Abstract Title Page

Not included in page count.

Title: Intact school matching in education: Exploring the relative importance of focal and local matching.

Authors and Affiliations:

Vivian C. Wong
Currie School of Education
University of Virginia

Kelly Hallberg
Department of Human Development and Social Policy
Northwestern University

Thomas D. Cook
Institute for Policy Research
Northwestern University

Abstract Body

Background / Context:

Description of prior research and its intellectual context.

The nested data structure inherent in education (i.e. students nested in schools nested in districts) makes intact school matching an appealing approach in observational studies of educational interventions and policies for both theoretical and practical purposes. From a theoretical perspective, intact group matching seeks to minimize the difference between the treated and untreated populations while maximizing overlap on key observable characteristics. Comparison schools selected on the basis of similar pretreatment achievement or because they are in the same district as the treatment school may still vary on observed characteristics, but the total bias will likely be less than if students from intervention schools are matched with individual students from a different, less similar population (Cook, Shadish, & Wong, 2008). From the perspective of practice, school level matching can also be efficient. Applied education researchers often try to prospectively match schools using extant school level data to identify schools that are similar to those implementing the treatment. Researchers can then gather outcome data as well as data on implementation in both treatment and matched comparison schools.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

This paper provides guidance to applied education researchers who are employing intact school matching. First, using a within study comparison (WSC) methodology we examine whether intact school matching is able to replicate experimental results. In addition, we examine what approaches to intact school matching lead to the greatest level of bias reduction. Specifically, we estimate the ability to reproduce RCT results when comparison units are matched:

- ***Aim 1.*** On all school level characteristics (school composition, geography, and past academic achievement – each set of covariates alone and in combination with the others)
- ***Aim 2.*** Within school district.
- ***Aim 3.*** Using an approach (described below) that prioritizes local matches unless the schools are too divergent on observable characteristics

Setting:

Description of the research location.

The study draws on data from two RCTs conducted in education and data on other schools and students in the states in which the RCTs were conducted. The RCTs were conducted in Tennessee and Indiana and include schools and students from across the state in both cases.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features, or characteristics.

Data from schools (and students attending those schools) that participated in the two RCTs were used to calculate the experimental benchmark. In the case of the Indiana RCT, the quasi-experimental comparison group was drawn from all other schools in the state that did not participate in the RCT. In the case of the Tennessee RCT, the comparison group was constructed from within the RCT using a synthetic WSC design.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.

The Indiana RCT was designed to examine the effect of Indiana's benchmark assessment system on student achievement as measured by the state's annual Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) measures. Schools assigned to the treatment condition implemented regular formative assessments to students. From these assessments, teachers received immediate feedback on student performance that could be disaggregated in a variety of ways to inform instruction. In the Tennessee RCT, kindergartners and their teachers were randomly assigned within school to one of three conditions: small classes (13-17 students), regular size classes (22-25 students), and regular/aide (22-25 students) where the support from the teacher was supplemented by a full time aide.

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

Guidance from past within study comparisons suggests that intact school matching is most effective when schools are matched on important correlates of selection and the outcome (especially pretest measures of the outcome) and when matched schools are geographically proximal (Cook, Shadish, & Wong, 2008). The intuition behind the importance of matching on observable school-level characteristics is analogous to the logic underpinning individual case level matching: to increase the plausibility of the strong ignorability assumption, the goal is to obtain balance on important predictors of selection and the outcome across the treated and comparison cases. This approach can only address observable characteristics of the schools and must assume that all of the important correlates of selection and the outcome have been measured. The logic behind local matching is that schools that are geographically proximal are often similar in both observable and unobservable ways. Schools within the same school district, for example, often have a similar observed percent of students who qualify for free and reduced price lunch and are also similar in unobserved ways, such as district policies, community perceptions of schools and the importance of schooling, and the labor markets which graduates of the public schools will enter. Matching within district can be seen as analogous to including a district fixed effect in an OLS regression in that it rules out confounding that results from a correlation of the district level error term and treatment.

In practice, however, as a result of the relatively small sample of schools within a school district, researchers may face a tradeoff between matching on observables and finding a geographically proximal match (Stuart & Rubin, 2008). That is, applied researchers find themselves in a situation in which they could match a treatment school to a school that is nearby, but varies on a variety of observable characteristics, or to a school from a district in a different part of the state that looks more similar on observable characteristics, but may vary in unmeasured ways as a result of being geographically distant. Thus, a goal of the proposed line of research is to provide researchers with empirical guidance on the relative importance of matching on observable characteristics and finding geographically local matches.

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

The means for achieving the study goals was to conduct two within study comparisons (WSC): one using the Indiana RCT as the causal benchmark and one using the Tennessee RCT as the benchmark. Within-study comparisons estimate the extent of bias remaining in non-experimental causal studies after attempts either to select non-experimental comparison groups as similar as possible to the treatment group or after various matching or regression techniques have been applied to adjust for observed group differences. Each dataset was analyzed independently to examine how bias reduction is affected by school level matching on a set of school level covariates (all school-level covariates, school composition, geography, and past academic achievement) and using within district matching.

To implement the intact school matching on observable covariates (Aim 1), we employed a propensity score matching approach. Each treatment school was matched with the school with the closest propensity scores. We implemented the matching with replacement, so that a given comparison school could serve as a match to multiple treatment schools. We implemented the match with and without a caliper of 0.25 standard deviations of the propensity score to examine the effect of more or less coarse intact school matches. Treatment on the treated weights were used in all of the analyses using these matched sets of schools.

