classrooms and their matched comparison cases in each of the 11 schools. Additionally, when conducting the pooled analysis we used hierarchical linear modeling, treating students as nested within schools.

Findings / Conclusions

Initial results suggest that matching while maintaining intact groups offers advantages over previous methods for matching at the student level. We can also reduce bias further when using geographic distance to supplement the “off-the-shelf” covariates that were collected at the time of the Project STAR study.

Our results should be put into context however. There are other strategies that remain preferable to intact local matching when attempting to reduce bias in observational studies. Where are method is useful is in situations where the data available is similar to that of Project STAR, i.e., information is missing on students’ pretest information and access to covariates is limited. In these cases, we believe we have provided a potentially useful improvement over student-level matching.

* The Census variables are taken from the 1980 Census. In 1980, only households in urban areas (>50,000) have census tracts, block groups, and blocks attached to them; it wasn't until 1990 that all areas were assigned blocks. Because we did not want to lose data on non-urban areas, our Census variables are aggregated at the city level (which means that schools within the same FIPS-MCD have the same aggregate values for each of the housing and community variables).
Appendices
Not included in page count.

Appendix A. References
References are to be in APA version 6 format.


Title: How Can Comparison Groups Strengthen Regression Discontinuity Designs?

Author(s): Coady Wing and Thomas D. Cook
Abstract Body

Limit 5 pages single spaced.

Background / Context:
Description of prior research and its intellectual context.

The regression discontinuity design (RDD) is a valid basis for causal inference under statistical and behavioral assumptions that often are plausible and at least partly testable. In applied work, the RDD has some well-known limitations. One critique is that the RDD identifies treatment effects only for the narrow sub-population of units defined by the cut-off value of the assignment variable. This limitation has practical importance when the cut-off sub-population is not the population that matters for decision-making. The RDD typically does not provide a credible basis for extrapolation from the cut-off sub-population to other sub-populations or to the general population.

A second weakness of the RDD is that unbiased estimates of treatment effects depend on functional form assumptions that describe the relationship between the outcome and the assignment variable and sometimes between the treatment variable and the assignment variable. Flexible estimation methods, such as local linear and polynomial regression and global polynomial series regressions, allow researchers to avoid this problem in very large samples. Although these methods are logically appealing, in applications without very large sample sizes they may lack the statistical power to differentiate between alternative functional forms.

A third and related problem arises because RD based estimates are less statistically efficient than classical randomized control trials, primarily because of the need to estimate both treatment effects at the cut-off and the effect of the assignment variable near the cut-off.

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

In this paper, we examine some of the ways that different types of non-equivalent comparison groups can be used to strengthen causal inferences based on RDD. First, we consider a design that incorporates pre-test data on assignment scores and outcomes that were collected either before the treatment became available or before the practice of assigning treatments based on a cut-off score began. The idea is to use these pre-test data to establish a baseline estimate of the relationship between the outcome variable and the assignment variable.

Second, we evaluate a design that incorporates data on the assignment scores and outcomes of a single contemporaneous comparison group of units that are always ineligible for treatment. Here the idea is to establish a baseline differences in the relationship between outcomes and assignment scores that prevails in the RD group and the comparison group.

Third, we consider how unit level and group level covariates might be used to form an optimal control group from a pool of several candidate control groups. We explore how various methods of matching and reweighting can be used to construct a control group in which the functional relationship between the outcome and the assignment score closely resembles the relationship that prevails in the RD sample below the assignment cut-off value.
In all three cases, we evaluate the statistical and behavioral assumptions that are required for the comparison group augmented RDD to produce unbiased estimates of specific treatment effects of interest. We also compare the assumptions required to extrapolate from the cut-off sub-population to other sub-populations in the augmented RDD with the assumptions required to extrapolate from standard RDD.

**Research Design:**
*Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).*

We describe several control group augmented regression discontinuity designs, which could be used in applied work in education, economics, political science, and public healths. The method applies to settings in which a standard RDD is feasible and there is also data from a pre-intervention period or from one or more candidate comparison groups of units that are ineligible for treatment at any level of the assignment score.

**Conclusions:**
*Description of conclusions, recommendations, and limitations based on findings.*

Under specific assumptions, a comparison group can provide a firmer basis for extrapolation to other sub-populations, improve the credibility of functional form assumptions, and increase the statistical efficiency of the estimated treatment effects.