Abstract Title Page
Not included in page count.

Title:

Constructing Counterfactuals in a Multisite Observational Study using Propensity Score Matching and Multilevel Modeling: An Empirical Example Looking at the Effect of 8th Grade Algebra across Students and Schools

Author(s):

Jordan H. Rickles
University of California, Los Angeles

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Certain questions about the effectiveness of an educational policy, program, or practice, ask researchers to make causal inferences under the following conditions: (1) random assignment is not practical or feasible; (2) assignment to the treatment condition is highly selective; (3) the assignment mechanism can vary across sites; and (4) the treatment effect can vary across students and sites. For example, school administrators might want to know whether an Advanced Placement curriculum affects student learning compared to the standard curriculum, or whether an out-of-school suspension affects subsequent student behavior differently than in-school detention. District decision makers might want to know whether a dropout prevention program keeps students in school, or whether grade retention improves long-term student outcomes. If researchers are asked to address these questions with pre-existing state or district data, they must confront the standard barriers to causal inferences that stem from a non-randomized design, and face further complication from the highly selective, and variable, nature by which students end up in the treatment condition. It may also be important to study how different school and classroom level factors mediate any treatment effect because the stable-unit-treatment-value assumption (SUTVA) is not likely to hold under these conditions.

This paper presents a methodology for estimating causal effects from a multi-site observational study that takes advantage of school-level variation in the assignment mechanism to construct balanced treatment and control groups. The methodology extends Stuart and Rubin's (2007) work on matching with multiple control groups when the treatment is within one school to the setting where the treatment is within multiple schools. I show how one useful tool for investigating treatment effect heterogeneity across students, classrooms, and schools is to impute each student's counterfactual outcome (Schafer and Kang, 2008) and use multilevel modeling to examine treatment effect variance. This technique allows researchers to examine whether the treatment effect varies across student characteristics (e.g., high vs. low ability students), across classrooms (e.g., peer and teacher characteristics), and across schools (e.g., treatment assignment policies).

The methods are demonstrated through an empirical example that seeks to determine whether students are better prepared for high school mathematics by taking a formal algebra course or a pre-algebra course in 8th grade. Prior research on the effects of early access to algebra used a variety of regression-based methods to adjust for selection bias. Ordinary least squares regression models (Gamoran & Hannigan, 2000), path analysis (Smith, 1996), and hierarchical linear growth models (Ma, 2005; Wang & Goldschmidt, 2003) have all been used to estimate the effects of algebra after controlling for various confounding factors. Additionally, propensity score methods have been employed recently to examine the related issue of curricular intensity (Attewell & Domina, 2008; Leow, Zanutto & Boruch, 2004).

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

This study seeks to demonstrate a method for treatment effect estimation in a multisite observational study where the treatment is highly selective and the assignment mechanism varies across sites. The method is demonstrated by addressing three primary research questions about the effect of 8th grade algebra:

1. For students who take algebra in 8th grade, what is the average effect of taking algebra in 8th grade on algebra achievement by the end of 9th grade?
2. Does the average effect vary across students with different levels of demonstrated 7th grade mathematics achievement and propensity for taking 8th grade algebra?
3. Does the average effect vary across classrooms and schools?

Through these three research questions, I focus on preprocessing the data with propensity score matching (Ho, Imai, King & Stuart, 2007) and imputation of the counterfactual (Schafer & Kang, 2008), and on the exploration of treatment effect heterogeneity with multilevel modeling. This paper recognizes, but does not directly address, the importance of sensitivity analysis.

Setting:
Description of the research location.

Data for this study cover a cohort of students from 54 middle schools within a California school district that spans urban and suburban communities.

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

The analysis is based on a cohort of 22,468 students who were 8th graders during the 2006-07 school year: 12,824 took algebra and 9,644 took pre-algebra. Key student characteristics for this cohort are presented in the first two columns of Table 1.

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

In California, and the district under study, about half of all 8th graders take algebra in 8th grade and the other half take a lower-level mathematics course. In the district under study, the lower-level, pre-algebra, mathematics course is called algebra readiness and primarily consists of 7th grade mathematics standards. The "treatment" in question is the assignment of 8th graders to an algebra course instead of this pre-algebra course.