To implement within district matching (Aim 2), the samples were limited to only those treatment schools that had at least one comparison school in the same school district. The comparison schools in the same district served as the match for the treatment schools in district. Treatment on the treated weights were again implemented to ensure that the estimates using this approach are analogous both to those from the other matching approaches and the RCT benchmark.

Finally, we implemented a matching approach in which schools were matched within district if they were not too divergent on observable characteristics and to schools outside of the district if suitable within district matches are not available (Aim 3). To implement this approach, we first calculated a propensity score for each school using the best available set of school level covariates. We then matched each school to closest comparison schools within district using several caliper values. For example, if the caliper was set at .25 standard deviations of a propensity score, we would match the treatment school to schools that had the closest propensity score as long as they do not differ by more than .25 standard deviations on the propensity score. If a within district comparison school that met this criterion was not available, we took the closet match on the propensity score from any school in the state until each treatment school had a matched comparison schools. In initial analyses (presented below) we implemented this strategy using a caliper of 0.70. In future work we will implement this strategy using various other caliper rules. By changing the caliper rules we implicit change the preference given to local/within district matches relative to matches that more closely balance on observable characteristics. Implementing this approach will provide empirical guidance for understanding the tradeoff between local matches and those with greater similarity on observable characteristics.

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

This study provides evidence of the performance of each of these approaches to intact school matching drawing on data from two empirical WSCs: one using the Indiana Formative Assessment System RCT as the causal benchmark and the other using the Tennessee STAR RCT for this purpose.

Findings / Results:

Description of the main findings with specific details.

Table 1 below presents the results for the Tennessee dataset. Estimated effects, standard errors, and t statistics are presented for the RCT benchmark as well as each of the matching approaches. The final column of the table presents a simple difference between the RCT benchmark and each of the quasi-experimental estimates. While this measure provides an easy way to compare each of the approaches, the meaning of these estimates should be interpreted with some caution as both the RCT and the quasi-experimental point estimates are estimated with error.

(Insert Table 1 here)

The matching approach in which schools were matched within district if they were not too divergent on observable characteristics and to schools outside of the district if suitable within district matches are not available (referred to as the "super team" approach in the table) produced an effect estimate that was most similar to the experimental benchmark. The bias associated with the within district match and the observable covariates were of roughly the same magnitude but in the opposite direction. Matching only on geographic covariates (i.e. urbanicity, latitude and longitude) performed worst suggesting that matching on geographic proximity along while not accounting for policy structures (e.g. school districts) or observable covariates is insufficient for bias reduction.

Tables 2 and 3 present the results for the Indiana data for the reading outcome and Tables 4 and 5 present the analyses for the math outcome. It is important to note the relevant RCT benchmark for the within district match differs from the other models because some of the schools in the sample were the only elementary school in their district. We again see that matching on geographic covariates performs poorly. Matching on school level pretest measures of the outcome (not a possibility in the Tennessee dataset), most closely replicates the experimental benchmark. The other matching approaches come relatively close to the experimental benchmark, but given the large standard errors it is difficult to distinguish a consistent pattern of performance across the estimates.

(Insert Tables 2-5)

Conclusions:

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

This paper provides preliminary evidence that intact school matching can replicate experimental estimates and provides some guidance for which approaches are more likely to effectively reduce bias.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies often produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724-750.

Stuart, E.A. and Rubin, D.B. (2008). Matching with multiple control groups and adjusting for group differences. *Journal of Educational and Behavioral Statistics* 33(3): 279-306.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Results from Tennessee Dataset

	Treatment Effect	Standard Error	T-stat	Bias
RE Benchmark	8.08	3.41	2.37	n/a
School composition	6.45	6.07	1.06	-1.63
Geography & urbanicity	1.49	5.11	0.29	-6.59
All school level covariates	10.28	7.60	1.35	2.20
Within district	6.06	4.84	1.25	-2.02
Super Team	8.26	5.74	1.44	0.18

Table 2. Results from Indiana Dataset – Reading Outcome

	Treatment Effect	Standard Error	T-stat	Bias
RE Benchmark	1.97	3.58	0.55	n/a
School composition	4.11	4.09	1.00	2.14
Geography & urbanicity	-2.82	3.87	0.73	-7.07
Pretests	1.82	2.78	0.65	-0.15
All school level covariates except pretests	-0.33	3.71	0.09	-2.30
Complete	0.14	2.59	0.06	-1.83
Super Team	-4.91	5.23	0.94	-6.88

Table 3. Results from Indiana Dataset – Reading Outcome

	Treatment Effect	Standard Error	T-stat	Bias
RE Within-District Benchmark	2.73	4.26	0.64	n/a
Within-district	0.19	2.92	0.07	-2.54

Note: Estimates only include treatment schools in districts in which there is at least one non-treatment school.

Table 4. Results from Indiana Dataset – Math Outcome

	Treatment Effect	Standard Error	T-stat	Bias
RE Benchmark	8.15	4.90	1.66	n/a
School composition	10.86*	5.04	2.16	2.71
Geography & urbanicity	3.81	5.37	0.71	-9.39
Pretests	4.29	3.80	1.13	-3.86
All school level covariates except pretests	2.71	5.23	0.52	-5.44
Complete	2.72	3.68	0.74	-5.43
Super Team	1.67	5.75	0.29	-6.48

Table 5. Results from Indiana Dataset – Math Outcome

	Treatment Effect	Standard Error	T-stat	Bias
RE Within-District Benchmark	12.57*	5.68	2.21	n/a
Within-district	0.98	4.07	0.24	-11.59

Note: Estimates only include treatment schools in districts in which there is at least one non-treatment school.