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

The goal of the proposed methodology is to not just estimate an average causal effect, but to examine variability in the causal effect estimate across students and schools. If possible, an

effective design would be a multisite randomized design (or randomized block design with schools as the blocking variable). Since randomization was not possible for the current treatment under study, a propensity score matching design was developed to preprocess the data in a way that would approximate a multisite randomized design and allow for a description of student-level treatment effect variation. This required two research design innovations.

The primary obstacle to overcome in the design stage was the highly selective nature of student assignment to algebra. On average, students in 8th grade algebra (treatment group) exhibited significantly higher mathematics achievement in 7th grade than students in pre-algebra (control group). For example, 45% of the treatment students score proficient or advanced on the 7th grade mathematics CST, while only 8% of the control students score in those top two performance levels. If ignored, the lack of covariate overlap between treatment and control students within the same school means regression-based treatment effect estimates will be very dependent on modeling assumptions and extrapolation. Preprocessing the data with propensity score matching can lessen our dependence on parametric modeling assumptions (Ho, Imai, King & Stuart, 2007).

To construct matched treatment and control students within each school, I adapted a multiple control group strategy developed by Stuart and Rubin (2007). For each school, treatment students were first matched to a control student within the same school. Treatment students without an adequate within-school match, were then matched to students outside the school. To account for possible school effect bias resulting from outside-school matches, the outcome score for outside-school control students was adjusted by a school effect estimate derived from a third match between control students. A more detailed description of the matching process is presented in Table 2.

The matching process was repeated over all 54 schools, resulting in a matched sample of 10,744 treatment students. A description of the key student characteristics for the matched treatment and control students is presented in the fourth and fifth columns of Table 1. This table shows that average treatment and control group differences in key characteristics decreased dramatically after matching. Furthermore, Figure 1 shows that the matched treatment and control students exhibit very similar propensity score distributions after matching and Figure 2 shows that balance on the propensity score also holds within schools.

After the data preprocessing stage, each treatment student's counterfactual potential outcome (i.e., outcome score under control condition) was imputed from a random-intercept multilevel regression model based on the control students. Imputing each treatment student's counterfactual potential outcome is encouraged by Schafer and Kang (2008) and provides two main benefits over standard effect estimation methods for this study in particular. First, it facilitates post-hoc analysis of treatment effect variation where the treatment effect for student i in school j is defined as $\delta_{ij} = y(1)_{ij} - \hat{y}(0)_{ij}$, where $y(1)$ is the outcome under treatment and $\hat{y}(0)$ represents the imputed counterfactual. Second, it helps overcome complications with duplicated control students in the matched sample because the analysis is only based on treatment students. Future research will examine methods for incorporating uncertainty in $\hat{y}$ into the analysis through multiple imputation.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

The analysis utilizes longitudinal student-level data that are maintained by the school district. For the 2006-07 cohort of 8th graders, the data cover the 2004-05 through 2007-08 school years (i.e., 6th grade through 9th grade). The data include standard student demographics and performance on the annual state ELA and mathematics standardized tests, as well as course grades, school attendance, and school suspensions.

The data allow one to determine which school and classes (a combination of teacher, course, and period) a student was in each semester. Students were defined as taking algebra in 8th grade (treatment) if they were in an algebra 1 math course during the fall semester of their 8th grade year and were defined as taking pre-algebra (control) if they were in an algebra readiness math course during the fall semester of their 8th grade year. Available student characteristics and academic performance during the 6th and 7th grade years were used to estimate the propensity score.

The outcome was defined as a student's scale score on the algebra 1 CST. A student could have taken the algebra 1 CST in either 8th grade or 9th grade depending on whether the student was in the treatment or control group. Some treatment students took the algebra 1 CST in both 8th and 9th grades because they had to repeat algebra 1. Since the objective of the study is to assess algebra knowledge by the end of 9th grade, the 9th grade year CST was used if the student repeated algebra.

After the data were preprocessed and the treatment group counterfactual outcome was imputed (see Table 2), the treatment effect for 8th grade algebra students was analyzed with general descriptive statistics and multilevel linear models to estimate variation in the treatment effect across students, classroom, and schools. For example, the unconditional multilevel model for the treatment effect took the following form for student *i* in classroom *k* in school *j*:

$\delta_{ikj} = \pi_{0kj} + e_{ikj}$, $\pi_{0kj} = \beta_{0j} + r_{0kj}$, $\beta_{0j} = \gamma_0 + u_{0j}$, where $e_{ikj} \sim N(0,\sigma^2)$, $r_{0kj} \sim N(0,\tau_\pi)$, $u_{0j} \sim N(0,\tau_\beta)$

The extent to which certain student, classroom, and school characteristics explain the effect heterogeneity was assessed by adding those characteristics to the unconditional model.

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

The selective nature of 8th grade algebra and the lack of covariate overlap between algebra and pre-algebra students became apparent when preprocessing the data through propensity score matching. Table 1 shows large differences in the original algebra and pre-algebra means across key characteristics and Figure 1 shows how the propensity score log-odds distributions differ between the algebra and pre-algebra students. This lack of overlap between the treatment and control groups implies that little information exists to estimate treatment effects for students at the low and high range of, for example, 7th grade mathematics achievement. In fact, after preprocessing the data, it was clear that one cannot estimate the average treatment effect for all students without extrapolating. As a result, average treatment effect findings reflect a more

restricted population of 8th grade algebra students, where very high and very low achieving students are less represented. While this results in a loss of generalizability, it should increase the internal validity of the findings.

For the propensity score matched 8th grade algebra students, taking algebra in 8th grade had a positive effect on algebra achievement by the end of 9th grade. The magnitude of the average effect did, however, vary across students, classrooms, and schools. Results from the unconditional multilevel model of the 8th grade algebra effect are presented in Table 3. For the average student in the matched sample, taking algebra in 8th grade instead of a pre-algebra course resulted in a 15.78 scale score (or about 0.29 standard deviation) increase on the algebra CST. About 79% of the estimated treatment effect variance was between students, 14% was between classrooms and 6% was between schools. Both the classroom-level and school-level variance components are significantly larger than zero (p-value < 0.001).

About half of the student-level variance can be explained by differences in the treatment effect across student 7th grade mathematics performance, with higher achieving 7th grade students experiencing a higher treatment effect. However, even students who scored below basic on the 7th grade mathematics CST experienced a positive treatment effect, on average. Similarly, students with a higher propensity for taking algebra in 8th grade had a higher treatment effect. Preliminary exploration of treatment effect variance across the classroom- and school-level did not reveal a statistically significant relationship between the treatment effect and the composition of students in the classroom or school.

Conclusions:

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

The findings from this study suggest that, on average, students will attain more algebra knowledge by the end of 9th grade if they have the opportunity to take algebra in 8th grade instead of a pre-algebra course that repeats 6th and 7th grade mathematics content. The effectiveness of 8th grade algebra, however, varies across students, classrooms, and schools. Students entering 8th grade with higher mathematics aptitude are more likely to benefit from a formal algebra course, but even relatively low performing students are also likely to experience positive benefits from 8th grade algebra. Variability across classrooms and schools points to the importance of instructional quality and course content above and beyond a course title. More research is required to understand the mechanisms causing these differential effects and sensitivity analyses are required to see how the findings hold up when critical assumptions are relaxed.

While the findings are subject to the key identifying assumptions required for non-experimental research, the methods employed in this study allow for a detailed exploration of the 8th grade algebra effect through the potential outcomes framework. Further research is required to refine the methodology, but, when random assignment is not feasible, the combination of propensity score matching across multiple schools and multilevel modeling is a promising tool to examine causal effect heterogeneity in educational settings.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Attewell, P., & Domina, T. (2008). Raising the bar: curricular intensity and academic performance. *Educational Evaluation and Policy Analysis*, 30, 51-71.

Gamoran, A., & Hannigan, E. C. (2000). Algebra for everyone? Benefits of college-preparatory mathematics for students with diverse abilities in early secondary school. *Educational Evaluation and Policy Analysis*, 22 (3), 241-254.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reduced model dependence in parametric causal inference. *Political Analysis, 15* (3), 199-236.

Leow, C., Marcus, S., Zanutto, E., & Boruch, R. (2004). Effects of advanced course-taking on math and science achievement: addressing selection bias using propensity scores. *American Journal of Evaluation*, 25 (4), 461-478.

Ma, X. (2005). Early acceleration of students in mathematics: does it promote growth and stability of growth in achievement across mathematical areas? *Contemporary Educational Psychology*, 30, 439-460.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods*, *13*(4), 279–313.

Smith, J. (1996). Does an extra year make any difference? The impact of early access to algebra on long-term gains in mathematics attainment. *Educational Evaluation and Policy Analysis*, 18, 141-153.

Stuart, E. A., & Rubin, D. B. (2007). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3), 279-306.

Wang, J., & Goldschmidt, P. (2003). Importance of middle school mathematics on high school students' mathematics achievement. *Journal of Educational Research*, 97 (1), 3-19.

Appendix B. Tables and Figures

Not included in page count.

Table 1. Means for key student characteristics by 8th grade algebra and pre-algebra students in the original and matched samples.

	Original Sample			Matched Sample*		
	Alg. 1	Pre-Alg	Diff	Alg. 1	Pre-Alg	Diff
# of Students	12,824	9,644		10,744	10,744*	
% Female	53%	47%	5%	53%	52%	1%
% GATE	23%	3%	20%	17%	15%	1%
% Students w/ Disabilities	2%	11%	-8%	3%	4%	-2%
7th Grade Math GPA	2.50	1.35	1.15	2.35	2.06	0.29
7th Grade, Semester 2 Math Mark (%):						
A	26%	4%	22%	21%	17%	4%
B	27%	11%	15%	26%	24%	2%
C	25%	25%	0%	28%	29%	-1%
D	12%	28%	-16%	14%	16%	-2%
F	9%	31%	-22%	11%	15%	-4%
6th Grade Math CST Scale Score	345.81	288.98	56.83	335.42	327.44	7.98
7th Grade Math CST Scale Score	343.45	288.98	54.48	333.53	328.99	4.54
7th Grade Math CST Performance Level (%):						
Advanced	11%	0%	10%	6%	5%	1%
Proficient	35%	8%	27%	32%	30%	2%
Basic	32%	27%	5%	36%	35%	1%
Below Basic	18%	43%	-25%	21%	24%	-3%
Far Below Basic	4%	21%	-17%	5%	6%	-1%
6th Grade ELA CST Scale Score	334.19	295.48	38.71	327.33	323.75	3.57
7th Grade ELA CST Scale Score	344.37	300.96	43.41	336.96	331.72	5.24
Propensity Score (log odds)	2.46	-1.76	4.22	1.77	1.43	0.33
Propensity Score	0.80	0.26	0.54	0.77	0.72	0.05

* Number of students and mean statistics for the matched pre-algebra group based on the weighted matched sample. 5,477 unique pre-algebra students comprise the matched pre-algebra group. Within any school, the matched treatment and control students are unique. However, some matched control students are duplicated when looking across schools because the outside-school match for each school was conducted on all available control students.

Table 2. Description of steps to construct the matched treatment and control sample.

Step	Description
1.	Estimate the propensity score for each student *i* in school *j* based on the following model: $\log[p_{ij}/(1-p_{ij})] = \beta_{0j} + \mathbf{X_{1}}_{ij}\boldsymbol{\beta_{1}}_{j} + \mathbf{X_{2}}_{ij}\boldsymbol{\beta}_{2}$, $\beta_{0j} = \gamma_0 + u_{0j}$, $\boldsymbol{\beta_{1}}_{j} = \boldsymbol{\gamma_{1}} + \mathbf{u_{1}}_{j}$, where pij = probability of taking algebra in 8th grade $\mathbf{X_1}$ = vector of grand-mean centered indicator variables for the student's 6th grade mathematics CST performance level $\mathbf{X_2}$ = vector of grand-mean centered student characteristics covering student demographics, prior academic achievement, 7th grade school attendance, and ever suspended in 7th grade.
2.	For school j, conduct a caliper 1-to-1 propensity score match without replacement using a caliper of 0.25 sd of the propensity score log-odds. Call this matched set $M1_j$.
3.	For treatment students in school j that are not in $M1_j$, conduct a caliper 1-to-1 propensity score match without replacement using a caliper of 0.25 sd of the propensity score log-odds with all control students not in school j. Call this matched set $M2_j$.
4.	For control students in M1j, conduct a caliper 1-to-1 propensity score match without replacement using a caliper of 0.25 sd of the propensity score log-odds with all control students not in school j. Call this matched set MC_j.
5.	Repeat steps 2 through 4 for all schools.
6.	Combine the M1 and M2 files for all schools into one data file (M) and combine the MC files for all schools into one data file (MC).
7.	Using the MC file, estimate school effects based on a multilevel linear model where the outcome is a linear function of student propensity score log-odds and an intercept that varies across schools.
8.	Adjust the observed outcome for control students in M2 for the difference between the estimated school effect for the control student and the estimated school effect for the matched treatment student. For example if the estimated school effect for school 1 is 10 and the school effect for school 2 is 15, then the outcome value for a control student in school 2 matched to a treatment student in school 1 would be discounted by 5.
9.	Estimate the counterfactual outcome for treatment students in M using on a multilevel linear model based on control students in M. The multilevel model includes the same student characteristics as in step 1 and allows the intercept to vary across schools.

Note: steps 2 through 8 adapted from Stuart and Rubin (2007).

Table 3. Grand-mean and variance estimation of the 8th grade algebra treatment effect from an unconditional 3-level model.

Fixed Effect	Coef.	df	se	t-ratio	p-value
Average effect of algebra, γ_0	15.78	53	1.57	10.03	0.000
Random Effect	**Variance**	**df**		**χ^2**	**p-value**
Students (level 1), e_{ikj}	1251.17				
Classrooms (level 2), r_{0kj}	226.83	480		1237.31	0.000
Schools (level 3), u_{0j}	97.39	53		213.69	0.000
Variance Decomposition (Percentage by Level)					
Students (level 1)	79.4%				
Classrooms (level 2)	14.4%				
Schools (level 3)	6.2%				

Note: based on 10,744 treatment students in matched sample.

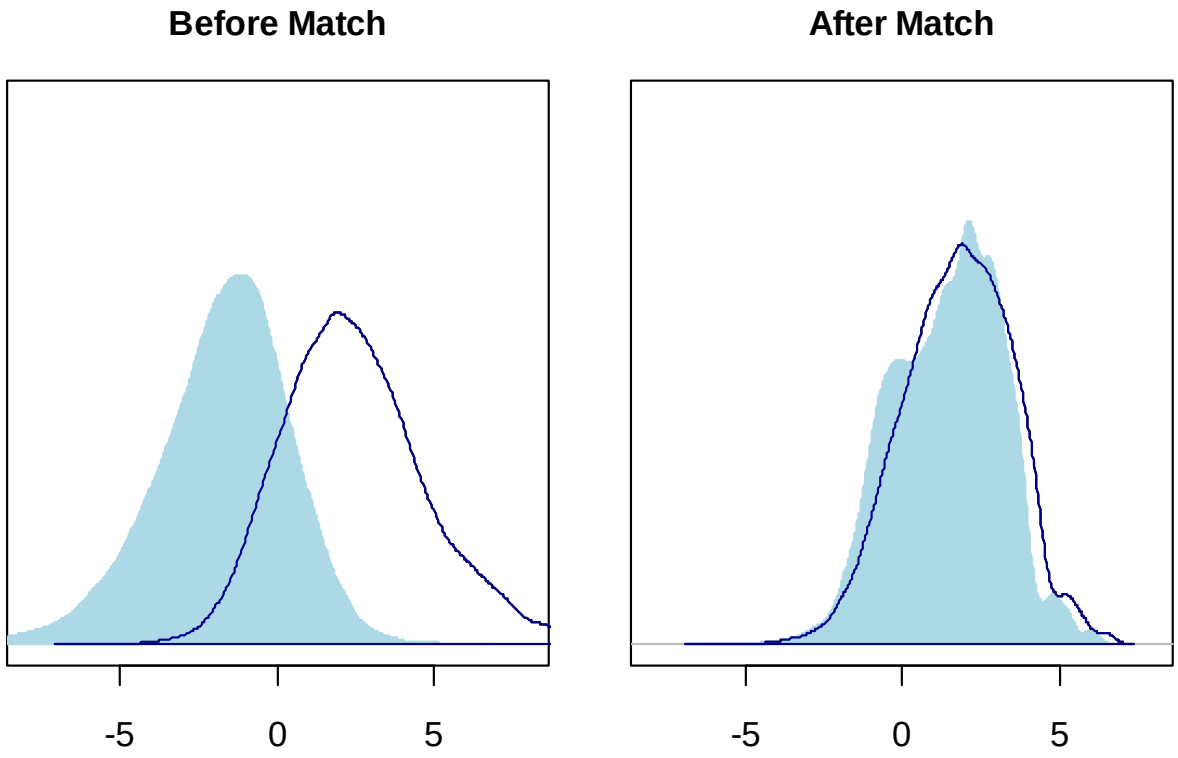

Figure 1. Empirical kernel density plot of the propensity score log-odds for 8th graders in algebra (dark blue line) and pre-algebra (shaded light-blue area) before and after matching.

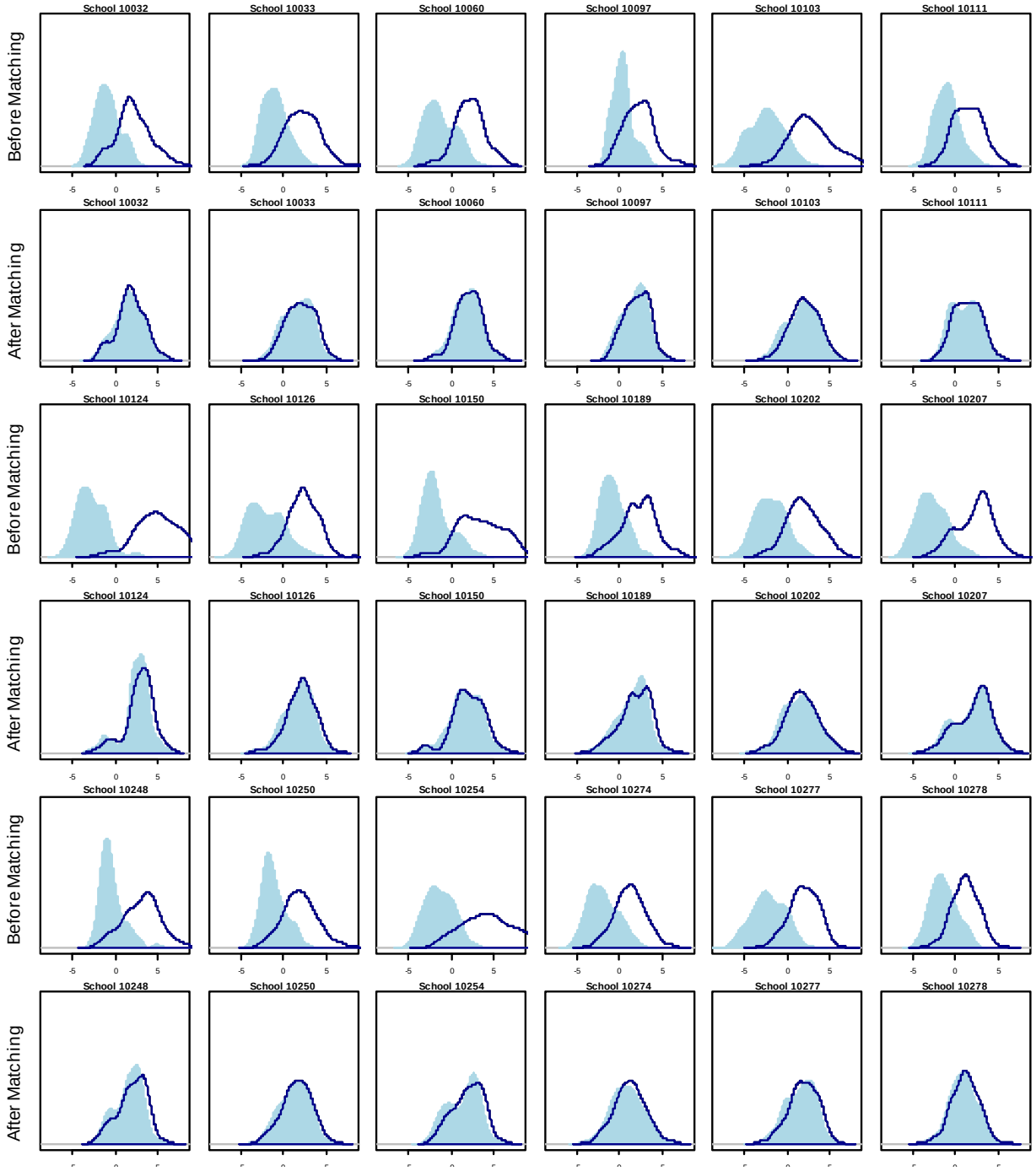

Figure 2. Empirical kernel density plots of the propensity score log-odds for 8th graders in algebra (dark blue line) and pre-algebra (shaded light-blue area) before and after matching by school. Note: only the first 18 schools are shown.